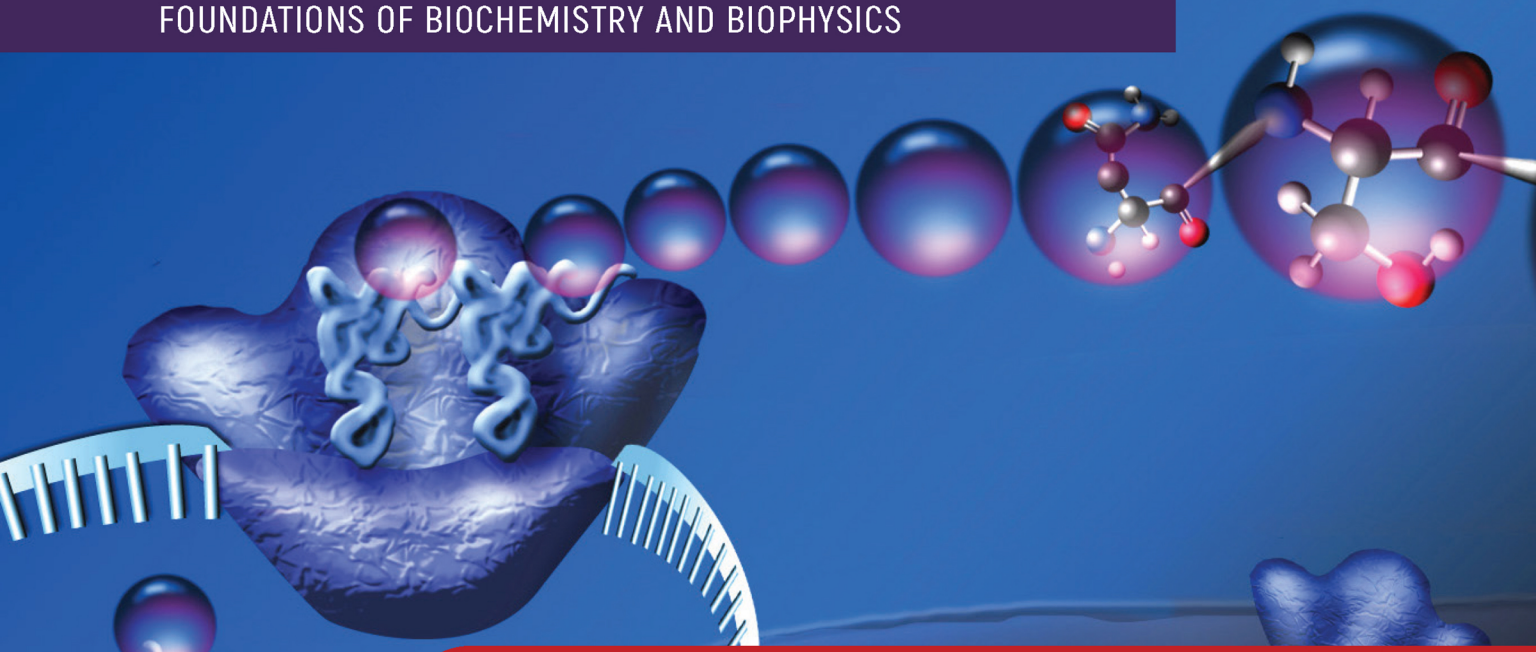
FOUNDATIONS OF BIOCHEMISTRY AND BIOPHYSICS

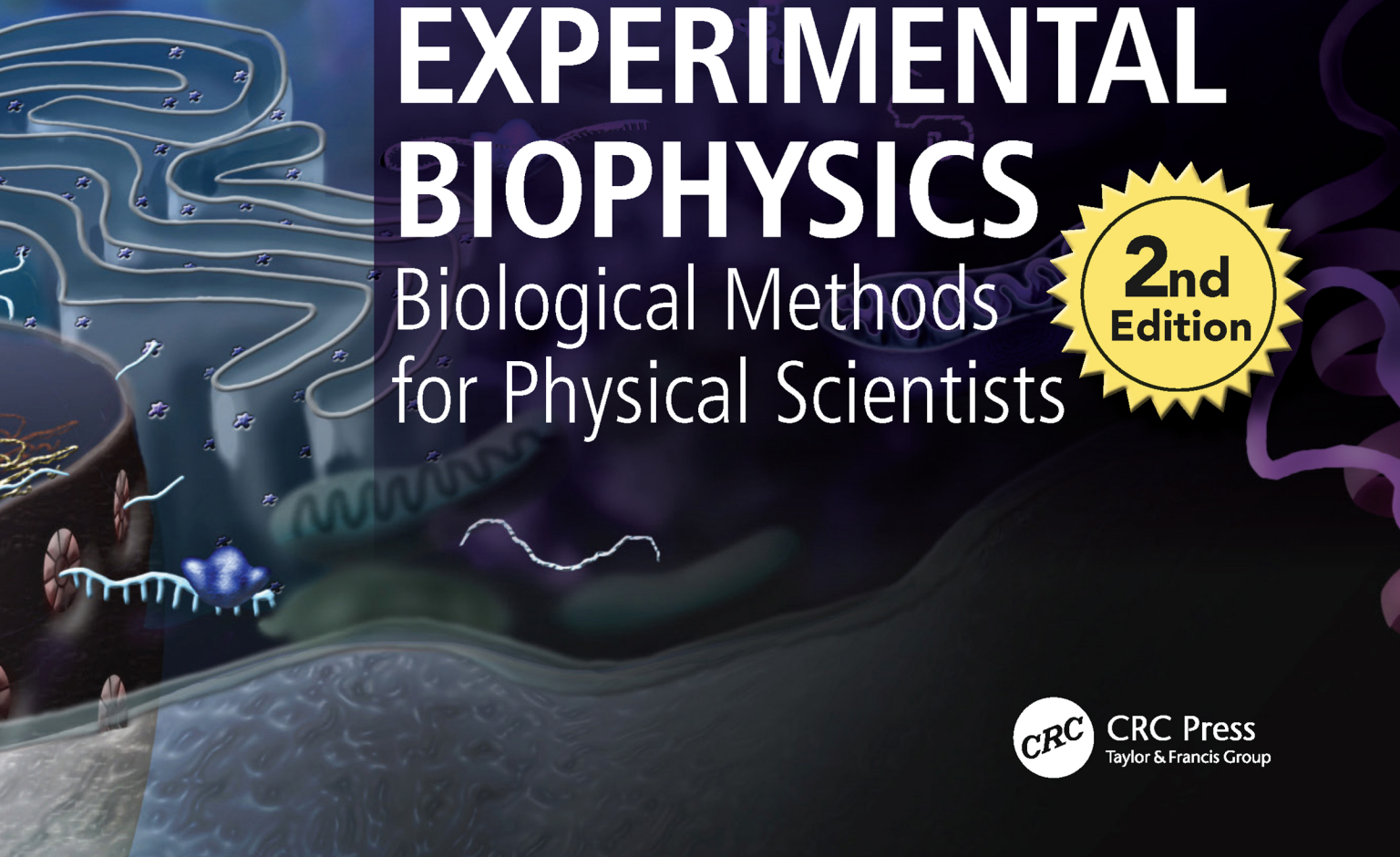


JAY L. NADEAU

Introduction to EXPERIMENTAL BIOPHYSICS

Biological Methods for Physical Scientists

2nd Edition



CRC Press
Taylor & Francis Group

Introduction to Experimental Biophysics

Biological Methods for Physical Scientists, Second Edition

FOUNDATIONS OF BIOCHEMISTRY AND BIOPHYSICS SERIES

Introduction to Experimental Biophysics:
Biological Methods for Physical Scientists,
Second Edition
Jay L. Nadeau

Introduction to Single Molecule Biophysics
Yuri L. Lyubchenko

Biomolecular Thermodynamics: From Theory to Application
Douglas Barrick

Biomolecular Kinetics: A Step-by-Step Guide
Clive R. Bagshaw

An Introduction to Biophysics:
Quantitative Understanding of Biosystems,
Second Edition
Thomas M. Nordlund and Peter M. Hoffmann

Introduction to Experimental Biophysics

Biological Methods for Physical Scientists, Second Edition

Jay L. Nadeau

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper

International Standard Book Number-13: 978-1-4987-9959-1 (Paperback); 978-1-138-08815-3 (Hardback)

Library of Congress Cataloging-in-Publication Data

Names: Nadeau, Jay L., author.
Title: Introduction to experimental biophysics : biological methods for physical scientists / Jay L. Nadeau.
Other titles: Experimental biophysics | Foundations of biochemistry and biophysics.
Description: Second edition. | Boca Raton, FL : CRC Press, Taylor & Francis Group, [2017] |
Series: Foundations of biochemistry and biophysics
Identifiers: LCCN 2017010261| ISBN 9781138088153 (hardback) | ISBN 1138088153 (hardback) |
ISBN 9781498799591 (pbk. ; alk. paper) | ISBN 1498799590 (pbk. ; alk. paper)
Subjects: LCSH: Biophysics--Experiments--Technique.
Classification: LCC QH505 .N247 2017 | DDC 572--dc23
LC record available at https://lccn.loc.gov/2017010261

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Detailed Contents

Series Preface

Biophysics encompasses the application of the principles, tools, and techniques of the physical sciences to problems in biology, including determination and analysis of structures, energetics, dynamics, and interactions of biological molecules. Biochemistry addresses the mechanisms underlying the complex reactions driving life, from enzyme catalysis and regulation to the structure and function of molecules. Research in these two areas is having a huge impact in pharmaceutical sciences and medicine.

These two highly interconnected fields are the focus of this book series. It covers both the use of traditional tools from physical chemistry, such as nuclear magnetic resonance (NMR), x-ray crystallography, and neutron diffraction, as well as novel techniques including scanning probe microscopy, laser tweezers, ultrafast laser spectroscopy, and computational approaches. A major goal of this series is to facilitate interdisciplinary research by training biologists and biochemists in quantitative aspects of modern biomedical research and teaching core biological principles to students in physical sciences and engineering.

Proposals for new volumes in the series may be directed to Lu Han, senior publishing editor at CRC Press, Taylor & Francis Group (lu.han@taylorandfrancis.com).

Preface

The second edition has been revised and updated to reflect changes in the fields between 2010 and 2016, with references, suppliers, and software all brought up to date. The study questions at the back of each chapter have been thoroughly revised and expanded, and a solutions manual is available.

The book has also been restructured to make a clear distinction between the basic techniques and more advanced approaches that are usually not accessible to an undergraduate laboratory. The book can be used on two levels: as an introductory course with only the basic techniques covered or as a more advanced course that requires access to more sophisticated equipment. The advanced material may also be used for self-study.

The advanced material is included within selected chapters as callouts, as well as forming the basis of five entirely new chapters: advanced molecular biology techniques (**Chapter 4**), advanced light microscopy (**Chapter 8**), holographic microscopy (**Chapter 9**), biomedical applications of gold nanoparticles (**Chapter 13**), and microfabrication techniques (**Chapter 17**).

A large fraction of the basic course material provides the basis for a one-semester or summer course on introductory molecular biology techniques.

This textbook is bundled with a laboratory companion guide. It is structured according to the chapters in the book, although it refers to the first edition of this book. The series of 14 experiments presents a wide variety of techniques that may be performed during a semester-long, three-credit course or during a 1-month intensive.

Acknowledgments

I thank all of the people who made this book possible. The biggest thanks are to the chapter authors, who provided years of firsthand experience on how to do things right (or wrong). Sections of some chapters were also contributed by colleagues. I am grateful to Chris Ratcliffe of the National Research Council, who wrote the section on solid-phase nuclear magnetic resonance (NMR) in **Chapter 16**, and to Jonathan Saari of McGill, who contributed the section on time-resolved absorption spectroscopy in **Chapter 16**.

A special mention also goes to Jenna Blumenthal, who was a senior undergraduate in physics/physiology when she helped to proofread the first edition and prepare the first version of the glossary. Ildiko Horvath of McGill drew some of the illustrations in **Chapter 1**. Thanks also to my former graduate students Samuel Clarke, Xuan Zhang, and Daniel Cooper, whose thesis material is incorporated into several of the chapters.

Another thanks goes to all of the people who provided figures, both published and unpublished, to help illustrate this work. When approached out of the blue, they responded with data, micrographs, and other material that allowed the illustrations to be as beautiful, relevant, and practical as I hoped they might be.

Thanks to those who helped to proofread, and special hugs to Susan Foster, the world's best copy editor.

Finally, this book would not have been possible without my editor, Lu Han, who helped develop the book's idea, encouraged me throughout its evolution, and solicited the second edition long before I had started to think about it.

Author

Jay L. Nadeau is an associate professor of physics at Portland State University (PSU). Prior to PSU, she was a research professor in the Graduate Aerospace Laboratories (GALCIT) at the California Institute of Technology (2015–2017) and an associate professor of biomedical engineering and physics at McGill University (2004–2015). Her research interests include nanoparticles, fluorescence imaging, and development of instrumentation for the detection of life elsewhere in the solar system.

She has published over 70 papers on topics ranging from theoretical condensed matter physics to experimental neurobiology to the development of anticancer drugs and, in the process, has used almost every technique described in this book. Her work has been featured in *New Scientist*, *Highlights in Chemical Biology*, Radio Canada's *Les Années Lumière*, *Le Guide des Tendances*, and educational displays in schools and museums. Her research group features chemists, microbiologists, roboticists, physicists, and physician–scientists, all learning from each other and hoping to speak each other's language. A believer in bringing biology to physicists as well as physics to biologists, she has created two graduate-level courses: methods in molecular biology for physical scientists and mathematical cellular physiology. She has also taught pharmacology in the medical school and was one of the pioneers in the establishment of multiple mini-interviews for medical school admission.

She retains adjunct positions at McGill and Caltech and has collaborators in industry and academia in the United States, Europe, Australia, and Japan. She has given several dozen invited talks at meetings of the American Chemical Society, the American Geophysical Union, the International Society for Optics and Photonics (SPIE), the Committee on Space Research, the American Association of Physics Teachers (AAPT), and many others. Before her time at McGill, she was a member of the Jet Propulsion Laboratory's Center for Life Detection, and previous to that, a Burroughs Wellcome postdoctoral scholar in the laboratory of Henry A. Lester at Caltech. She earned a PhD in physics at the University of Minnesota in 1996.

Contributors

Edward S. Allgeyer
Department of Chemical and Biological Engineering
University of Maine
Orono, Maine

Manuel Bedrossian
Graduate Aerospace Laboratories
California Institute of Technology
Pasadena, California

Oliver M. Baettig
Department of Biochemistry
McGill University
Montreal, Quebec, Canada

Albert M. Berghuis
Department of Biochemistry
McGill University
Montreal, Quebec, Canada

Lina Carlini
Laboratory of Experimental Biophysics
EPFL, Switzerland

Gary Craig
Department of Chemical and Biological Engineering
University of Maine
Orono, Maine

Michael W. Davidson
National High Magnetic Field Laboratory
Florida State University
Tallahassee, Florida

Jeremy Grant
Department of Chemical and Biological Engineering
University of Maine
Orono, Maine

Sanjeev Kumar Kandpal
Department of Chemical and Biological Engineering
University of Maine
Orono, Maine

Thomas Knöpfel
Riken Brain Science Institute
Saitama, Japan

Christian Lindensmith
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Michael D. Mason
Department of Chemical and Biological Engineering
University of Maine
Orono, Maine

Joshua A. Maurer
Department of Chemistry
Washington University
St. Louis, Missouri

Orad Reshef
Department of Physics
Harvard University
Cambridge, Massachusetts

CHAPTER 1

Introduction and Background

1.1 BASIC BIOCHEMISTRY

Molecules important to molecular biophysics

The chemicals of life are *organic* compounds, or compounds that contain carbon. Carbon (C, atomic number 6) is one of the few *tetravalent* atoms, meaning that it has four *valence* electrons available to form bonds with other atoms. Each of the four atoms to which it bonds can be different and can include other carbons. Carbon is thus central to the formation of complex, three-dimensional molecules, and it makes up about 10.7% of the atomic ratio of living matter. Other molecules necessary for the building blocks of life are hydrogen (H, atomic number 1, monovalent, 60.5%); oxygen (O, atomic number 8, divalent, 25.7%); nitrogen (atomic number 7, trivalent, 2.4%); phosphorus (P, atomic number 15, trivalent up to hexavalent, 0.17%); and sulfur (S, atomic number 16, divalent, tetravalent, or hexavalent, 0.13%).

The valence of the key elements forms the basis of the *structural model* of organic chemistry that permits us to predict which combinations of atoms will combine to form stable molecules. **Figure 1.1** shows the classes of organic molecules that are most important in biochemistry and their functional groups. If the letter *R* is used to designate any chemical moiety besides hydrogen, then an *amine* has the general formula RNH_2 (for a *primary amine*), R_2NH (for a *secondary amine*), or R_3N (for a *tertiary amine*). A *carboxylic acid* is RCOOH; at physiological pH, it will usually dissociate into a free proton (H^+) and a negatively charged ion $RCOO^-$ (called a *carboxylate*). A *ketone* is RCOR where the second R is not an OH group. *Phosphates* in biology have the form RPO_3^{2-}; when R is OH, this is referred to as inorganic *phosphate* or P_i. *Alcohols* are ROH where R can be nearly anything; any biomolecule with a name ending in *-ol* terminates in an OH group. A *sulfhydryl*, also known as a thiol group, is RSH. Thiols are also known as mercaptans. Finally, an *aromatic* group is a planar ring that may be made of carbon only or of carbon plus oxygen, nitrogen, or sulfur (called *heterocyclic* compounds). The simplest example is the six-carbon benzene ring.

The structural and functional makeup of a cell results from combinations of four basic molecular types, each of which falls into one or more of the categories in

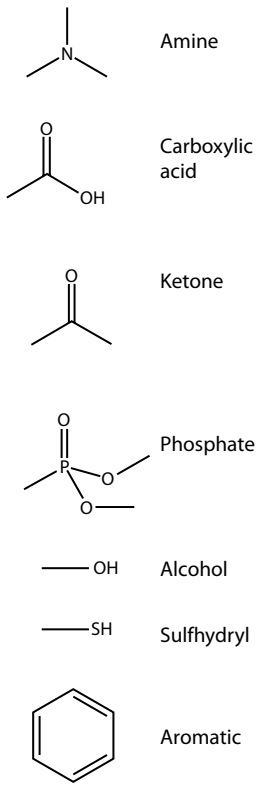


Figure 1.1 Functional groups seen in biochemistry.

Figure 1.1; these molecules join end to end (*polymerize*) to achieve their final active form:

- *Amino acids* (polymerize to form *peptides* and *proteins*). There are twenty naturally occurring amino acids, whose structure consists of a central carbon atom with a carboxylic acid on one end and a primary amine on the other, and a *side chain* that branches off the first carbon after the amine. The side chain determines the amino acid's identity and ranges from a hydrogen (glycine) to complex charged or aromatic groups (**Figure 1.2**). Short chains of amino acids are called peptides and may be synthesized by organisms like fungi in order to kill bacteria. The example shown is bacitracin, which is a cyclic peptide active against many bacteria; it is often found in first-aid creams. Some peptides are available from biological suppliers, and custom peptides are also available, though costly. Full-length proteins are encoded genetically and synthesized as a long polypeptide chain. They then fold to form their final *tertiary structure*; the example shown is *green fluorescent protein*, or GFP, which has 238 amino acids. The physics of protein folding still remains largely a mystery. Proteins usually cannot be purchased but must be expressed and purified by the experimenter (**Figure 1.3a**).

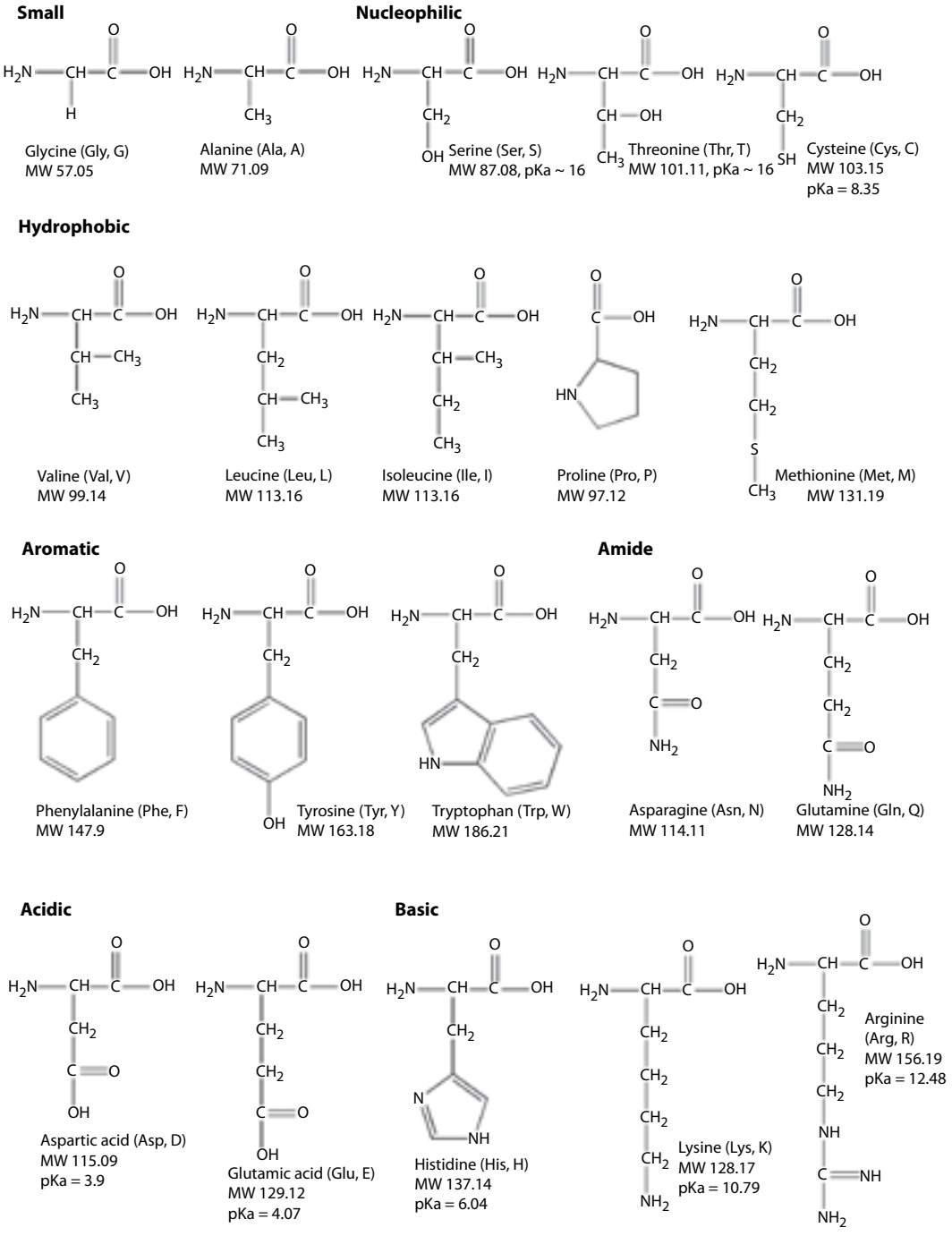


Figure 1.2 The 20 naturally occurring amino acids, showing their one- and three-letter abbreviations, their molecular weights, and their acid dissociation constants (pKa values).

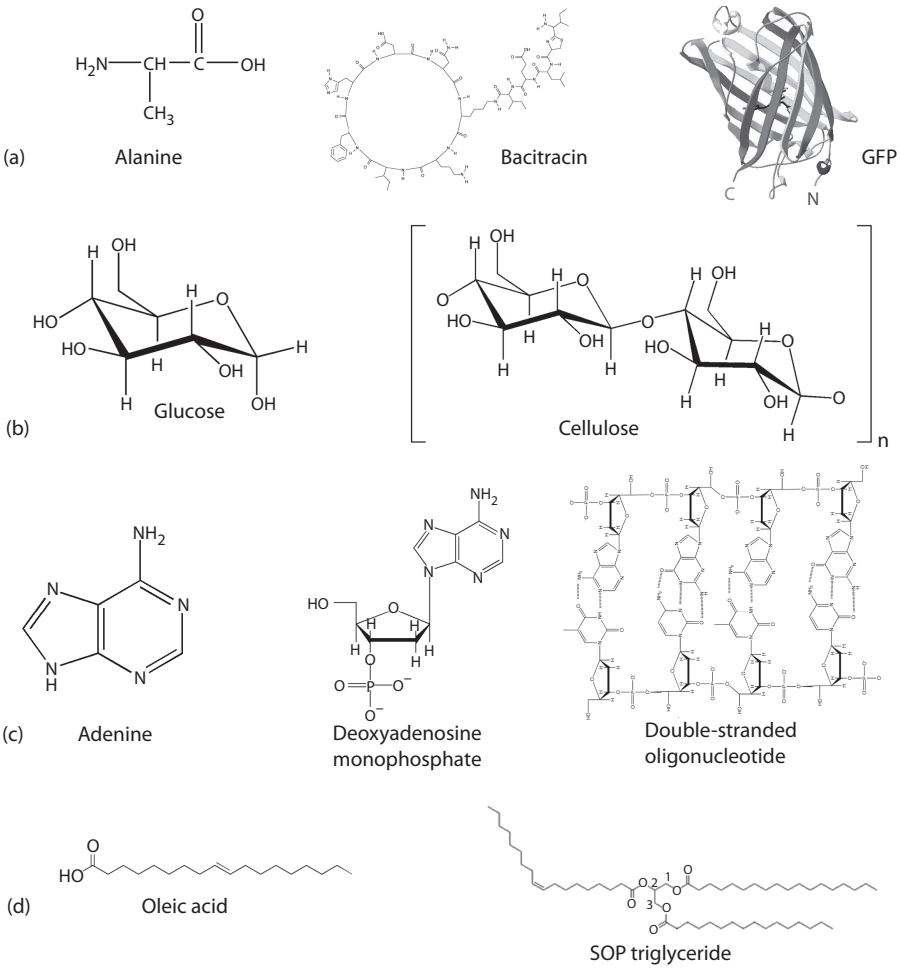


Figure 1.3 Monomers and polymers of living systems. Images are not to scale with each other. (a) An amino acid (alanine; side chain CH_3), a peptide (bacitracin), and a protein (GFP). (b) A monosaccharide or simple sugar (glucose), and the polymer of glucose (cellulose). (c) A DNA base (adenine), a nucleotide (deoxyadenosine monophosphate), and a double-stranded oligonucleotide. (d) A fatty acid (oleic acid) and a triglyceride (SOP: steric, oleic, and palmitic acid).

- *Monosaccharides* (polymerize to form *polysaccharides*). Shown in **Figure 1.3b** is glucose (also known as grape sugar or corn sugar), which is the major source of fuel for every living cell on Earth. The active form in biology is right-handed and polarizes light to the right; thus, it is often called simply dextrose, especially in the food industry (see **Advanced Topic 1.1**). Monosaccharides can polymerize to form important storage and structural molecules. Storage molecules include starch and glycogen; the latter is what provides energy after carbo loading. Structural molecules include some of the most abundant natural materials in the world: chitin (a polymer of a glucose derivative found in fungi, arthropods, crustaceans, and insects) and cellulose (a polymer of glucose, the primary component of wood; **Figure 1.3b**).

- *Nucleotides* (polymerize to form *nucleic acids* [DNA, RNA]). DNA is made of four *nitrogenous bases*: adenine, guanine, cytosine, and thymine (abbreviated A, G, C, and T). Adenine is shown in **Figure 1.3c**. A and G are *purines*, while T and C are *pyrimidines* (**Figure 1.4**). When the base is linked to a sugar (in the case of DNA, this sugar is *deoxyribose*; in the case

ADVANCED TOPIC 1.1: CHIRALITY

Many organic compounds are not identical to their mirror images. These molecules are called *chiral*, from the Greek *cheir* ("hand"), since human hands are also mirror images of each other but not superposable. In general, any tetrahedral atom with four different groups attached to it will be chiral. This includes all of the amino acids except glycine, all the monosaccharides, and many other compounds (**Figure A1.1.1a**).

Chirality is of great importance in chemistry and biology for several reasons. First, the chemistry of the right- and left-handed *enantiomers* of the same compound is not identical. Although they have the same molecular weight, solubility properties, index of refraction, and melting and boiling points, they behave differently when they interact with other chiral compounds or with light. The easiest way to observe chirality is to use a polarimeter to observe the rotation of plane-polarized light as it passes through the substance in question. A clockwise rotation is characteristic of a dextrorotatory or right-handed substance; a counterclockwise rotation indicates a levorotatory or left-handed enantiomer.

In biology, one enantiomer or the other is preferred almost exclusively. With a few exceptions in bacteria, sugars in biological systems are D- and amino acids are L- (where D and L refer to structure and not necessarily to the way in which they polarize light). Enzymes are all correspondingly chiral. The opposite-handed compounds have no nutritional value, and large amounts of D-amino acids may be harmful. Some drugs are hazardous only in one enantiomeric form; the best example may be thalidomide, which acts as a sedative and appetite enhancer in its right-handed form but whose left-handed form is highly teratogenic (**Figure A1.1.1b**). Different enantiomers also often taste and smell different, reflecting their differing interactions with our receptors and enzymes.

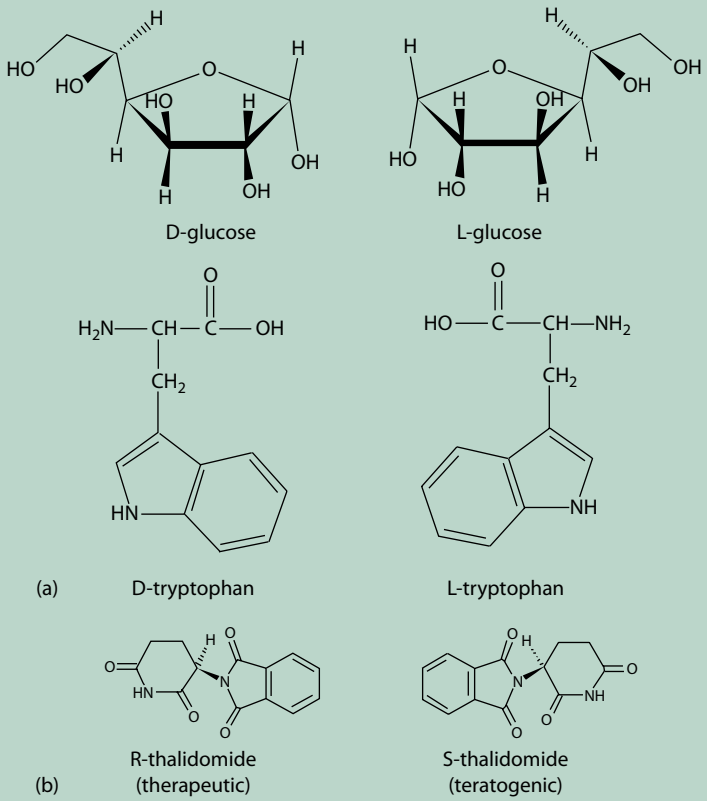


Figure A1.1.1 Chirality. (a) Sugars and amino acids are chiral because their mirror images cannot be superposed. (b) Thalidomide is a good example of how different enantiomers react differently with biological systems.

(*Continued*)

ADVANCED TOPIC 1.1 (CONTINUED): CHIRALITY

The origin of this exclusiveness, called *homochirality*, is unknown and widely studied because of its implications for the evolution of life on Earth and for the search for life on other planets. It is possible that the "choice" of one enantiomer or another was an evolutionary accident—i.e., an enzyme happened to evolve for an L-amino acid, thereby driving selection for all L-amino acids in the future. If the former is true, then life on other planets would be expected to be homochiral, but not necessarily in the same way as Earth life. Organic molecules that form from abiotic processes, however, should not show this homochirality but instead exist as *racemic mixtures* of both entantiomers. (Indeed, chiral mixtures left to their own devices are found to eventually *racemize*; this fact can be used as a dating technique.)

However, some physicists believe that the observed forms of these molecules are thermodynamically favored, possibly by an asymmetry in the weak force. If this is true, all molecules throughout the universe would be expected to be homochiral, or at least have an *enantiomeric excess*. Thus, finding homochirality on another planet would not be a sign of life. Finding the solution to this problem will have important implications for the design of orbital and landed extraterrestrial missions.

SUGGESTED READING

Bakasov, A., Ha, T.K., and Quack, M. (1998). Ab initio calculation of molecular energies including parity violating interactions. *Journal of Chemical Physics* 109, 7263–7285.

Barron, L.D. (2008). Chirality and life. *Space Science Reviews* 135, 187–201.

Borchers, A.T., Davis, P.A., and Gershwin, M.E. (2004). The asymmetry of existence: Do we owe our existence to cold dark matter and the weak force? *Experimental Biology and Medicine* 229, 21–32.

MacDermott, A.J. (2000). The ascent of parity-violation: Exochirality in the solar system and beyond. *Enantiomer* 5, 153–168.

of RNA, *ribose*), it is called a *nucleoside*: e.g., adenine becomes adenosine (in RNA) or deoxyadenosine (in DNA). Addition of one or more phosphate groups makes it a *nucleotide*. Short chains of A, C, G, and T nucleotides form *oligonucleotides* (if there are few, usually 20 or fewer bases) or *polynucleotides* (for longer chains). Oligonucleotides may be purchased from many suppliers and are inexpensive. As provided, they are single-stranded. However, DNA in nature is usually double-stranded, with A being complementary to T and C to G due to complementary hydrogen bonding (**Figure 1.3c**). *Complementary* oligonucleotides can be made to hybridize into double strands by simply heating them to 95°C and then allowing them to cool. However, if they are not fully complementary, the final double-stranded form is much less stable, and the strands can separate at relatively low temperatures. This fact forms the basis of much of molecular cloning and many types of *biosensors*.

- *Fatty acids* (form diglycerides and triglycerides by *dehydration synthesis*). Free fatty acids are molecules with a long carbon chain terminated in a carboxylic acid (**Figure 1.3d**). Fatty acids are crucial components of every cell, as they are the principal constituents of *cell membranes*. Most dietary fats, as well as fats stored in our own bodies, are in the form of triglycerides, which is a *glycerol* head attached to three fatty acid tails. The composition of these tails varies widely and plays an important role in the taste and texture of fatty foods. The number of double bonds in a fatty acid is called its degree of unsaturation and determines its melting temperature. Fully *saturated* fats (no double bonds) are solid at room

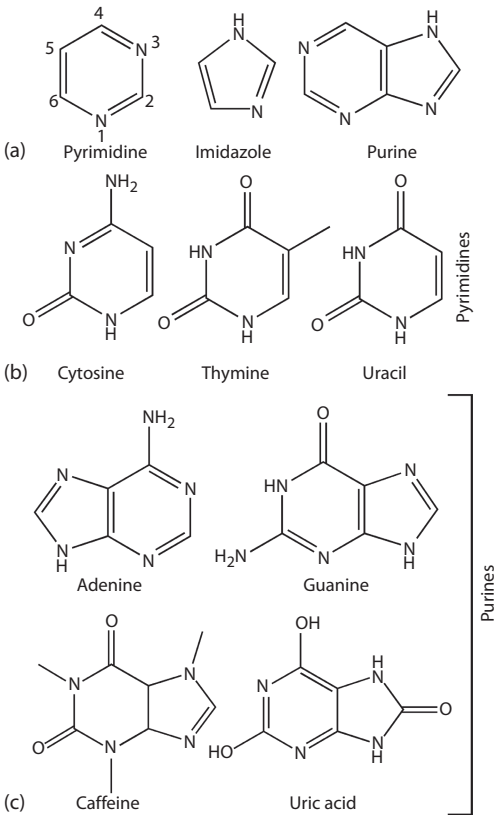


Figure 1.4 Purines and pyrimidines. (a) The structure of the pyrimidine ring is shown with its carbon-numbering convention. A pyrimidine ring bound to imidazole makes purine. (b) Cytosine, thymine, and uracil are derivatives of pyrimidine. (c) There is a very large number of biologically relevant purines, including the bases adenine and guanine as well as caffeine, uric acid, and many more.

temperature (butter is >50% saturated) whereas *unsaturated* fats are liquid (canola oil is nearly 95% unsaturated). A mix of different numbers of double bonds in the three chains allows triglycerides to have very complex melting properties. The triglyceride shown in **Figure 1.3d** is one found in cocoa butter, and its melting properties are responsible for the "melt in your mouth, not in your hand" nature of chocolate.

Making use of functional groups

The different functional groups of the molecules in **Figure 1.1** can be manipulated to create new bonds. Some of these groups are highly reactive, and simple reagents known as *cross-linkers* can catalyze their reactions with a complementary group. For example, a carboxylic acid and an amine can be joined in an *amide bond*; a carboxylic acid and an alcohol can be linked to form an *ester*, or a carboxylic acid and a thiol to a *thioester*; a phosphate can be linked to two other molecules via a *phosphodiester* bond; or two thiols can form a *disulfide* bond (**Figure 1.5a**). Sulfur also forms strong bonds to gold by mechanisms that are still being investigated. These principles can be used to adhere biomolecules to a surface or a particle, a process called *biofunctionalization*; to label a biomolecule with a dye (many dyes are sold that are made prereactive to a specific functional group; see **Chapters 7 and 8**); or simply to join two or more biomolecules (**Figure 1.5b**). Biofunctionalization of nanoparticles will be covered more fully in **Chapters 11 through 13**, and surface functionalization is treated in **Chapter 14**. This is a broad and complex field and is the subject of several excellent review articles and textbooks referenced at the end of each of these chapters.

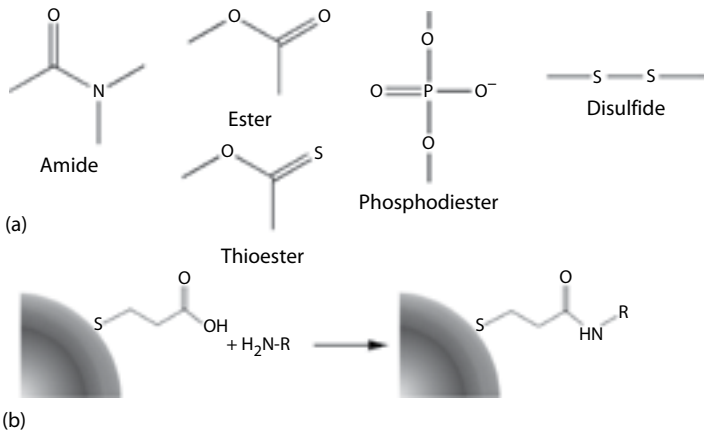


Figure 1.5 Linking biomolecules. (a) Types of bonds that can be made by linking amines, carboxylic acids, alcohols, phosphates, and/or thiols. (b) Biofunctionalization example. A gold-covered surface—which may be a nanoparticle, slide, tip, cantilever, etc.—is coated with a molecule containing a thiol group, one or more carbon atoms, and a carboxylic acid. It is then reacted with any molecule containing a primary amine to give an amide bond. Note that all proteins contain both primary amines and carboxylic acids, as they are made up of amino acids.

1.2 ENERGIES AND POTENTIALS

Biologically relevant energy scales

The structural model is empirical; it was developed by August Kékulé, Archibald Scott Couper, and Alexander M. Butlerov independently between 1858 and 1861. It does not provide any mechanistic description of bond formation, which had to wait until the invention of quantum mechanics for the development of a theory of *orbital* formation based upon electron wave functions (see **Advanced Topic 1.2**). Quantum mechanics also describes several other types of interatomic and intermolecular forces besides *covalent bonds*, all of which are equally crucial to biology, and without which molecules such as DNA could not exist. **Table 1.1** lists some examples of these forces and their relative energies. For comparison, kT at room temperature is 2.5 kJ/mol.

Ionic bonds

An *ionic bond* can be thought of as a covalent bond in which one of the partners is more *electronegative* than the other—that is, it has a stronger affinity for the shared

Table 1.1

Types of Interatomic/Intermolecular Interactions and Their Relative Strengths

Type of Interaction	Example	Bond Energy (kJ/mol)
Covalent bond	C–C	200–400
Covalent double bond	C=C	600–800
Ionic bond	Na–Cl	700–1000
Hydrogen bond	O–H	10–40
Ion–dipole interaction	Na^+–H_2O	40–600
Dipole–dipole interaction	HCl–HCl	5–25
Ion–induced dipole interaction	Cl^-–hexane	3–15
Dipole–induced dipole interaction	H_2O–Ar	2–10
London dispersion interaction	Hexane–octane	.05–2
Cation–pi interaction	Benzene…Na^+	2–50

ADVANCED TOPIC 1.2: A QUANTUM MECHANICAL DESCRIPTION OF BONDING: MOLECULAR ORBITAL THEORY

The first major conceptual breakthrough in the quantum mechanics of chemical bonding was the idea that bond energy results from exchange (*resonance*) of electrons between two nuclei. For example, for the hydrogen molecule, Heitler and London expressed the wave function of the two electrons as a spin singlet part Ψ_s and spin triplet part Ψ_t:

$$\begin{aligned}\Psi_s(1,2) &= N_s\left[\phi_n(r_1)\phi_m(r_2)+\phi_m(r_1)\phi_n(r_2)\right]\chi_s(s_1,s_2)\\ \Psi_t(1,2) &= N_t\left[\phi_n(r_1)\phi_m(r_2)-\phi_m(r_1)\phi_n(r_2)\right]\chi_t(s_1,s_2),\end{aligned} \tag{A1.2.1}$$

where the N's are normalization constants, the ϕ's are spatial wave functions, and the χ's are spin wave functions. They then calculated energy shift resulting from bonding as a perturbation:

$$\Delta E = \int \Psi^* H \Psi$$

where H is a sum of the interactions between the two electrons (1, 2), the electrons and the nuclei (a, b), and the two nuclei (separated by distance R):

$$H = \frac{e^2}{R}+\frac{e^2}{r_{12}}-\frac{e^2}{r_{a2}}-\frac{e^2}{r_{b1}}. \tag{A1.2.2}$$

The results show that the singlet state has an energy level lower than that of the ground state of a single atom. (It is called the *bonding orbital*.) The triplet state has a higher energy and so is called the *antibonding orbital* (**Figure A1.2.1a**). The sum of the energies of these orbitals gives an attractive potential that approaches r^{-6} behavior at large distances, but which becomes steeply repulsive at short distances (**Figure A1.2.1b**). This type of potential is seen in all diatomic molecules, and its general form is also seen in other types of interactions other than covalent.

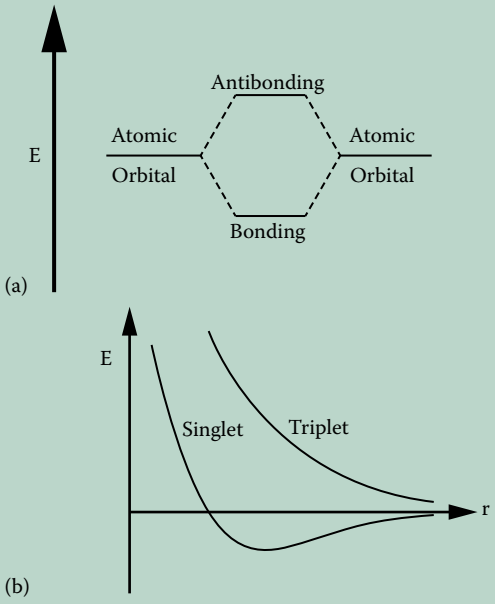


Figure A1.2.1 Molecular orbital energies for diatomic hydrogen. (a) Relative energies of atomic hydrogen and of the singlet state of molecular hydrogen (bonding) and triplet state (antibonding). (b) Energy versus distance for the singlet and triplet state of hydrogen using the Heitler–London model.

(*Continued*)

ADVANCED TOPIC 1.2 (CONTINUED): A QUANTUM MECHANICAL DESCRIPTION OF BONDING: MOLECULAR ORBITAL THEORY

Molecular orbital calculations become highly complex for molecules more sophisticated than hydrogen. The Hückel approximation, developed in 1931, lends itself to analytic solutions but is a poor approximation for most problems. More sophisticated quantum chemistry approaches rely upon modern computational power as well as upon development of appropriate approximation techniques; the 1998 Nobel Prize in Chemistry was awarded to John Pople "for his development of the density-functional theory" and to Walter Kohn "for his development of computational methods in quantum chemistry." Many software packages, both open-source and commercial, make use of a variety of approaches to solve these quantum chemistry problems. Density-functional theory is the least computationally intensive; the "functional" refers to the energy of the molecule as a function of electron density as a function of position. Another very common approach to molecular structure calculations is the *Hartree–Fock* approximation. Rather than electron density, Hartree–Fock looks at the wave function of the molecule as a product of the wave functions of each electron. Simplification of this tremendous task can be achieved using semiempirical methods (e.g., spectroscopy data) or ab initio approaches, which use mathematical simplifications to facilitate processing. Improvements to the Hartree–Fock approximation are called post-Hartree–Fock methods. In an advanced version of this course, one or more computational chemistry packages will be provided or suggested to you for the solution of test problems.

SUGGESTED READING

Albright, T.A., Burdett, J.K., and Whangbo, M.-H. *Orbital Interactions in Chemistry*. Edn. 2. Wiley-Interscience, 2013.

Fleming, I. *Molecular Orbitals and Organic Chemical Reactions.* Wiley, 2009 (student edition also available).

Kotz, J.C., Treichel, P.M., and Weaver, G.C. Bonding and molecular structure: Orbital hybridization and molecular orbitals. *Chemistry and Chemical Reactivity*. Thomson Brooks/Cole, Belmont, CA, 457–466, 2006.

electron. In the extreme case, the electron is almost entirely localized around this partner. Nearly all covalent bonds have some ionic character. An ionic bond can be described using the same quantum mechanical formulations as covalent bonds, with an alteration in the electrostatic term. The potential energy between two ions falls off as $1/r$.

Ion–dipole interactions

Most atoms do not have permanent dipoles, but many molecules do, meaning that there is an uneven distribution of charge along the molecule. In addition, both atoms and molecules can show induced dipole moments caused by exposure to an electric field (which may come from ions or dipolar molecules). Permanent dipole interactions are stronger than induced dipole interactions and we will consider these first. A *polar* molecule with a dipole moment m interacts with an ion of charge q with the potential

$$V(r) = -\frac{mq}{4\pi\varepsilon_0 r^2}\cos\theta, \tag{1.1}$$

where

r is the distance between them,

ε_0 is the permittivity of free space, and

θ is the angle of the dipole (**Figure 1.6a**).

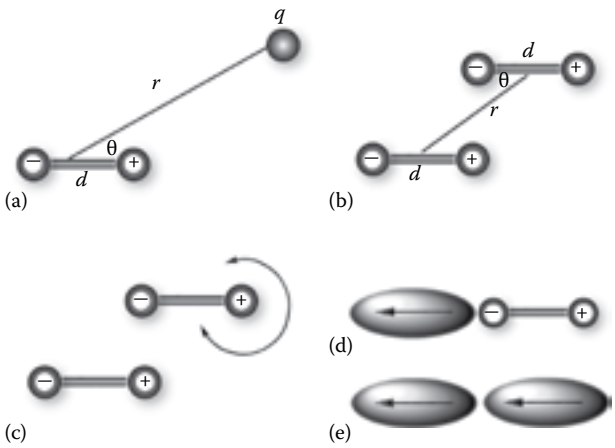


Figure 1.6 Intermolecular interactions. (a) A dipole–point charge interaction. When the two are far enough apart, d can be considered negligible relative to r. (b) Two parallel dipoles at a fixed angle. This sort of arrangement occurs in a solid. (c) If one or more of the dipoles is completely free to rotate, integration over all angles gives an interaction energy of 0. However, real molecules are limited in their rotations, so dipole–dipole interactions in a liquid or gas are nonzero. (d) A dipole can polarize a nonpolar molecule. (e) Two nonpolar molecules can have induced dipole moments and interact with each other.

If the dipole is free to rotate, thermal averaging results in a potential that falls off more rapidly with distance:

$$V(r) = -\frac{1}{3k_B T}\frac{(mq)^2}{(4\pi\varepsilon_0)^2 r^4}, \tag{1.2}$$

where k_B is the Boltzmann constant and T is the absolute temperature.

A key example of an ion–dipole interaction is the interaction between water and dissolved ions in solution. These relatively strong forces give rise to the *energy of hydration* of these ions, which needs to be overcome if an ion is to be separated from its surrounding water molecules. This energy is why most ions permeate through biological pores and channels in a hydrated state. (For polar solvents other than water, this can be generalized to an *energy of solvation*.)

Dipole–dipole interactions

The potential between two permanent dipoles m_1 and m_2 (assuming their own radii are negligible) falls off with distance more quickly than that between a dipole and a charge (**Figure 1.6b**):

$$V(r) = \frac{-2m_1 m_2}{4\pi\varepsilon\varepsilon_0 r^3} \tag{1.3}$$

This formula is valid for fixed dipoles, as in a solid. However, if the dipoles are completely free to rotate, their attractive and repulsive components cancel, and the net interaction is 0. An important concept is that in liquids and gases, rotation is not completely free but is weighted by the Boltzmann distribution (**Figure 1.6c**). Keesom showed that the average dipole–dipole interaction for rotating molecules at temperature kT is given by

$$V(r) = \frac{-2m_1^2 m_2^2}{3(4\pi\varepsilon_0)^2 k_B T r^6} \tag{1.4}$$

where

m_1 and m_2 are the dipole moments,

r is the distance between the dipoles,

ε_0 is the permittivity of free space,

k_B is the Boltzmann constant, and

T is absolute temperature.

This is the familiar form of the dipole–dipole interaction encountered in energy transfer experiments, e.g., fluorescence resonance energy transfer (FRET; see **Chapter 16**). It is called the *Keesom interaction* and is the strongest of the $1/r^6$ interactions. A derivation is given in **Advanced Topic 1.3**.

A single dipolar molecule can also induce an instantaneous dipole in a nonpolar molecule, with an induced dipole moment m_i related to the molecule's polarizability α and the applied electric field E as $m_i = \alpha E$. The average interaction for a dipole and a nonpolar molecule in a liquid or gas (**Figure 1.6d**) is

$$V = \frac{-m^2\alpha}{4\pi\varepsilon_0 r^6} \tag{1.5}$$

This formula does not need to be thermally averaged, because the direction of the induced dipole follows the permanent dipole; thermal fluctuations do not affect it. It is called the *Debye interaction* and is about 1/10 the strength of the Keesom interaction.

If both molecules are nonpolar, induced dipole–induced dipole interactions can still occur. These are called *London dispersion interactions* and, although weak, are responsible for the only possible interactions between nonpolar species such as noble gases. Their distance dependence is also $1/r^6$, and they are about 1/10 as strong as the Debye interaction (**Figure 1.6e**).

Collectively, these noncovalent interactions with $1/r^6$ dependence (Keesom, Debye, London) are called *van der Waals interactions* and may be parametrized by a single equation. The attractive $1/r^6$ potential is only valid at relatively long distances relative to the size of the molecule. To better describe what happens at short distances, a repulsive term must be added. The exact form can vary, but the Lennard-Jones potential is often used because it simply describes a very steep repulsion that occurs within a certain radius (called steric hindrance):

$$V = -\frac{A}{R^6} + \frac{B}{R^{12}} \tag{1.6}$$

Values of A and B are determined empirically or computationally for different molecules and can be found in journal articles and books. The arguments relating to dipoles also relate to higher multipoles (**Advanced Topic 1.4**).

Hydrogen bonds

Hydrogen bonds are a special case of fixed dipole–dipole interaction. A hydrogen attached to an electronegative atom (usually oxygen, fluorine, or nitrogen)

ADVANCED TOPIC 1.3: DERIVATION OF KEESOM INTERACTION

Derivation of the Keesom interaction is not done in most textbooks. The averaging is nontrivial. First, it must be noted that the averaging is not simply over all angles, which would give a result of 0; the averaging is Boltzmann-weighted, which gives greater weight to the lower-energy configurations.

To derive **Equation 1.4**, start from looking at the generalization of **Equation 1.3** for dipoles each free to rotate in the plane (θ_1, θ_2) as well as to twist relative to each other (φ). This expression becomes

$$V(r,\theta_1,\theta_2,\varphi)=\frac{-m_1m_2}{4\pi\varepsilon_0 r^3}\left[2\cos\theta_1\cos\theta_2-\sin\theta_1\sin\theta_2\cos\varphi\right].$$

This formula is a product of a radial part and an angular part

$$V(r,\theta_1,\theta_2,\varphi)\equiv V_0(r)f(\Omega),$$

where

$$V_0(r)=\frac{m_1m_2}{4\pi\varepsilon_0 r^3}.$$

Also define

$$\beta=\frac{-m_1m_2}{4\pi\varepsilon_0 r^3(k_BT)}.$$

This gives the Boltzmann-weighted average as

$$<V>=\frac{V_0(r)\int d\theta_1\,d\theta_2\,d\varphi\sin\theta_1\sin\theta_2 f(\Omega)e^{\beta f(\Omega)}}{\int d\theta_1\,d\theta_2\,d\varphi\sin\theta_1\sin\theta_2 e^{\beta f(\Omega)}},$$

where the integral over φ goes 0 to 2π and the integrals over θ go 0 to π. Now, using the properties of the natural logarithm, allows us to write

$$<V>=V_0(r)\frac{d}{d\beta}\ln\left[\int d\theta_1\,d\theta_2\,d\varphi\sin\theta_1\sin\theta_2 e^{\beta f(\Omega)}\right].$$

Since we make the assumption that $\beta f(\Omega)\ll 1$, we can expand the exponential as a Taylor series and keep only the first two terms:

$$e^{\beta f(\Omega)}\sim 1+\beta f(\Omega)+\frac{\left(\beta f(\Omega)\right)^2}{2}.$$

So the integrals we need to evaluate are

$$A=\int d\theta_1\,d\theta_2\,d\varphi\sin\theta_1\sin\theta_2=8\pi$$

(*Continued*)

ADVANCED TOPIC 1.3 (CONTINUED): DERIVATION OF KEESOM INTERACTION

$$B = \int d\theta_1\, d\theta_2\, d\varphi \sin\theta_1 \sin\theta_2 \beta f(\Omega)$$
$$= \int d\theta_1\, d\theta_2\, d\varphi \sin\theta_1 \sin\theta_2 [2\cos\theta_1 \cos\theta_2 - \sin\theta_1 \sin\theta_2 \cos\varphi] = 0$$

$$C = \frac{1}{2}\int d\theta_1\, d\theta_2\, d\varphi \sin\theta_1 \sin\theta_2 \left[\beta f(\Omega)\right]^2$$
$$= \frac{\beta^2}{2}\int d\theta_1\, d\theta_2\, d\varphi \sin\theta_1 \sin\theta_2 [2\cos\theta_1 \cos\theta_2 - \sin\theta_1 \sin\theta_2 \cos\varphi]^2 = \frac{8}{3}\pi\beta^2$$

This gives

$$< V > = V_0(r)\frac{d}{d\beta}\ln\left[8\pi\left(1+\frac{\beta^2}{3}\right)\right] = V_0(r)\left(1+\frac{\beta^2}{3}\right)^{-1}\frac{2}{3}\beta \sim V_0(r)\left(\frac{2}{3}\beta\right)$$
$$= \frac{-2(m_1 m_2)^2}{3(4\pi\varepsilon_0)^2 (k_B T) r^6},$$

which is the final result.

SUGGESTED READING

The original papers reporting this derivation are as follows:

Keesom, W. H., *Physik. Z.* 22, 129, 1921.
Keesom, W. H., *Physik. Z.* 22, 643, 1921.

develops an effective positive charge. This hydrogen is then called the donor or *proton donor*. It interacts with an electronegative atom that doesn't necessarily have any hydrogens and might be of a different species (**Figure 1.7a**). The strongest hydrogen bonds are between HF and F. Hydrogen bonds are crucial in biology: they hold double-stranded DNA into its helical structure and polypeptides into helical protein conformations, and assist in all varieties of specific binding interactions (antibody/antigen, receptor/ligand, DNA/transcription factor, etc.) (**Figure 1.7b**).

The (strept)avidin/biotin interaction

One of the strongest noncovalent interactions known is the interaction between *avidin* and *biotin*. Avidin is a protein produced in the oviducts of egg-laying animals and deposited into the egg white to protect the developing embryo from bacterial invasion by sequestering biotin, an essential vitamin in the B family that bacteria need to grow (mammals and birds also require biotin; it is possible to acquire a deficiency by eating raw egg whites). In its typical state, avidin is a tetramer in which each subunit binds biotin with equal affinity (**Figure 1.8a**). This gives an overall dissociation constant of ~10^{-15} M! A similar protein isolated from a bacterium is called *streptavidin* from its origin (*Streptomyces*), and shows

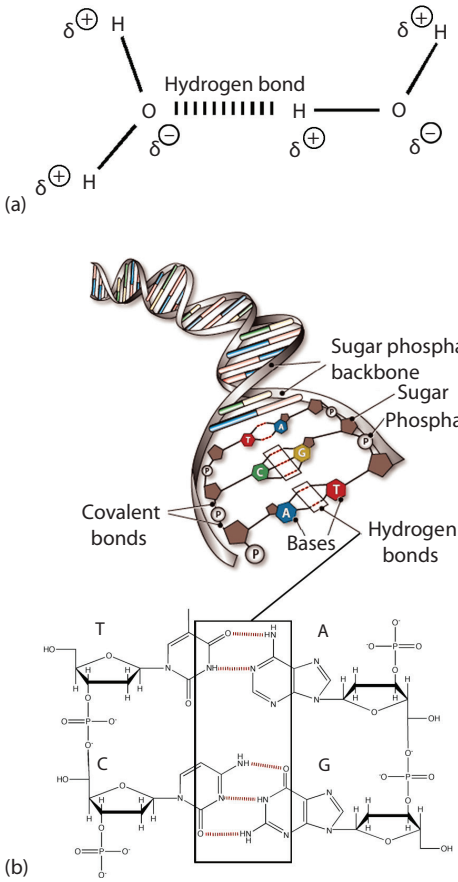


Figure 1.7 Hydrogen bonding. (a) Schematic of a hydrogen bond between two water molecules, where hydrogen acts as the proton donor. (b) Hydrogen bonds hold together the helices of the DNA double helix. Two hydrogen bonds link each A–T pair, and three link each G–C pair; no other combinations are possible. Important features of a DNA fragment, such as its melting temperature, are determined by the strength of these bonds. The upper image shows the hydrogen bonds between A–T and G–C in the context of the double helix; the lower inset shows the details of the structure.

similar chemistry but is easier to handle, with less aggregation and unwanted interactions with cells.

This interaction is exploited in biology for just about any type of experiment where specific, irreversible binding is desired. Biotin is a small molecule (molecular weight, 244.31) and may be chemically linked to any other molecule of choice, a process called *biotinylation*. Because it is so small, it very rarely affects the function of the molecule to which it is attached. The biotinylated molecule will then bind avidin or streptavidin specifically with great affinity.

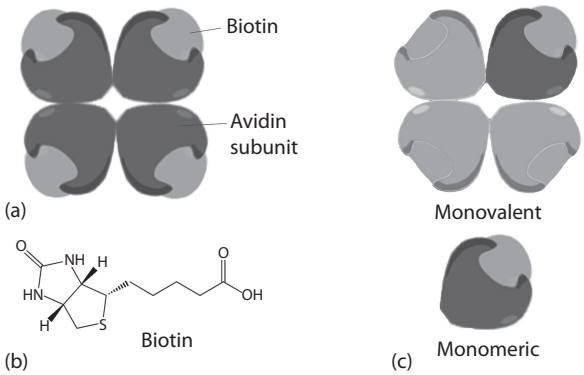


Figure 1.8 Avidin/biotin. (a) Schematic of tetrameric structure of avidin/streptavidin, with four binding sites for biotin. (b) Chemical structure of biotin. (c) Monovalent streptavidin has one active site for biotin and three blocked or inactive sites (the protein remains the same size). (d) Monomeric streptavidin is one-quarter the size of the full protein and has only one binding site in consequence.

ADVANCED TOPIC 1.4: CATION–PI, ANION–PI, AND PI–PI INTERACTIONS

Some molecules with no permanent dipole moment have a permanent quadrupole moment, and interactions of this moment with charges can be comparable in strength to that of a hydrogen bond. The most notable example is benzene, which has no dipole moment due to symmetry. However, the quantum mechanical delocalization of electrons above and below the plane of the molecule creates a quadrupole moment and a significant attractive potential to positive ions (**Figure A1.4.1a**). Because two amino acids (tryptophan and tyrosine) contain benzene rings, the *cation–pi* interaction has important implications for protein structure. This interaction is also believed to play an important role in the binding of the *neurotransmitter acetylcholine* to its receptor, and in the ability of nicotine to act as a full agonist for this receptor. (Acetylcholine and nicotine are pictured in **Figure A1.4.1b**.) More recently, a quadrupole–anion interaction was proposed, and potentially stabilizing anion–pi interactions were found in over 3000 proteins, the majority of the sample tested. Quadrupole–quadrupole interactions between aromatic rings result in *pi stacking*, or arrangement of the rings in energetically favorable conformations. Pi stacking is important in both nucleic acid and protein structure; it also is important for the concentration-dependent optical properties of many dyes in solution, since the majority of fluorescent compounds are aromatic.

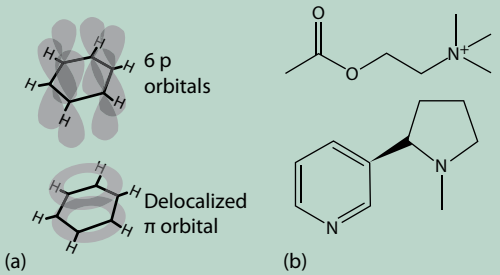


Figure A1.4.1 The cation–pi interaction. (a) Delocalized electron clouds around the benzene plane create a quadrupole moment. (b) Acetylcholine and nicotine.

SUGGESTED READING

Jones, G.J., Robertazzi, A., and Platts, J.A. (2013). Efficient and accurate theoretical methods to investigate anion–pi interactions in protein model structures. *The Journal of Physical Chemistry B* 117, 3315–3322.

Lovas, S., He, D.Z., Liu, H., Tang, J., Pecka, J.L., Hatfield, M.P., and Beisel, K.W. (2015). Glutamate transporter homolog-based model predicts that anion–pi interaction is the mechanism for the voltage-dependent response of prestin. *The Journal of Biological Chemistry* 290, 24326–24339.

Wilson, K.A., Wells, R.A., Abendong, M.N., Anderson, C.B., Kung, R.W., and Wetmore, S.D. (2016). Landscape of pi–pi and sugar–pi contacts in DNA–protein interactions. *Journal of Biomolecular Structure and Dynamics* 34, 184–200.

Zhao, Y., Li, J., Gu, H., Wei, D., Xu, Y.C., Fu, W., and Yu, Z. (2015). Conformational preferences of pi–pi stacking between ligand and protein, analysis derived from crystal structure data geometric preference of pi–pi interaction. *Interdisciplinary Sciences* 7, 211–220.

Zhong, W., Gallivan, J.P., Zhang, Y., Li, L., Lester, H.A., and Dougherty, D.A. (1998). From ab initio quantum mechanics to molecular neurobiology: A cation–pi binding site in the nicotinic receptor. *Proceedings of the National Academy of Sciences U S A* 95, 12088–12093.

Variations of avidin and streptavidin have been developed specifically for molecular biology. *Monovalent* streptavidin is still a tetramer but has only one active binding site for biotin. *Monomeric* streptavidin has been dissociated into its components (**Figure 1.8b**). The advantages of both are that they cannot cross-link in reactions. Monomeric streptavidin has the additional advantage of being only one-fourth as large as the tetramer, an advantage for some applications. The disadvantages are much weaker binding (dissociation constants, ~10^{-7} M). Other variants have been developed for pH-sensitive or reversible binding.

We will revisit biotin/streptavidin in many chapters of this book. The system is useful for labeling dyes (**Chapters 7 and 8**) and nanoparticles (**Chapters 11 through 13**), and for biofunctionalization (**Chapter 14**).

1.3 PRINCIPLES OF SPECTROSCOPY

What can be measured

The various types of *spectroscopy* used in molecular biophysics probe the interactions described previously, confirming the accuracy of the physical description and allowing for atomic-scale insight into bond lengths and angles, functional groups, molecular symmetry, and electronic structure. All of the methods listed here will be covered in more detail in **Chapter 16**; the following is an introduction to what they are capable of measuring.

Individual atoms and molecules have different quantized electronic energy states in which electrons are permitted to exist. *Transitions* to a higher or lower state require the absorbance or emission of a precise amount of energy. These energies (as well as lifetimes of excited states before they decay to lower-energy states) are characteristic of the atomic system in question. Complex molecules have other types of transitions as well, resulting from the spatial degrees of freedom of the atoms with respect to one another. These are rotations and vibrations about the bonds, giving rise to rotational and vibrational transitions (**Figure 1.9**).

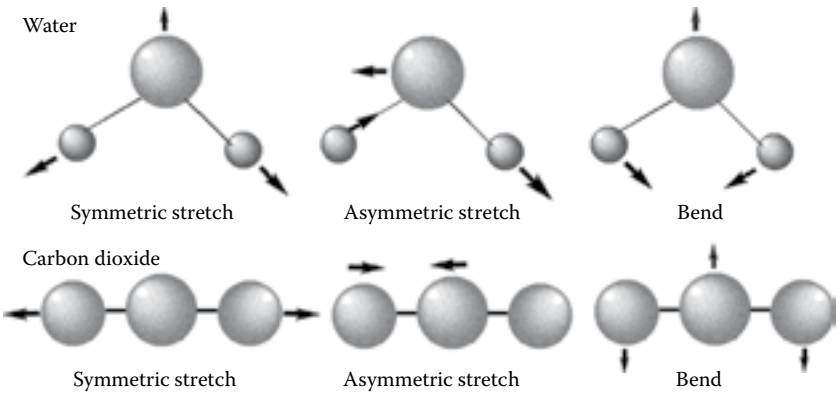


Figure 1.9 Vibrational modes of a polar triatomic molecule (water) and a nonpolar, linear triatomic molecule (carbon dioxide). Note that the symmetric stretching mode of CO_2 does not change the dipole moment, so it does not appear on an infrared spectrum.

The allowed rotational and vibrational states are quantized and can be estimated with quantum mechanical rigid rotor and harmonic oscillator approximations, respectively; we will return to this in **Chapter 16**.

How transitions are measured

When a molecule is exposed to electromagnetic radiation, its dipole interacts with the electric field, and energy can be transferred if the dipole moment's mode and the electric field have the same frequency and phase. If enough energy is transferred, a transition can occur. Calculation of the transition dipole moment between two states can predict which transitions will be experimentally observed; these selection rules will be revisited in more detail in **Chapter 16**.

The energy needed to excite electronic transitions is comparable to that of photons in the x-ray and ultraviolet–visible (UV–Vis) range. Vibrational transitions have energies comparable to that of infrared (IR) light. Finally, rotational transitions are the least energetic and can be measured using microwaves (**Figure 1.10**). Irradiating a sample to excite it and observing the resulting transitions is called spectroscopy, and the transitions are spectroscopic transitions. Some common forms of spectroscopy, which we will cover in **Chapter 16**, are as follows:

- *UV–Vis absorbance spectroscopy*, which provides characteristic signals for molecules that have electronic transitions within this energy range. These molecules are usually aromatic, and if the signal is in the visible, the solution will appear colored. Metals also provide useful signals in this range.
- *Fluorescence spectroscopy* is the complement of absorption spectroscopy as it measures transitions from the excited state to the ground state. Absorbance of a visible or UV photon can excite a molecule to a higher electronic level; relaxation to the ground state causes the emission of a less energetic photon. The photons emitted can vary slightly in energy as relaxation can occur into different vibrational levels of the ground

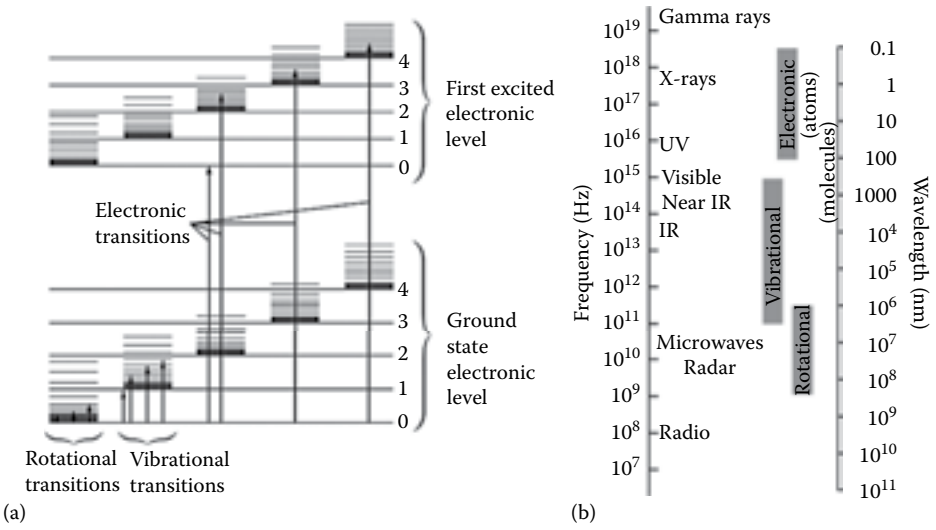


Figure 1.10 Energies of electronic, vibrational, and rotational transitions (a) relative to each other and (b) relative to the electromagnetic spectrum.

electronic state; this can be used to probe vibrational energies. Many organic molecules are fluorescent in the UV or visible, especially those with aromatic rings.

- Any vibrational modes of the molecule that result in a change in the permanent dipole moment are IR active. (It can be inferred from **Figure 1.9** that water has a very complex IR spectrum.) The energies of these transitions are characteristic for each type of covalent bond. Thus, *IR spectroscopy* can be used to give a "fingerprint" of the types of bonds present in the sample. It can be used to measure very complex materials. *IR fluorescence* can also be used to probe vibrational levels.

- When an intramolecular bond breaks, the result may be an unpaired electron, known as a *free radical*. Free radicals are highly reactive, and specific species, such as oxygen radicals, have been implicated in cell death, cancer, and aging. They are also of interest because when exposed to microwave radiation, an unpaired electron can move between parallel and antiparallel states in a magnetic field. This is the basis for *electron paramagnetic resonance* (EPR) spectroscopy.

1.4 CELLS

Life can be thought of as a controlled series of reactions in which the building block molecules are taken up, assembled, recycled, and broken down in precisely controlled order and proportion. The overall process is called metabolism, where building-up processes are anabolism and breaking-down ones are catabolism. A lack of key elements or disruption in the balance of assembly and breakdown leads to death. The first requirement for life is thus that it be enclosed within some kind of container to allow for nonequilibrium concentrations of key elements. The second is that this container must be able to assimilate all of the needed building blocks. Hence the *cell*, the unit upon which all Earth life is based. A cell is a water-impermeable lipid container bearing a series of passive and active (i.e., energy-dependent) transport systems for import and export of key molecules.

A single cell placed in a surrounding "infinite" sea with a certain nutrient concentration will be limited in the extent of its reactions by the rate of diffusion of nutrients to its surface. The poorer the sea, the smaller the cell must be to meet its requirements. In ocean water, calculations have shown this limit to be approximately a cubic micron.

Once inside the cell, the nutrients also must diffuse to specific sites to undergo reactions. This limits the size of the cell unless it develops subcellular containers dedicated to specific metabolic processes. We can distinguish two distinct domains of life: those that do not have these subcellular compartments (called *organelles*, including a *nucleus*) and those that do. The former are called Prokarya, and the latter, Eukarya. The Prokarya include the kingdoms Bacteria and Archaea, and the Eukarya are Animalia, Plantae, and Fungi.

An example of a prokaryotic cell is molecular biology's key organism, the bacterium *Escherichia coli*, usually just called *E. coli*. The structure of an *E. coli* cell is shown in **Figure 1.11a,b**; it consists principally of a *cell wall*, *cell membrane*, circular DNA *chromosome* packaged into a *nucleoid* region, and *ribosomes*. Within its fluid contents or *cytoplasm* are all of the enzymes needed to replicate

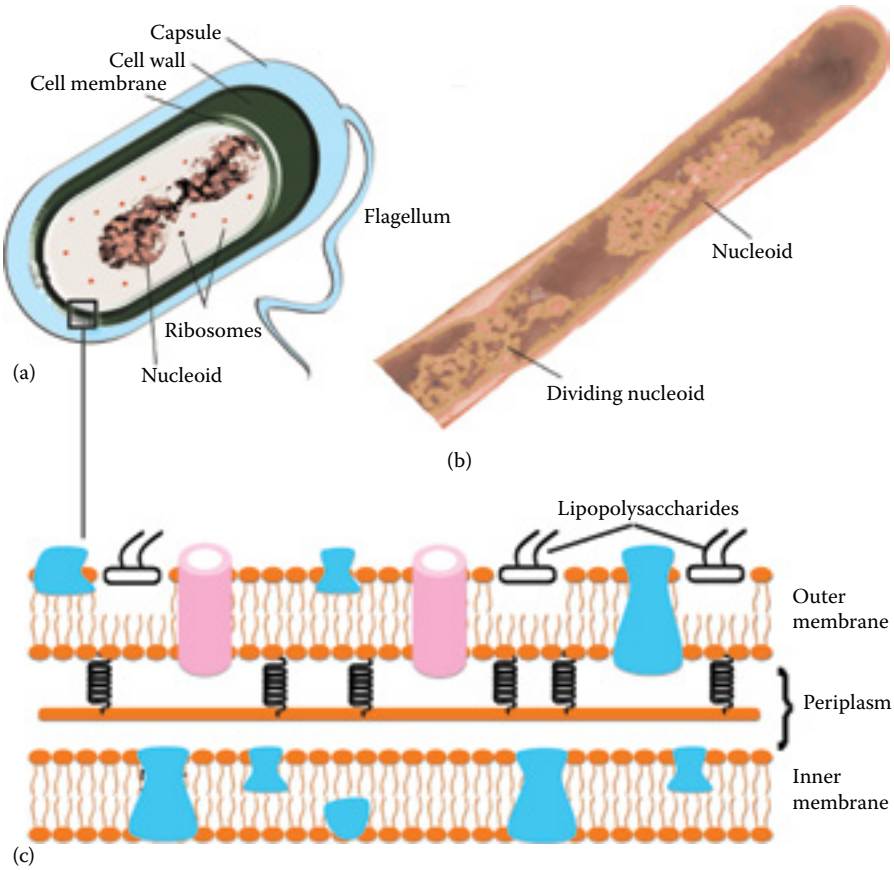


Figure 1.11 Structure of a typical prokaryote, the bacterium *E. coli*. (a) Schematic of cross-section through the cell. (b) Transmission electron microscopic image (TEM) of dividing cell. (c) Schematic of inner membrane, periplasm (or periplasmic space), and outer membrane.

DNA and *transcribe* DNA to RNA. RNA is *translated* to protein in the ribosomes. The cytoplasm also contains disequilibrium concentrations of ions, particularly potassium, leading to a nonzero *membrane potential* (see **Chapter 15** for more on the origin and measurement of membrane potentials). The space between the membrane and cell wall is called the *periplasm* and may comprise 40% of the cell's volume; many important reactions, such as neutralization of *antibiotics*, occur in this space. The structure of the *E. coli* membrane and periplasm is characteristic of the class of bacteria called *Gram negative* (**Figure 1.11c**).

A eukaryotic cell is much larger and more complex. Plant and fungal cells are surrounded by rigid cell walls, whereas animal cells have only a lipid membrane. This significantly changes the way in which the different types of cells are handled in the laboratory. Introducing a foreign agent through a rigid cell wall is challenging, and so we will not deal with culture or transfection of fungal or plant cells in this book. Animal cells used for molecular biophysics include those from normal human tissue or from human cancers and rat and mouse cells. These cells are usually immortalized into *cell lines*, usually by infecting with viral agents. This causes the cells to divide essentially forever, so that the experimenter can simply maintain a flask of cells and extract some as needed for each experiment. The cells are seeded onto the substrate of choice and allowed to grow to the desired density before the experiment takes place. Many different cell lines are available from commercial suppliers, along with specialized media that permit their growth. A cell line is chosen based upon the desired experiment: some cell lines simply

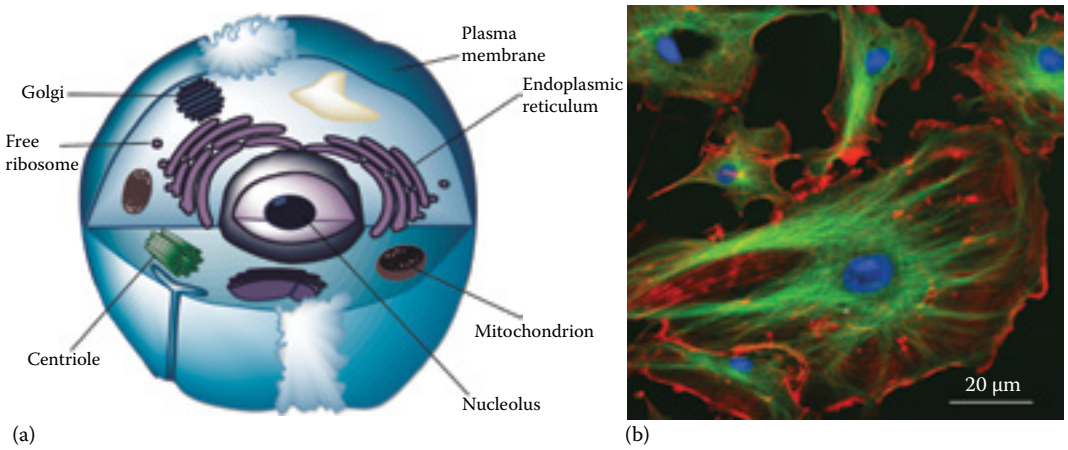


Figure 1.12 Animal cells. (a) Schematic of a typical animal cell showing organelles. (b) Light micrograph of cultured cell line showing labeled cytoskeleton (microtubules in green, actin in red) and nuclei (blue).

express foreign DNA well; others express specific receptors or ion channels; some are models for disease processes, especially cancer. The alternative to cell lines is *primary cells*, which are taken directly from an animal and cultured in a dish. This requires obtaining ethical approval for sacrifice of the animals as well as the expertise needed to extract, isolate, and culture the cells. Depending upon the cell type, this can range from easy (e.g., blood or immune cells) to very difficult (e.g., specific neuronal populations). Primary cells are more difficult to manipulate than cell lines, as will be discussed in **Chapter 3**. Several textbooks on cell lines and culture of specific types of primary cells are available.

A typical animal cell is shown in **Figure 1.12**. Organelles include the nucleus, *mitochondria*, *lysosomes*, *Golgi apparatus*, and smooth and rough *endoplasmic reticulum* (ER). Ribosomes may exist free or bound to the ER (which is what makes the rough ER "rough" under electron microscopy). The function of these organelles will be discussed in the following sections. Eukaryotic cells are also characterized by a complex *cytoskeleton* of *microfilaments*, *intermediate filaments*, and *microtubules*. These molecules are often studied in molecular biophysics.

1.5 DNA, RNA, REPLICATION, AND TRANSCRIPTION

The structure and function of DNA and RNA

In cells, nucleic acids are used to encode information on when, where, and in what quantities proteins and other nucleic acids will be made. Their chemistry is what makes this possible. DNA usually exists in cells in a stable double-helical form with a hydrophilic sugar–phosphate backbone surrounding the hydrogen-bonded nitrogenous bases. The phosphate terminus is called the *5′ end* of each strand, and the hydroxyl terminus is the *3′ end*. The double helix is *antiparallel*: that is, what is considered the top strand runs 5′ to 3′, and the bottom strand runs 3′ to 5′. A only bonds to T and G only to C; these are complements (**Figure 1.13a**). If this helix is separated (*denatured* or melted), free nucleotides that are complementary to the single strand can hydrogen-bond to the single-stranded DNA to regenerate the double helix. Thus, the molecule encodes its own copy

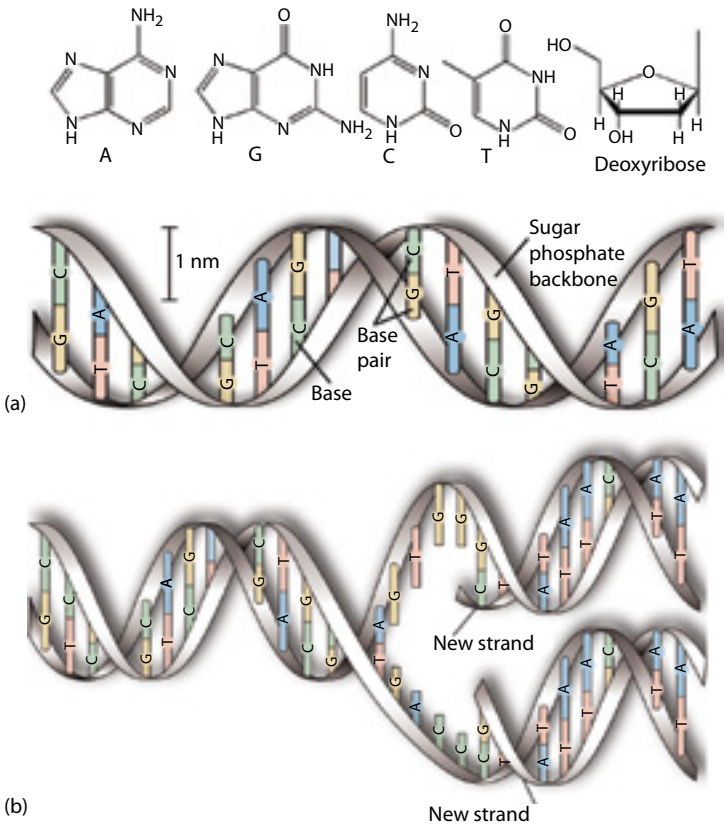


Figure 1.13 Structure and replication of DNA. (a) DNA bases and deoxyribose. (b) Principles of DNA replication. Complementary strands form from each strand of denatured DNA. Each of the resulting copies is then made up of one old strand and one new strand. This is called *semiconservative* replication.

(**Figure 1.13b**). The process is thermodynamically favorable and is accelerated in cells by enzymes that *catalyze* the reactions. Of course, it implies that there must be a certain concentration of free nucleotides available whenever DNA needs to be synthesized. In prokaryotic cells, these are found throughout the cytoplasm, whereas they are compartmentalized in eukaryotic cells according to the need for them. Both prokaryotes and eukaryotes take free nucleotides up from the environment (ingest or eat them); organ meats, seafood, and legumes are good dietary sources. Duplication of DNA is called replication.

RNA is similar to DNA except that each of its sugars contains an extra oxygen (ribose instead of deoxyribose) and it contains the base uracil in the place of thymine, which lacks a carbon (**Figure 1.14a**). The extra oxygen makes RNA less stable, as it is more susceptible to hydrolysis. In every creature on Earth except some viruses, RNA does not encode the genome but serves as a messenger between the stable DNA chromosome and downstream applications. RNA nucleotides can hydrogen-bond to single-stranded DNA in the same way as DNA nucleotides can bind to each other, and thereby, the information contained in DNA is transmitted (**Figure 1.14b**). The process of making RNA from DNA is called transcription, and it relies upon free RNA nucleotides being present. Because RNA is single-stranded, it can have sequences that are self-complementary, causing it to fold back in on itself and form very complex *secondary structures*. These structures are important for many of RNA's roles, including transcription; RNA molecules can also act as enzymes and even catalyze their own replication (**Figure 1.14c**).

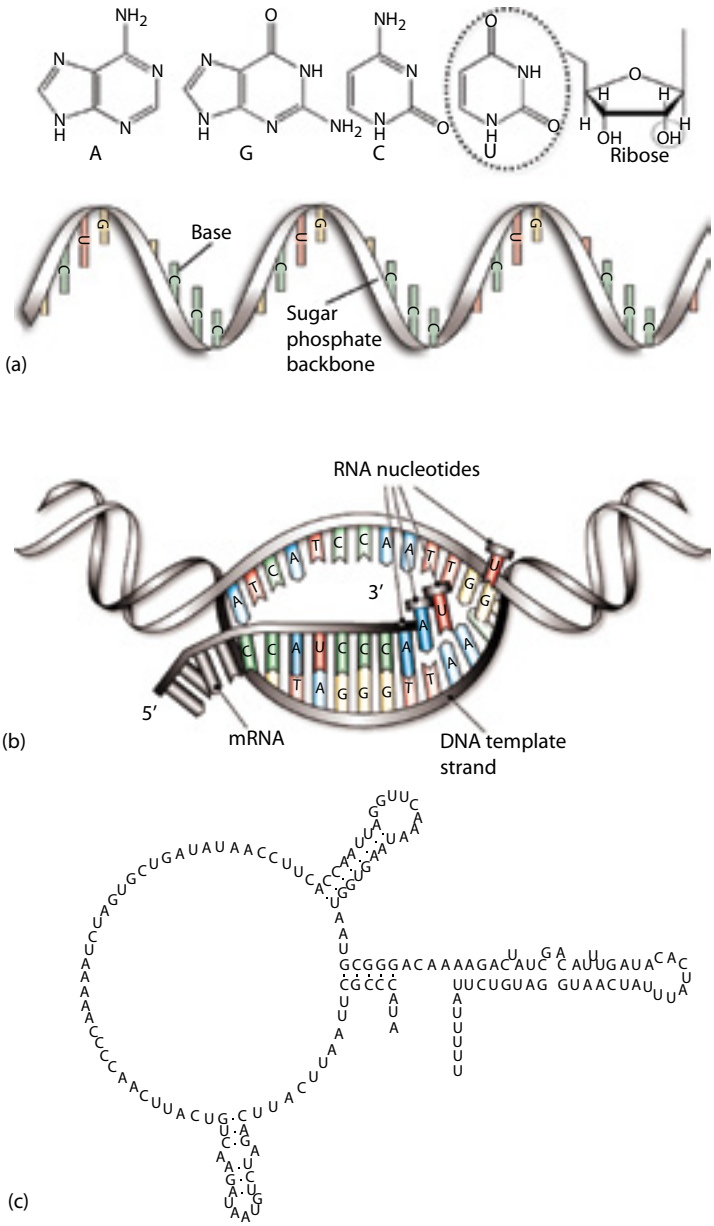


Figure 1.14 Structure and functions of RNA. (a) RNA bases and ribose. The circles show the differences with respect to DNA. (b) Principle of transcription. RNA bases can hybridize with single-stranded DNA, creating a single-stranded RNA messenger (messenger RNA [mRNA]) that carries the information. (c) Secondary structure. Since RNA is single-stranded, its self-complementary sequences can bind to each other, creating loops and hairpins. This feature allows RNA to perform roles such as catalysis that DNA cannot. (This image is of a telomerase RNA.)

Replication

Enzymes serve as *catalysts* to increase the rate of reactions and the local concentration of needed ingredients. Without enzymes, DNA replication and transcription would occur too slowly and haphazardly to be practical for life as we know it. The enzymes involved in DNA processing have been identified, purified, and studied at the atomic level in many organisms, particularly in bacteria, and the physical chemistry of many of the reactions has been elucidated in detail. The protein structure of the enzymes, combined with the physics and chemistry specific to given DNA sequences, allows for careful regulation of when and how

DNA will be replicated, transcribed into RNA, or otherwise modified. We will give an overview here that is focused on prokaryotes (bacteria). Many of the same principles apply in eukaryotes, though the processes are often significantly more complex.

In *E. coli*, there are at least 30 proteins involved in replication. We will see many of these enzymes in later chapters as their purified forms play important roles in molecular biology. For example, *topoisomerase* uncoils DNA from its stably packaged *supercoiled* form. *DNA helicase* breaks open hydrogen bonds to unzip the DNA. *Single-strand binding proteins* (SSBs) stabilize the DNA in a single-stranded configuration so that it does not rewind while transcription is taking place or self-bind as RNA can do. After complementary free nucleotides bind, DNA polymerase joins them to each other with phosphodiester bonds.

Additional complications require even more enzymes. *DNA polymerase* only works in the 5′ to 3′ direction. This means that the parent strand with the free 3′ end elongates continuously; it is called the *leading strand*. The strand with the free 5′ end, called the *lagging strand*, is made in small fragments, which are joined later by *DNA ligase*. In addition, DNA polymerase can only add a base to an existing strand; it cannot start from zero. Thus, a primer (usually made of RNA) and a *DNA primase* are required to start replication. The primers are later removed by *endonucleases*. Replication is shown schematically in **Figure 1.15a**.

Replication occurs in both directions along the DNA strand from a location called a *replication fork*. Multiple replication forks usually occur in an *E. coli* chromosome at the same time. This allows the entire 4.7-megabase genome to be replicated in 20 min even though the rate of replication is only 1000 base pairs/s. **Figure 1.15b** shows a schematic of replication forks along with a micrograph of replication in a bacterium.

DNA replication is extremely accurate thanks to the system of complementary hydrogen bonding, but it is not perfect. Polymerases have proofreading mechanisms

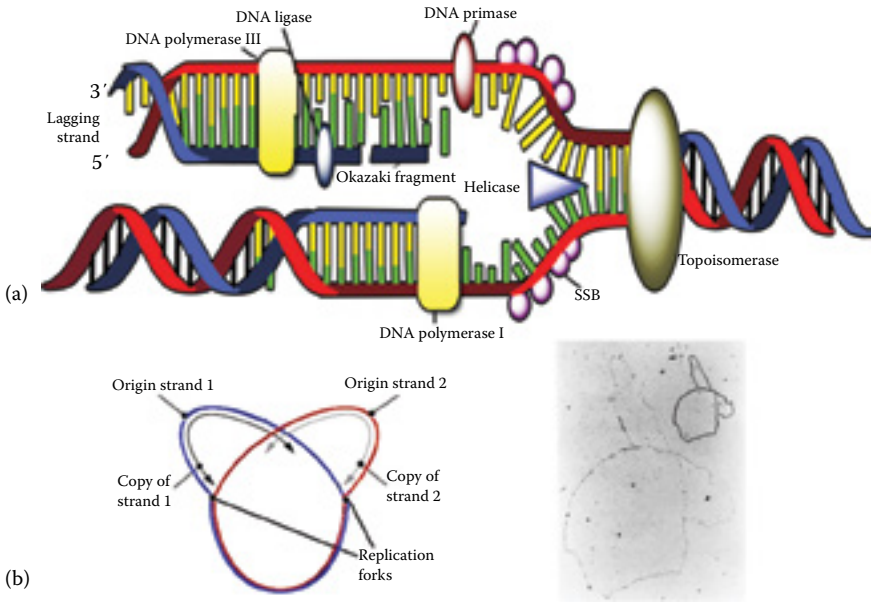


Figure 1.15 DNA replication in *E. coli*. (a) Schematic of replication, showing some of the enzymes involved and the synthesis of short fragments on the lagging strand versus continuous synthesis on the leading strand. (b) Schematic and electron micrograph of bidirectional replication of circular bacterial chromosome. Each strand has its own origin, and synthesis proceeds bidirectionally at each origin.

that allow them to trim off misplaced bases; this is called 3′ → 5′ *exonuclease* activity and is an important feature of these enzymes that is used in molecular biology. The proofreading ability of DNA polymerase I, operating on the leading strand, is about 20-fold better than that of DNA polymerase III on the lagging strand. Nonetheless, *E. coli* can replicate its genome with only about one error every 10 million base pairs. These errors are called *mutations*, and an organism's mutation rate is crucial for its ability to survive. It shouldn't be too high, or too many individuals will be nonviable mutants, but it must not be 0, or the organism would not be able to evolve.

Transcription

The generation of a single-stranded RNA molecule from its DNA complement is called transcription. One generally speaks of transcription of *genes*, since as a general rule, each sequence of DNA that encodes for a specific protein is called a gene. (However, the original idea of "one gene, one protein" is an oversimplification, as a single gene can encode for multiple proteins.) The parts of DNA that are directly transcribed into complementary RNA are called coding regions. Noncoding regions include *promoters*, *enhancers*, *repressors*, and many others, some still being discovered. (Identifying all of the *regulatory regions* for specific genes is a major challenge in the biology of eukaryotes.) Regulatory regions are usually upstream of the coding region, or toward the 5′ end of the DNA molecule, and are involved in controlling the activity of coding regions.

A promoter is a key regulatory region found in every gene; it is what permits the enzyme complex involved in transcription to recognize the gene and bind to it. In *E. coli*, there is only one enzyme that transcribes DNA, called *RNA polymerase*. It binds nonspecifically and weakly to DNA, effectively confining its diffusion to one dimension along the organism's chromosome. When it encounters a promoter, it binds tightly (**Figure 1.16a**). The equilibrium binding constant determines the strength of the promoter and can range from 10^6/M for weak promoters to 10^9/M for strong promoters. The first step of transcription is when the promoter DNA and the polymerase form a *closed complex*, in which the DNA is still double-stranded. However, transcription cannot take place until the DNA is made single-stranded so that RNA nucleotides can bind. This occurs in an isomerization reaction, in which the closed complex becomes the *open complex*. During this reaction, a short sequence (12–17 base pairs) of the DNA just upstream to the gene is opened or melted. This open fragment is called the *transcription bubble* (**Figure 1.16b**). Along with DNA melting, the polymerase undergoes a large conformational change during this reaction, developing a hole or pocket allowing the DNA template strand to move inside. The exact details of this reaction remain largely a mystery and may be different for different promoters. The formation of the open complex is the rate-limiting step in the initiation of transcription; rate constants for formation range from 10^{-3}/s for weak promoters to 0.1/s for strong promoters.

The binding of free RNA nucleotides can now begin. However, they will be limited to fragments 7–9 base pairs in length (called abortive transcripts) unless the RNA polymerase manages to dissociate from the promoter and bind to the nonspecific DNA downstream. This step is called *promoter clearance* and requires a protein to displace the strong polymerase–promoter binding subunit σ. Transcription has now been initiated, and elongation has begun (**Figure 1.16c**). The RNA transcribed from the template strand is called *messenger RNA*, or mRNA, because it serves to relay the information encoded in the genome to other areas of the cell.

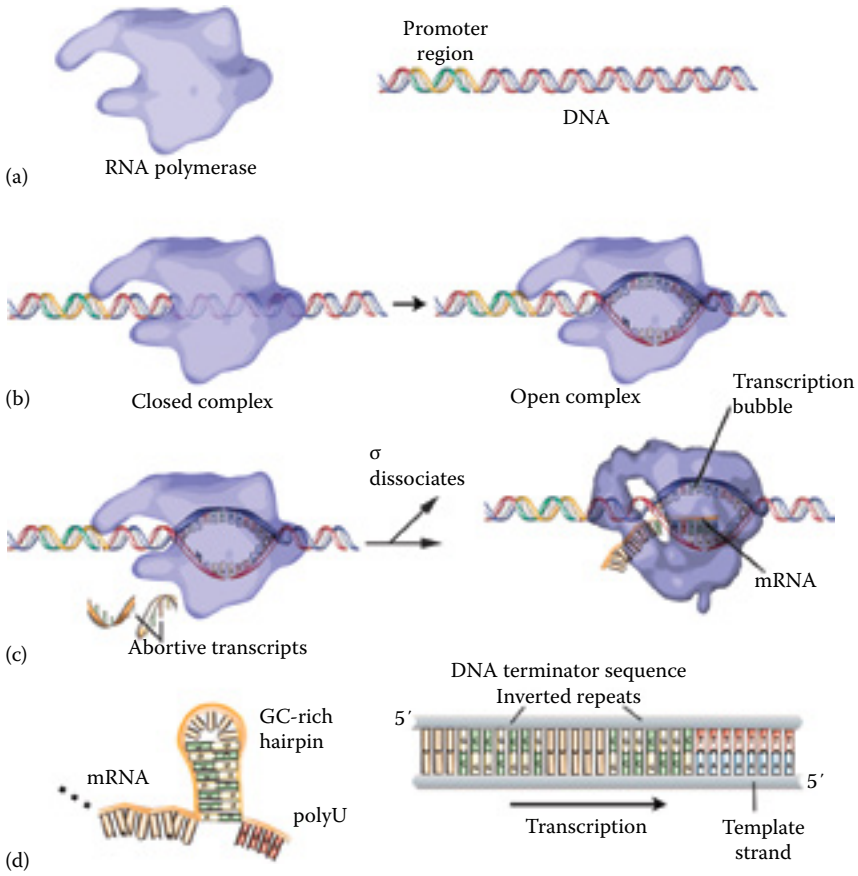


Figure 1.16 DNA transcription in *E. coli*. The images of the polymerase are schematics inspired by electron crystallography data of the protein. (a) The first step involves the RNA polymerase binding strongly to a fragment of DNA containing a promoter. A bacterial promoter sequence is shown. (b) The rate-limiting step is the melting of a 12- to 17-base-pair *bubble* to form the open complex. (c) As long as the polymerase is bound to the promoter, only short abortive transcripts can be made. For promoter clearance, the sigma subunit of the polymerase must dissociate, and the enzyme must clear the promoter. (d) Termination occurs at a specific DNA sequence (shown) that leads to a hairpin RNA with a poly-U tail. (Red indicates T or U, blue represents A, yellow represents G, green represents C, and orange indicates any base.)

Recent studies have found that mRNA elongation is not a boring, steady process but instead very complex, with different types of paused complexes.

Termination of transcription is assisted by the chemistry of RNA, which allows it to form three-dimensional secondary structures. A *terminator* is a sequence on the DNA that contains two sequences that are inverted repeats, followed by a *poly-A* (multiple adenines). When the RNA is transcribed from the repeats, it bends into a hairpin, probably causing the polymerase to pause. The poly-A creates the weakest possible DNA–RNA hybrid because there are only two hydrogen bonds per base. The combination of these factors causes the mRNA to dissociate from the DNA, and transcription is finished (**Figure 1.16d**).

1.6 TRANSLATION AND THE GENETIC CODE

After the synthesis of mRNA is complete, the single strand travels to the ribosomes to be translated into protein. The mRNA is read in units of 3 base pairs, each of which is a *codon* that encodes for a single specific amino acid. At each codon, a *transfer RNA* (tRNA) binds to the mRNA at one end, while bound to the specific amino acid at the other end. The successive tRNAs form a polypeptide chain that emerges from the ribosome (**Figure 1.17**).

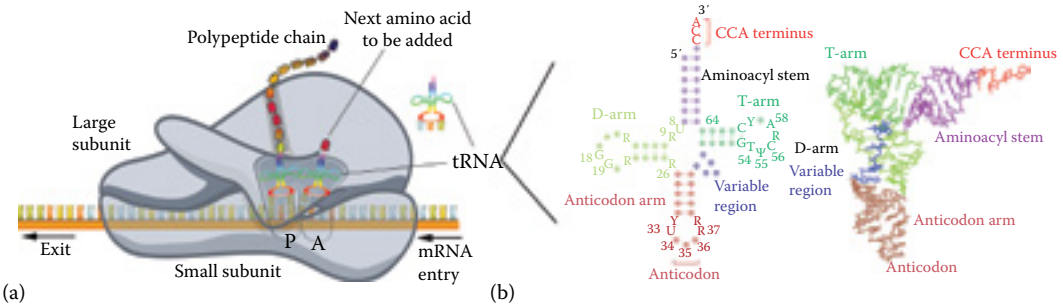


Figure 1.17 Schematic of translation in *E. coli*. (a) The ribosome is made up of a large subunit and a small subunit. The *A site* binds aminoacyl-tRNA (tRNA with an amino acid attached); the *P site* binds peptidyl-tRNA (tRNA that's bound to the nascent peptide chain), with an exit tunnel for the new protein to emerge. There is also an *E site* for tRNA preparing to exit, which is not shown here. The mRNA moves through the ribosome being translated one codon at a time. It is possible for more than one ribosome to bind a single mRNA. The tRNA features an anticodon at one end, which is the complementary sequence to the codon in the mRNA, and an attachment site for an amino acid on the other end. (b) Detailed structure of tRNA. (From Hori, H., Methylated Nucleosides in tRNA and tRNA Methyltransferases, *Front. Genet.* 23 May 2014, http://dx.doi.org/10.3389/fgene.2014.00144.)

There are several features of this process that are key in molecular biology and play major roles in the design of most experiments. The first is the genetic code. Because there are 4 unique RNA bases, the number of combinations of n bases is 2^n; thus, 3 is the smallest number that can encode for the 20 amino acids. However, this creates a lot of degeneracy, as there are now 64 available codons for the 20 amino acids. Three codons are *stop* or termination codons, leaving 61 for the 20 amino acids (**Table 1.2**).

Table 1.2
The Genetic Code

	Second Nucleotide				
First Nucleotide	**U**	**C**	**A**	**G**	**Third Nucleotide**
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Note: The amino acid abbreviations are given in **Figure 1.2**.

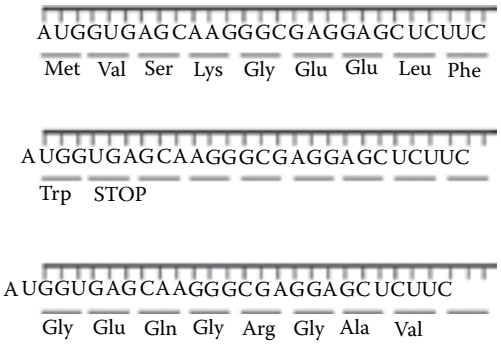


Figure 1.18 Reading frames. Any double-stranded DNA has, in principle, six reading frames (three on the top strand, three on the bottom). The three frames of this sample of mRNA encode very different peptides; the second one terminates after a single amino acid.

This degeneracy helps to protect organisms from harmful mutations. Of the 64 codons, 32 require only the first 2 base pairs to be specified, and thus are fourfold degenerate; these correspond to the most common amino acids, such as alanine and glycine. Thus, no mutation in the third base would have any effect on these codons. Mutations are discussed further in **Chapter 2**, as they can be used as a valuable tool in the manipulation of DNA. They are also important in determining molecular mechanisms of disease, as mutations leading to large changes in hydrophobicity or charge of a single amino acid can have disastrous consequences to a protein and often to the entire living organism.

The degeneracy of the code is also important when designing genes for expression in different systems. Nearly all organisms use the same genetic code; however, eukaryotes often prefer a different codon than prokaryotes for the same amino acid. Thus, a sequence that uses the bacterial codon would be expressed correctly in mammalian cells, but possibly at a low level. To increase expression levels, silent mutations must be inserted to optimize the codon for the expression system desired. This process is called *codon optimization* and needs to be considered any time genes are being expressed in cells very different from their species of origin.

Another important concept in translation is that of the *reading frame*. Since every three bases code for an amino acid, the exact position of the start point determines the frame in which the sequence is read. A single insertion or deletion will change every subsequent amino acid (**Figure 1.18**). The translation of mRNA does not begin at the beginning of the molecule; most mRNAs contain a 5' untranslated region (5' UTR). Instead, an AUG codon (which encodes the amino acid methionine) signals the start of the protein and determines the reading frame. Any additions to the protein must then be in frame with the start codon in order to be read correctly. We will return to this in more detail in **Chapter 2**.

It is important to note that tRNAs are small molecules and thus under the control of the resourceful chemist (**Advanced Topic 1.5**).

1.7 PROTEIN FOLDING AND TRAFFICKING

The ribosome produces a one-dimensional polypeptide chain that is biologically inactive. This is referred to as the protein's *primary structure* (**Figure 1.19a**). The *secondary structure* results from hydrogen bonding and creates typical forms such as *alpha helices* and *beta sheets* (**Figure 1.19b**). The *tertiary structure* is the final

ADVANCED TOPIC 1.5: UNNATURAL AMINO ACIDS

A tRNA bearing a normal codon can be coupled to something that is not an amino acid at all, or that is some sort of variation of an amino acid (caged, fluorescently tagged, etc.). This is called an *unnatural amino acid* and can be used to study the role of single amino acids in proteins. Many unnatural amino acids are available commercially (**Figure A1.5.1**).

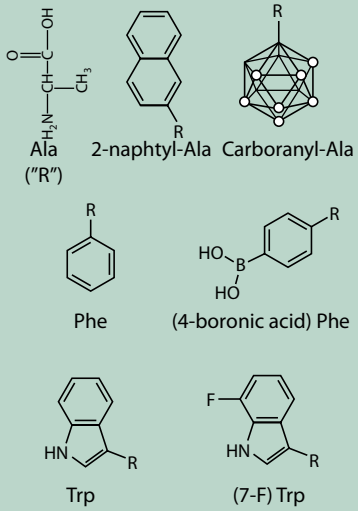


Figure A1.5.1 Some commercially available unnatural amino acids that are derivatives of alanine (Ala), phenylalanine (Phe), or tryptophan (Trp).

SUGGESTED READING

Adumeau, P., Sharma, S.K., Brent, C., and Zeglis, B.M. (2016). Site-specifically labeled immunoconjugates for molecular imaging—Part 2: Peptide tags and unnatural amino acids. *Molecular Imaging and Biology* 18(2), 153–165.

Brown, K.A., and Deiters, A. (2015). Genetic code expansion of mammalian cells with unnatural amino acids. *Current Protocols in Chemical Biology* 7, 187–199.

Dippel, A.B., Olenginski, G.M., Maurici, N., Liskov, M.T., Brewer, S.H., and Phillips-Piro, C.M. (2016). Probing the effectiveness of spectroscopic reporter unnatural amino acids: A structural study. *Acta Crystallographica Section D, Structural Biology* 72, 121–130.

Hino, N., Hayashi, A., Sakamoto, K., and Yokoyama, S. (2006). Site-specific incorporation of non-natural amino acids into proteins in mammalian cells with an expanded genetic code. *Nature Protocols* 1, 2957–2962.

Lang, K., and Chin, J.W. (2014). Cellular incorporation of unnatural amino acids and bioorthogonal labeling of proteins. *Chemical Reviews* 114, 4764–4806.

Lang, K., Davis, L., and Chin, J.W. (2015). Genetic encoding of unnatural amino acids for labeling proteins. *Methods in Molecular Biology* 1266, 217–228.

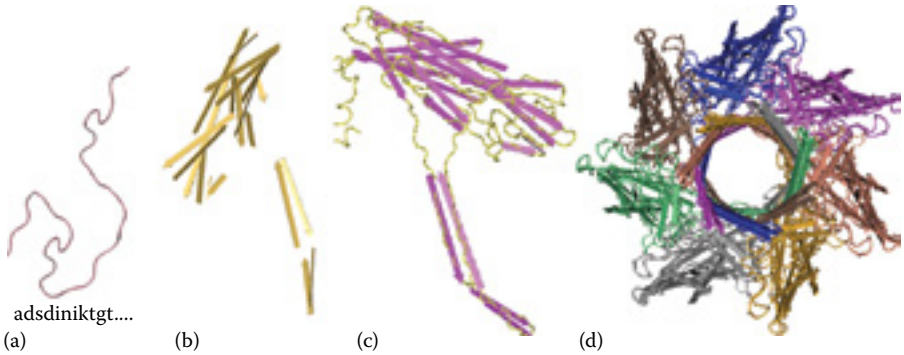


Figure 1.19 Protein structure. The example protein shown is alpha-hemolysin, a pore-forming protein from the bacterium *Staphylococcus aureus*. (a) The primary structure is an amino acid sequence, or a polypeptide chain. (b) The secondary structure consists of helices, sheets, and other forms made by hydrogen bonding. (c) The tertiary structure includes all interactions and leads to a specific *native* conformation under the right conditions. (d) Some proteins have quaternary structure. Alpha-hemolysin's final form consists of seven subunits identical to the structure in (c). (It is a homoheptamer.) Each of the colors represents an independent subunit.

three-dimensional form and results from all of the intermolecular interactions among the amino acid residues in the context of the correct temperature, pH, and ionic concentration (**Figure 1.19c**). A correctly folded protein is called the *native* conformation. Some proteins have a *quaternary structure*, which refers to the assembly of multiple subunits that fold independently (**Figure 1.19d**).

How proteins fold is an extremely complex subject, and many sophisticated resources exist on the topic, of which we reference a few at the end of this chapter. It was originally believed that all of the information for the tertiary structure was contained within the primary structure, but this turned out only to be the case for relatively small, soluble proteins. More complex proteins require the assistance of enzymes, called *chaperones*, to fold properly and reach their final destination without aggregating. Chaperones assist proteins in finding their native conformation, assist refolding of misfolded proteins, and bind to the hydrophobic surfaces of proteins in order to prevent aggregation. Other types of *posttranslational processing* of the polypeptide chain may also be necessary to produce an active protein. These include proteolytic cleavage (enzymes called *proteases* trim the ends of the polypeptide chain, or cut it into smaller pieces) and chemical modifications in which new chemical groups are added to specific amino acid residues. Some chemical modifications, such as *phosphorylation* (addition of a phosphate group), are simple and are performed by all organisms. More complex modifications such as *glycosylation* (addition of a large carbohydrate side chain) are only performed in eukaryotic cells.

In *E. coli*, chaperones are located in the cytoplasm, whereas in eukaryotic cells, folding takes place in a specialized membranous organelle called the endoplasmic reticulum. Despite this major difference, many types of eukaryotic proteins can be successfully produced in large amounts in *E. coli* and show normal folding and biological activity; these include human insulin (the first recombinant DNA pharmaceutical), bovine growth hormone (BGH), and many others.

However, many proteins fail to fold properly in *E. coli*, and attempts to optimize expression can be frustrating and fruitless. Reasons for failure include the following:

- The expressed protein contains many cysteine residues. The cytoplasm of *E. coli* is reducing, so disulfide bonds cannot form there. Thus, a protein that requires the formation of one or more disulfide bonds to attain its final tertiary structure may not fold properly.
- The protein has a complex structure with many turns. The formation of turns is a rate-limiting step, and rapid overproduction of complex proteins may lead to aggregation.
- The protein has a need for glycosylation or other complex posttranslational processing.

Failure to fold properly usually results in insoluble aggregates containing misfolded proteins and chaperones. The aggregates are called *inclusion bodies* and can be half as large as a bacterial cell. There are several possible approaches to avoiding inclusion body formation, including simple incubation at lower temperatures or producing lower amounts of protein; these will be discussed more fully in **Chapter 4**. It is also possible to purify protein from inclusion bodies and refold it, although this approach is not often successful. In many cases, *E. coli* cannot be used to express the protein of interest, and a eukaryotic expression system must be used. The most common of these systems are yeast cells and insect cells, and their cultivation and use will be covered in **Chapter 3**. Even in this case, posttranslational modifications may not be identical to those in mammalian systems, so proteins for therapeutic use may still be incorrectly expressed. Mammalian cell systems exist for expression and purification of these proteins, but these are costly and inefficient compared to the other culture types mentioned and should be used as a last resort.

The delivery of proteins to their sites of activity is known as *trafficking*. This is an immensely complex subject, especially in specialized eukaryotic cells such as neurons. More than 100 human diseases are known or believed to be the result of errors in protein trafficking, including some forms of cystic fibrosis, osteogenesis imperfecta, and Alzheimer disease. New proteins and mechanisms involved in trafficking are being continually discovered, many using the techniques of single-molecule biophysics, which we will discuss in later chapters. Here we will mention only the key concepts that have important implications for the design of molecular biology experiments. In *E. coli*, which has no intracellular compartments, trafficking is relatively simple: proteins either are cytoplasmic or are secreted into the periplasmic space or outside the cell. In order to be excreted, the proteins usually do not fold until they have passed through the cell membrane (although there are exceptions to this). In order to prevent premature folding, the proteins are tagged with a signal sequence that marks them as to be excreted. A chaperone protein binds the signal sequence and prevents premature folding, and the polypeptide is secreted through specialized protein pores. The chaperone is then cleaved, and the protein folds (**Figure 1.20**). Numerous signal sequences have been identified in *E. coli* and can be used in molecular biology experiments to tag proteins for excretion.

In eukaryotes, the presence of complex intracellular compartments necessarily makes targeting much more complex. An intracellular organelle, the Golgi apparatus or Golgi complex, is responsible for targeting of proteins to the

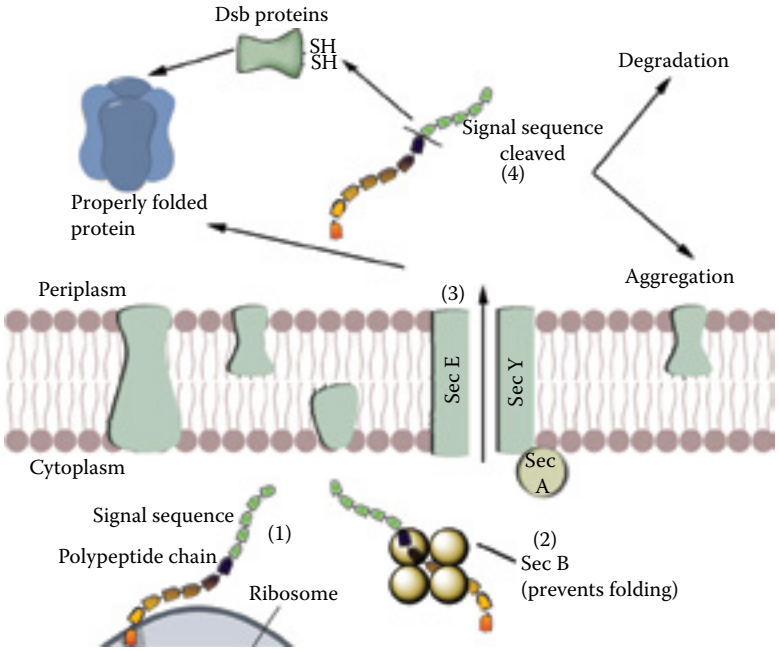


Figure 1.20 Protein trafficking in *E. coli*. A protein tagged for secretion will have a specific peptide sequence on one end as it emerges from the ribosome (1). It binds the protein SecB (2), which prevents premature folding, and is excreted through a membrane pore made of other proteins SecA, SecY, and SecE (3). Once in the periplasm, the signal sequence is cleaved (4). The protein can then fold correctly. If disulfide bonds must be formed, they are assisted by the dsb proteins. Incorrectly folded proteins aggregate and/or are degraded.

membrane or outside the cell via the *secretory pathway*. The proteins arrive at one face of the Golgi (called the *cis* face) in vesicles after translation in the ER, are processed, and then bud in vesicles from the other side (the *trans* face). Trafficking to other regions, such as the nucleus and mitochondria, is controlled via organelle-specific targeting sequences (**Figure 1.21**).

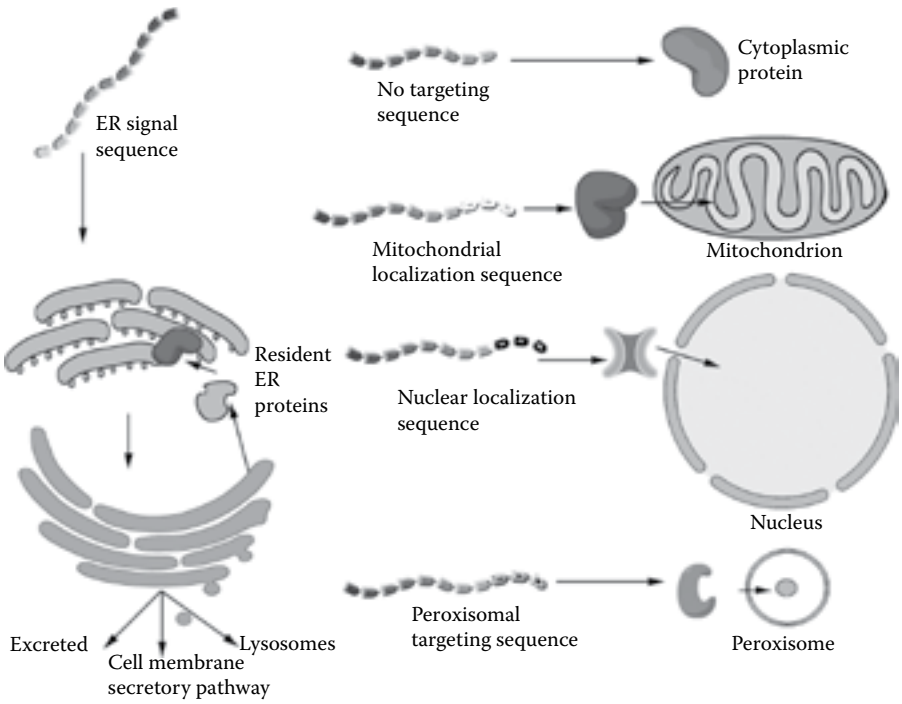


Figure 1.21 Protein trafficking in an animal cell. Proteins destined for the endoplasmic reticulum (ER) or the secretory pathway are tagged with an ER signal sequence. The *resident ER proteins* (needed for ER function) remain in the ER while the others pass to the Golgi. In the Golgi, they are processed further for the lysosomal, membrane, or excretion pathways. Multiple other types of tags allow proteins to target organelles without passing through the secretory pathway: there are nuclear, mitochondrial, and peroxisomal targeting sequences. Of course, many proteins are untagged and remain in the cytoplasm.

ADVANCED TOPIC 1.6: MOLECULAR MECHANICS

Molecular mechanics is a classical mechanical approach to modeling macromolecules. It creates a force field by summing all of the possible covalent and noncovalent potential energies in a system, where noncovalent includes electrostatic and van der Waals interactions. The simulation must include not only the molecule itself but also the solvent—usually water, though lipid membranes are often commonly needed as well. Empirical data are usually used to input parameters such as bond lengths. Molecular mechanics is usually used in the field of molecular dynamics, or simulations of the motion of macromolecules, including protein folding and conformational changes involved in protein function. The disadvantage of this technique is the immense computational power required; usually, only very short timescales can reasonably be modeled, short even on the timescale of folding or ligand binding. Many of the quantum chemistry software packages also contain molecular mechanics tools, though many do not; dedicated packages are also available.

SUGGESTED READING

Haile, J.M. *Molecular Dynamics Simulation: Elementary Methods*. Wiley-Interscience, 1997.
Leach, A. *Molecular Modelling: Principles and Applications*. Edn. 2. Prentice Hall, 2001.

The major implications of trafficking for molecular biology are as follows:

- Mammalian proteins that are targeted to the Golgi or other organelles are likely not to be properly processed or folded in *E. coli*.
- One solution is to target these proteins to the periplasm via a targeting sequence. Many times, this allows membrane proteins to fold properly.
- If this does not work, eukaryotic and possibly mammalian expression systems will be needed to produce these proteins.
- The expression of recombinant proteins may be inhibited even in mammalian systems if the designed protein inhibits processing in the ER or Golgi apparatus. For example, adding GFP to membrane proteins often results in a product that does not escape from the Golgi, hence never reaching its target and causing stress to the cell.
- Targeting sequences can be added to genes that do not otherwise have them, in order to obtain a desired result such as an extracellular protein or a specific organelle label, or in order to study the function of a protein in a particular organelle.

1.8 ALTERNATIVE GENETICS

Nearly all forms of known life have their genomic information stored in DNA. RNA acts as a messenger from DNA to protein. The few exceptions are found among the viruses, which have very small genomes—as small as 5 kilobase pairs (kb), compared to the very smallest bacterial genome of nearly 500 kb. The so-called *RNA viruses* use RNA exclusively as their genetic material; this RNA may be single-stranded or double-stranded. Single-stranded RNA viruses may be *positive sense*, meaning that they are transcribed directly into protein by the host cell, or *negative*

sense, meaning they are first turned into complementary RNA by a viral enzyme and then transcribed by the host cell. This viral enzyme is an *RNA-dependent RNA polymerase*, meaning a polymerase that makes RNA from RNA. Some of the most dangerous human and animal pathogens are single-stranded RNA viruses: Ebola virus, influenza virus, Lassa fever virus, rabies virus, measles, and mumps are all negative sense; poliovirus, hepatitis A and E, yellow fever, West Nile, and some varieties of common cold viruses are all positive sense.

RNA viruses that use DNA intermediates—that is, turn their RNA genomes into DNA—are called *retroviruses*, and the RNA-dependent DNA polymerase is called *reverse transcriptase*. Once the genome has been turned into double-stranded DNA inside a host cell, it integrates into the host genome, where it is propagated and transcribed with the host cells. It is estimated that 5–8% of the human genome is made up of endogenous retroviral sequences—that is, sequences that have been propagated for a long time after some ancestral infection. The most commonly known pathogenic retrovirus is the human immunodeficiency virus (HIV), the cause of acquired immune deficiency syndrome (AIDS). The use of HIV-based viral vectors as delivery vehicles for genes is described in detail in **Chapter 3**.

Unlike the DNA polymerases found in more complex organisms, RNA-dependent RNA polymerase and reverse transcriptase lack proofreading mechanisms. This accounts for the rapid mutation rates of RNA viruses and retroviruses and hence the difficulty of developing effective vaccines and treatments. The mutation rate is also increased in drug-resistant forms of reverse transcriptase, making future drug failures more likely. These features also influence the way these enzymes must be used in the laboratory; reverse transcriptase can be expected to make approximately 1 error every 2000 base pairs.

1.9 WHAT IS CLONING?

For the purpose of this book, we use the term *cloning* to refer not to the cloning of organisms but to molecular cloning. The terms *gene cloning*, *DNA cloning*, *molecular cloning*, and *recombinant DNA* all refer to the same process: the transfer of a DNA fragment of interest from one organism to a self-replicating genetic element, where it is then propagated in a foreign host cell. Like cloning of whole organisms, the output is a series of exact copies of the original. The *host* is nearly always *E. coli*, which is readily grown in the laboratory in large quantities. Over time, a multitude of *cloning strains* of *E. coli* have been developed, designed to take up foreign DNA readily and to replicate it at controlled numbers of copies. These strains are available commercially, and their attributes will be discussed in detail in **Chapter 2**.

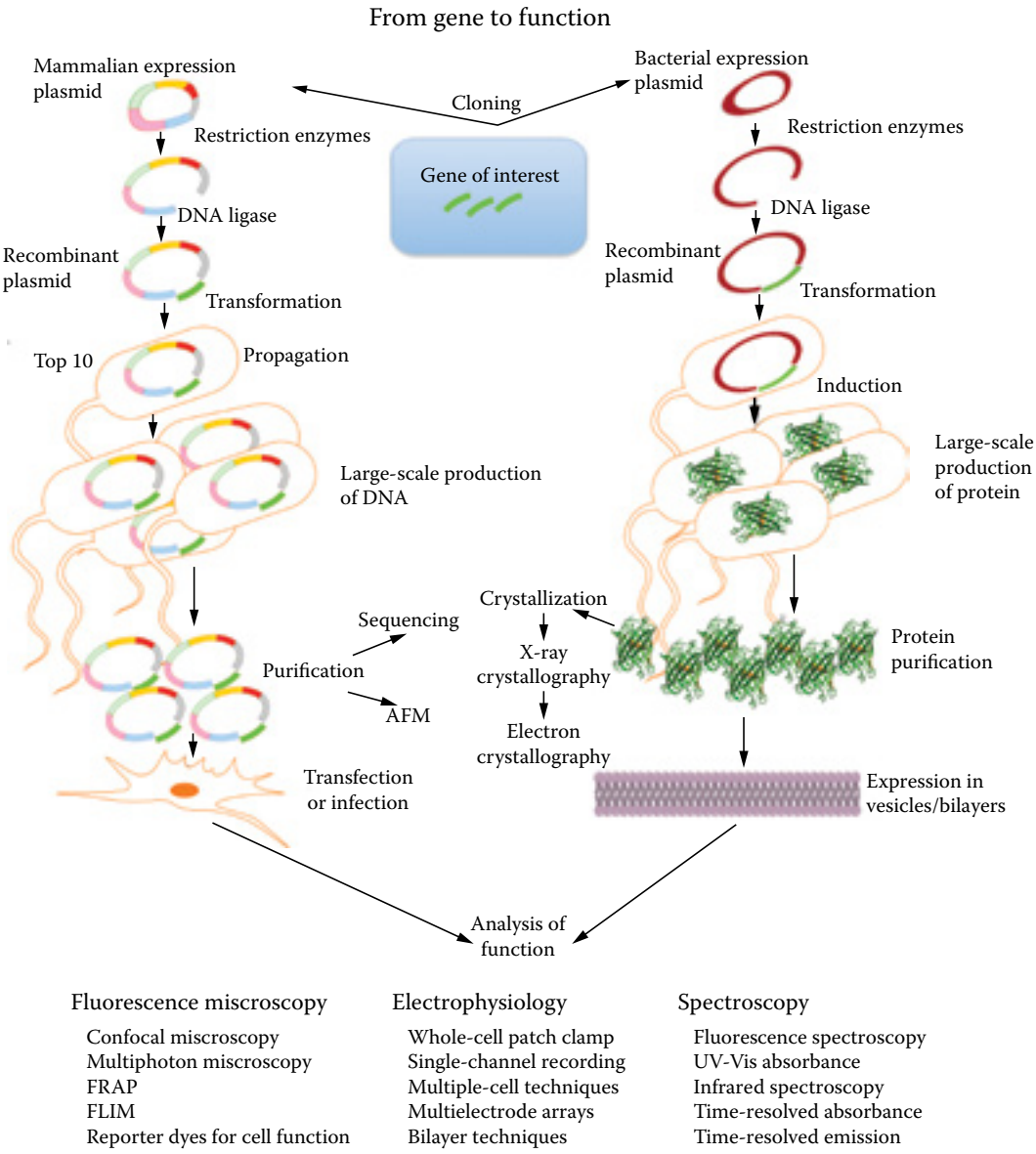
Purified forms of the enzymes involved in DNA and RNA replication, transcription, ligation, and other reactions are also available from commercial suppliers. Molecular biology consists of using these enzymes along with purified nucleotides, nucleosides, and other building blocks in order to amplify and manipulate DNA. Manipulations may be done inside *E. coli* cells or *in vitro* (i.e., in a test tube outside a living organism).

1.10 DESIGN OF A MOLECULAR BIOLOGY EXPERIMENT AND HOW TO USE THIS BOOK

The design of a molecular biology experiment depends upon the final downstream application or applications (**Scheme 1.1**). The questions to ask are as follows:

- What is the DNA sequence I want to clone?
- What modifications do I want to make to it?
- What is the desired final output (a DNA sequence, cell line expression, primary cell expression, or purified protein)?
- How much of the final output do I need? For example, a crystallization experiment requires at least several milligrams of highly purified protein. On the other hand, a cellular transfection experiment requires only micrograms of DNA, and the production and trafficking of the protein are all done by the cells. The total amount of protein made in this case is very small, however, so transfection of animal cell lines should not be used when purified protein is desired.
- How pure does it need to be? If it needs to be pure, how can I separate it from undesired contaminants?
- How will I confirm the desired output?

In the following chapters, we will proceed through the techniques needed at each step, beginning with obtaining and modifying DNA sequences and proceeding to different methods of expression, purification, and characterization. The chapters are intended to stand alone for each method, so that you can choose which ones to consult based upon your application. **Table 1.3** shows some examples of start-to-finish experiments and their corresponding chapters.



Scheme 1.1 Design of a molecular cloning experiment based upon desired outcome and downstream analysis techniques. A fragment of DNA is first inserted into either a mammalian (left) or a bacterial (right) expression vector. In the case of a mammalian vector, the DNA is amplified in *E. coli*. This DNA is purified, and may be sequenced or used in other experiments such as atomic force microscopy. It is then used to transfect mammalian cells, which may be subjected to many of the forms of microscopy, physiology, or spectroscopy listed at the bottom of the scheme. A bacterial expression vector is used when protein is to be purified (unless the protein of interest does not express properly in bacteria; see **Chapter 3**). Bacterial expression vectors can generate many milligrams of protein, which can be purified for spectroscopy or crystallography. Electrophysiological techniques can also be applied to pure protein in lipid bilayers, which ensures that only the protein of interest is being studied.

Table 1.3

Guide to Use of This Book

Desired Application(s)	Chapters to Read
Clone a gene, purify the protein, perform UV-Vis and/or infrared spectroscopy	**2**, **5**, **16**
Clone a gene, purify the protein, crystallize, perform x-ray or electron crystallography	**2**, **5**, **6**
Clone a very difficult gene, quantify its expression	**2**, **4**
Clone a gene, express in cell line, examine by fluorescence microscopy (with additional labels)	**2**, **3**, **7**–**8** [**11**–**13**]
Clone a gene, express in cell line, examine by electron microscopy (with additional labels)	**2**, **3**, [**11**–**13**], **16**
Clone a gene, express in cell line, examine by holographic microscopy	**2**, **3**, **9**
Clone a gene, purify the protein, express it in a lipid bilayer, perform electrophysiology	**2**, **5**, **15**
Clone a gene, express in cell line, perform electrophysiology (with additional labels)	**2**, **3**, **15** [**7**, **11**–**13**]
Clone a gene, purify the protein, biofunctionalize a surface, characterize the surface	**2**, **5**, **14**
Clone a gene, express in cell line, examine cellular toxicity	**2**, **3**, **10**
Clone a gene, express in cell line, pattern cells on microfabricated surface	**2**, **3**, **17**

End-of-Chapter Questions and Problems

Biochemistry

1. Where does *valence* come from? Can you tell the valence of an element by looking at its position in the periodic table? What are the valences of the following elements: H, He, O, K, Ar, Cl, Fe?

2. Look up the structural formulas of the following compounds, draw them, and identify their functional groups: cholesterol, vitamin E, dopamine, amphetamine, and aspartame. What are their alternative names, if any?

3. The concept of *acidity constant* K_a or pK_a comes from the equilibrium between the associated form of an acid (HA) and its dissociated form (A^-) in water:

$$HA + H_2O \Leftrightarrow H_3O^+ + A^-$$

$$K_a = \frac{[H_3O^+][A^-]}{[HA]}$$

$$pK_a = -\log K_a.$$

pH is simply a special case of the hydronium ion:

$$pH = -\log[H_3O^+].$$

Using the values of pK_a given in **Figure 1.2**, answer the following questions for glutamic acid, histidine, arginine, and lysine. (a) Which of these amino acids is the most acidic? (b) What percentage of each of these amino acids would be in its dissociated form at pH 7.4 (physiological pH)?

4. Draw the following functional groups and discuss their importance: amino, hydroxyl, carboxyl, and phosphate.

5. Match the following (more than one may apply!):

____Triacylglycerols

____Phospholipids

____Sphingomyelins

____Cholesterol

____Glycogen

____Olestra

a. Nonpolar lipids
b. Found in fat cells
c. Look like steroids
d. Have a glycerol backbone
e. Polar lipids
f. Glucose storage in animal cells
g. Insulate neuronal axons
h. Alter membrane fluidity
i. Found in biological membranes
j. Artificial fat used as a diet food additive

6. Draw and label a segment of lipid bilayer membrane showing the fatty acids, cholesterol, and membrane proteins.

Energies and Potentials

7. What is the speed of a molecule of nitrogen (N_2) at room temperature?

8. Derive **Equation 1.1** starting from Coulomb's law.

9. Derive **Equation 1.2**.

10. Polarizability of a nonpolar molecule may be modeled as a spherically symmetric electron cloud surrounding a positively charged nucleus. An externally applied electric field causes a shift in the electron cloud, resulting in an internal field developing at the nucleus to oppose it. Derive a formula for polarizability as a function of electron cloud radius, *a*, based upon this model. If you apply this formula to water, with an assumed radius of ~0.1 nm, what value do you obtain? Compare with the literature value of 1.66×10^{-40} mC^2/N.

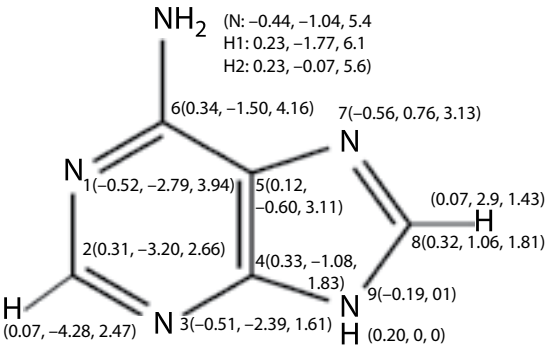
11. Which of the following molecules have a permanent dipole moment? Quadrupole moment?

a. H_2O
b. CO_2
c. CH_4
d. N_2
e. CO
f. NH_3

12. Contrary to most other substances, the density of water decreases as temperature decreases between 4°C and 0°C, and then even further upon freezing. Why? How much does the density change between 0°C and 4°C, and upon freezing?

13. Calculate the dipole moment of water given that the HOH angle is 104.5°, OH length is 0.98 Å, and partial charges are +0.4 on H and –0.8 on O. Express your answer in debyes (D).

14. Take a close look at one of the DNA bases, adenine, and calculate its dipole moment in debyes. Assume the molecule is planar. Then, given the partial charges and *x*–*y* coordinates in the picture below (given as *q* in esu, *x*, *y* in angstroms), calculate the dipoles in the *x* and *y* direction and the total moment. Draw an arrow showing the direction of the moment vector. Compare with the measured value of 3.0 D. (ADVANCED: Use molecular modeling software to calculate these partial charges.)



15. Calculate the electrostatic potential, $V(r)$, for two water molecules oriented (a) in a configuration permitting hydrogen bonding and (b) in a configuration unfavorable to hydrogen bonding, given that the HOH angle is 104.5°, OH length is 0.96 Å, partial charges are +0.4 on H and –0.8 on O, and hydrogen bond length is 2.5 Å.

16. (ADVANCED) Use molecular modeling software to predict the hydrogen bonds that can occur (a) between adenine and thymine and (b) between guanine and cytosine. Which of these actually occur in DNA?

17. What is the most stable conformation for pi–pi stacking? In which conformation is the interaction repulsive? (ADVANCED: Use modeling software to quantify.)

Spectroscopy

18. To study crystal diffraction, neutrons of energy approximately 0.33 eV are required. What is the corresponding energy of a photon of the same wavelength? An electron? Is it safe to treat the neutrons nonrelativistically?

19. (a) Show that a hydrogen atom has energy levels $E_n = -E_0/n^2$ where $E_0 = 13.6$ eV. (b) Calculate the wavelengths of the spectral lines corresponding to the 2→1 transition, the 3→1 transition, and the 3→2 transition. Which is the one corresponding to the common telescope filter?

Cells

20. Estimate the volume of a prokaryotic cell (radius, 1 μm) and of a eukaryotic cell (radius, 10 μm).

21. Estimate the minimum size of a complete living cell containing a genome of 1 megabase (each nucleotide pair is 0.34 nm long) and 15,000 ribosomes each 20 nm in diameter. How is the DNA packaged to fit inside the cell?

22. Estimate the maximum size of a bacterial cell. Hints: for a sphere of radius R, the flux across the surface is $Q = 4\pi DRC$, where D is the diffusion coefficient and C is the external concentration of the diffusing molecule; a reasonable metabolic rate is about 2 W/kg; the energy yield of glucose is about 260 kcal/mol; the diffusion coefficient of glucose is 6.7×10^{-10} m^2/s; and the external concentration is 50 mM. Does your answer seem reasonable? What factors could have made it too high or too low? (ADVANCED: Derive the formula for Q.)

23. If a cell of radius r divides into two cells with the interior volume being conserved, how much new membrane has to be made?

DNA, RNA, Proteins

24. Estimate how long it would take to replicate the *E. coli* genome from a single replication point. Explain how *E. coli* can divide every 20 min.

25. (a) Estimate how many mRNAs can be made in 20 min if the rate-limiting step is the formation of the open complex (RP_o) from the closed complex (RP_c). The rate equation is

$$[\text{RNAp}] + [\text{P}] \underset{k_{-1}}{\overset{k_1}{\Leftrightarrow}} [\text{RP}_c] \xrightarrow{k_2} [\text{RP}_o],$$

where [RNAp] and [P] are the concentrations of RNA polymerase and promoters, respectively. First, show that

$$[\text{RP}_o] = [\text{P}_{tot}](1 - \exp^{-k_{obs}t})$$

where

$$k_{obs} = \frac{k_1 k_2 [\text{RNAp}]}{k_1[\text{RNAp}] + k_{-1} + k_2} \equiv \frac{k_2[\text{RNAp}]}{[\text{RNAp}] + k_B^{-1}}$$

and

$$\text{P}_{tot} = [\text{P}] + [\text{RP}_c] + [\text{RP}_o] .$$

Now use values of $k_2 = 0.04$/s, $k_B = 10^7$/M, and [RNAp] = 30 nM to get a numerical result. Is this consistent with the replication rate?

(b) Discuss how you would distinguish experimentally whether k_{-1} or k_2 serves as the rate-limiting parameter.

(c) If the half-life of an mRNA molecule is 3 min before it is degraded by enzymes, what is the approximate equilibrium number of mRNAs in the cell?

26. Estimate the average translation rate (codons/s) in a cell if a cell is about 30% by weight protein and divides every 20 min. Then assuming 15,000 ribosomes, what is the translation rate per ribosome?

27. How many different polypeptides of 50 amino acids could, in principle, be produced? How

many different secondary structures could be produced from a given sequence, if each secondary domain is 10 amino acids long and can be either an alpha helix or beta sheet?

28. Match the following:

a. Replication fork
b. Origin of replication
c. Lagging strand
d. Leading strand
e. DNA helicase
f. Single-strand binding protein
g. DNA gyrase
h. Primase
i. DNA polymerase III holoenzyme
j. DNA polymerase I
k. DNA ligase

__Proofreads newly synthesized DNA

__Joins lagging strand edges

__Is an RNA polymerase

__Prevents unwound DNA from re-forming base-paired helix

__Is synthesized discontinuously

__Relieves stress induced by positive supercoiling

__Removes RNA primers

__Requires ATP to add negative supercoils to DNA

__First place where primosome functions

__Polarity of synthesis is opposite to replication fork movement

__Unwinds DNA at replication fork

__Synthesizes most of DNA during replication

29. Discuss some consequences of errors in DNA replication. Some types of mutations are deletions, substitutions, insertions, and frameshifts. Give short descriptions of each and discuss how they might affect the downstream protein.

30. Why is thymine used in DNA instead of uracil?

31. Which is faster, DNA polymerase I or III? What are their respective rates of replication? Why is this the case?

32. Describe one of the experimental methods that gave evidence that the genetic code is a triplet code.

33. Choose a protein–cofactor, protein–drug, or similar interaction of your choice (e.g., acetylcholine in the acetylcholine receptor). Go to the protein database (PDB) to look it up and import it into Protopedia. Then illustrate the polar, charged, and hydrophobic residues. Indicate which areas of the protein interact with the drug or cofactor, and what types of noncovalent interactions dominate. Discuss what else the protein interacts with and the biological implications.

Background Reading

Books

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. *Molecular Biology of the Cell*. Garland Science, 2008.—This is an outstanding textbook on cell biology that should be read by anyone with an interest in the biological sciences. Garland Science, New York, NY.

Atkins, P., and De Paula, J. *Atkins' Physical Chemistry*. Oxford University Press, 2006.—An excellent, mathematically intensive introduction to physical chemistry. Oxford University Press, Oxford, UK.

Carey, F., and Sundberg, R. *Advanced Organic Chemistry: Part A: Structure and Mechanisms*. Springer Science, 2007.—One of the few texts to cover advanced concepts in organic chemistry. Springer Science, Berlin, Germany.

Hettema, H. *Quantum Chemistry: Classic Scientific Papers*. World Scientific Books, 2000.—English translations of (originally German) papers by Born, London, and others, covering atoms, bonds, spectroscopy, and intermolecular interactions. World Scientific Books, Hackensack, NJ.

Kaplan, I.G. *Intermolecular Interactions*. Wiley, 2006.—Excellent coverage of the mathematics needed for intermolecular interaction calculations. Includes dozens of model potentials. John Wiley & Sons, Hoboken, NJ.

Lehninger, A., Nelson, D., and Cox, M. *Lehninger Principles of Biochemistry*. W. H. Freeman, 2008.—Classic biochemistry text with superb illustrations and problems. WH Freeman, New York, NY.

Pavia, D.L., Lampman, G.M., Kriz, G.S., and Vyvyan, J.A. *Introduction to Spectroscopy*. Brooks Cole, 2008.—Classic introduction to spectroscopy by the author of several chemical methods textbooks. Brooks Cole, Boston, MA.

Solomons, T.W.G., and Fryhle, C.B. *Organic Chemistry*. John Wiley & Sons, 2007.—Classic organic chemistry text with emphasis on applications to biology and very clear presentation. John Wiley & Sons, Hoboken, NJ.

Stone, A.J. *The Theory of Intermolecular Forces*. Clarendon Press, 1996.—A comprehensive and modern coverage of intermolecular forces in biophysical chemistry. Clarendon Press, Oxford, UK.

Tinoco, I., Sauer, K., Wang, J.C., and Puglisi, J.D. *Physical Chemistry: Principles and Applications in Biological Sciences*. Prentice Hall, 2002.—An excellent, highly physical introduction to biological concepts. Prentice Hall, Upper Saddle River, New Jersey.

Tuszynski, J.A. *Molecular and Cellular Biophysics*. Chapman & Hall, 2008.—A one-of-a-kind introduction to biological concepts from a physicist's point of view. Chapman & Hall, New York, NY.

Journal articles

Alfasi, S., Sevastsyanovich, Y., Zaffaroni, L., Griffiths, L., Hall, R., and Cole, J. (2011). Use of GFP fusions for the isolation of *Escherichia coli* strains for improved production of different target recombinant proteins. *Journal of Biotechnology* 156, 11–21.

Bayer, E.A., Skutelsky, E., and Wilchek, M. (1979). The avidin–biotin complex in affinity cytochemistry. *Methods in Enzymology* 62, 308–315.

Bratthauer, G.L. (2010). The avidin–biotin complex (ABC) method and other avidin–biotin binding methods. *Methods in Molecular Biology* 588, 257–270.

Izrailev, S., Stepaniants, S., Balsera, M., Oono, Y., and Schulten, K. (1997). Molecular dynamics study of unbinding of the avidin–biotin complex. *Biophysical Journal* 72, 1568–1581.

Koch, A.L. (1996). What size should a bacterium be? A question of scale. *Annual Review of Microbiology* 50, 317–348.

Kyratsous, C.A., Silverstein, S.J., DeLong, C.R., and Panagiotidis, C.A. (2009). Chaperone-fusion expression plasmid vectors for improved solubility of recombinant proteins in *Escherichia coli*. *Gene* 440, 9–15.

LaVallie, E.R. (2001). Production of recombinant proteins in *Escherichia coli*. *Current Protocols in Protein Science*, 00, 5.1, 5.1.1–5.1.8.

Mamat, U., Wilke, K., Bramhill, D., Schromm, A.B., Lindner, B., Kohl, T.A., Corchero, J.L., Villaverde, A., Schaffer, L., Head, S.R., Souvignier, C., Meredith, T.C., and Woodard, R.W. (2015). Detoxifying *Escherichia coli* for endotoxin-free production of recombinant proteins. *Microbial Cell Factories*, 14, 57.

McClure, W.R. (1980). Rate-limiting steps in RNA chain initiation. *Proceedings of the National Academy of Sciences U S A* 77, 5634–568.

Prescott, D.M., and Kuempel, P.L. (1972). Bidirectional replication of the chromosome in *Escherichia coli*. *Proceedings of the National Academy of Sciences U S A* 69, 2842–2845.

Rothman-Denes, L.B. (2013). Structure of *Escherichia coli* RNA polymerase holoenzyme at last. *Proceedings of the National Academy of Sciences U S A* 110, 19662–19663.

Shimada, K., and Koga, H. (2009). High-throughput production of the recombinant proteins expressed in *Escherichia coli* utilizing cDNA resources. *Methods in Molecular Biology* 577, 83–96.

Sivashanmugam, A., Murray, V., Cui, C., Zhang, Y., Wang, J., and Li, Q. (2009). Practical protocols for production of very high yields of recombinant proteins using *Escherichia coli*. *Protein Science* 18, 936–948.

Tougu, K., and Marians, K.J. (1996). The interaction between helicase and primase sets the replication fork clock. *Journal of Biological Chemistry* 271, 21398–213405.

Walls, D., and Loughran, S.T. (2011). Tagging recombinant proteins to enhance solubility and aid purification. *Methods in Molecular Biology* 681, 151–175.

Wilkinson, D.L., and Harrison, R.G. (1991). Predicting the solubility of recombinant proteins in *Escherichia coli*. *Biotechnology (N Y)* 9, 443–448.

Young, R., and Bremer, H. (1976). Polypeptide-chain-elongation rate in *Escherichia coli* B/r as a function of growth rate. *Biochemical Journal* 160, 185–194.

Online resources and software

Databases

Entrez. Search engine and database for biomedical journals (PubMed), DNA/RNA sequences (Nucleotide), protein sequences (Protein), structures (Structure), and more. The first place to go to find the sequence of a gene or protein.

SRS. Sequence retrieval system.

NIH Center for Molecular Modeling. National Institutes of Health site with links to software and databases of relevance to biochemistry and modeling. http://cmm.cit.nih.gov/modeling/

Software

Wikipedia maintains a list of computational chemistry software packages and their capabilities here: https://en.wikipedia.org/wiki/List_of_quantum_chemistry_and_solid-state_physics_software

The following are some commonly used commercial quantum chemistry packages:

Amsterdam Density Functional (ADF). Semiempirical, Hartree–Fock, and density functional theory.

Gaussian. All methods. Frequently updated. Very commonly used.

Jaguar. Ab initio and Density Functional Theory (DFT); focus on metal-containing systems.

SCIGRESS. Many methods; no Hartree–Fock.

Spartan. All methods. Easy-to-use Graphical User Interface (GUI).

TURBOMOLE. Ab initio methods, Hartree–Fock and post-Hartree–Fock.

The following are some commonly used quantum chemistry packages that are free for academic use but not necessarily open source:

DALTON. Hartree–Fock, post-Hartree–Fock, and density functional theory.

DIRAC. Hartree–Fock, post-Hartree–Fock, and density functional theory.

GAMESS. All methods.

The following are some commonly used open-source quantum chemistry packages:

MONSTERGAUSS. Started as open-source answer to Gaussian, with other features. Hartree–Fock.

PSI4. Hartree–Fock and density functional theory.

OpenAtom. Molecular mechanics and density functional theory.

CHAPTER 2

Basic Molecular Cloning of DNA and RNA

2.1 INTRODUCTION

Modern molecular biology is about manipulating small amounts of invisible, highly sensitive molecules where there is often no direct evidence that what is in the tube is what you think it is. It begins usually with a catalog, a list of naturally occurring enzymes that have been identified, purified, and in some cases mutated in order to facilitate the manipulation of DNA and RNA molecules. The currency of most cloning experiments is the *plasmid*, a circular piece of DNA found in bacteria that usually ranges from 2,000 to 14,000 base pairs (or 2–14 kb) in length. (For comparison, the genome of *Escherichia coli* is 4.6 Mb.) (See **Practical Tips 2.1**.) A plasmid usually contains all of the following features:

- An *origin of replication* (ORI). This is a 50- to 100-base-pair sequence to which host (*E. coli*) enzymes bind and signal the replication of the entire plasmid. This allows the plasmid to be replicated in *E. coli* cells, so that the experimenter can produce as much of it as desired.
- A promoter. As discussed in **Chapter 1**, a promoter is a sequence that permits RNA polymerase to bind, and thus the gene sequence that is downstream of this promoter to be expressed. Depending upon the experiment desired, this promoter can be weak or strong; inducible (i.e., requiring a nutrient or chemical to turn on) or constitutive (always on); and bacterial or eukaryotic. If a plasmid has a bacterial ORI but a mammalian promoter, the DNA of the plasmid itself is replicated in *E. coli*, but the genes it encodes are only expressed if the plasmid is put into mammalian cells.
- One or more genes of interest. These occur downstream of the promoter.
- A *selectable marker*, which is almost always a gene encoding for resistance to a specific antibiotic under the control of a bacterial promoter. All bacteria containing the plasmid will be resistant to the antibiotic, and those without it almost always will not be, allowing the experimenter to selectively amplify bacteria containing the plasmid in culture medium that contains that antibiotic. The most commonly used selection antibiotics in cloning are ampicillin (Amp), kanamycin (Kan), and tetracycline (Tet).

PRACTICAL TIPS 2.1: PLASMIDS

Bacteria reproduce asexually, with cell division giving rise to two genetically identical daughter cells each containing a circular genome (chromosome) on the order of a few hundred thousand to a few million base pairs in size. If they had no means of exchanging genetic information, their evolution would be restricted to the rate of random mutations, much too slow to adapt to environmental changes.

However, bacterial evolution is driven not by changes to the primary genome, but to what is called the *mobilome*: pieces of DNA that are separate from the primary chromosome but can interact with it as well as being carried from one bacterium to another via horizontal gene transfer (HGT). (The passage of DNA to a daughter cell is referred to as vertical gene transfer.) The size and composition of the pieces and the means of HGT vary widely; some are due to viruses and are carried by infection, and others are spread by direct *conjugation* of two bacteria or even by uptake from the environment (*transformation* or *transduction*, which we will take advantage of in the laboratory). A *plasmid* is a small (1–20 kb) circular piece of DNA that is usually passed between organisms by conjugation (certain plasmids, called conjugative plasmids, are able to initiate this process) (**Figure P2.1.1**).

Unless it bears a gene conferring resistance to an antibiotic or the ability to metabolize a common element in a restricted medium, the presence of a plasmid lowers the bacterium's fitness, as it takes energy to replicate and maintain. The term *plasmid* was coined in 1952 by Joshua Lederberg, who won the 1958 Nobel Prize in Physiology or Medicine for the discovery that bacteria can exchange genes. Since then, it has been found that HGT is responsible for crucial differences in bacterial properties, such as the variation in virulence among *E. coli* O157 strains and multidrug resistance in species of *Salmonella*, *Staphylococcus*, and others. However, only since the availability of full genome sequencing has HGT's importance in overall bacterial evolution become appreciated.

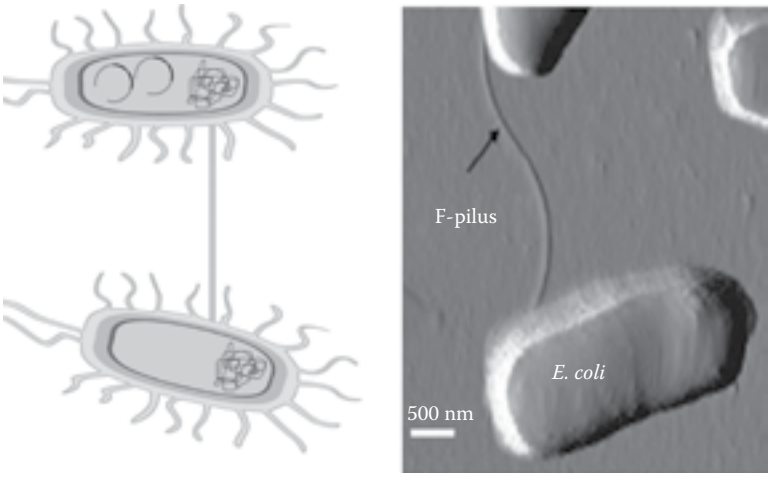


Figure P2.1.1 Plasmids and horizontal gene transfer. (a) Schematic of a bacterial cell containing plasmids (top) connected by a conjugation pilus to a cell with only genomic DNA and no plasmids (bottom). The plasmids themselves give rise to the pilus, which is a protein bridge permitting DNA to be exchanged. (b) Electron micrograph of conjugating *E. coli*.

RECOMMENDED REVIEW ARTICLES

Classic

Meynell, E., Meynell, G.G., and Datta, N. (1968). Phylogenetic relationships of drug-resistance factors and other transmissible bacterial plasmids. *Bacteriological Reviews* 32, 55–83.

Richmond, M.H. (1965). Penicillinase plasmids in *Staphylococcus aureus*. *British Medical Bulletin* 21, 260–263.

(*Continued*)

PRACTICAL TIPS 2.1 (CONTINUED): PLASMIDS

Modern

Ahmed, N., Dobrindt, U., Hacker, J., and Hasnain, S.E. (2008). Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. *Nature Reviews Microbiology* 6, 387–394.

Bower, D.M., and Prather, K.L. (2009). Engineering of bacterial strains and vectors for the production of plasmid DNA. *Applied Microbiology and Biotechnology* 82, 805–813.

Nikaido, H. (2009). Multidrug resistance in bacteria. *Annual Review of Biochemistry* 78, 19–146.

Hundreds to thousands of copies of a plasmid can be made inside a single bacterium, allowing for rapid and efficient amplification of this specific sequence. These plasmids can then be isolated from the bacterial cells, resulting in ultrapure solutions of plasmid in water at concentrations up to several milligrams per milliliter.

The trick for the molecular biologist is to insert a sequence of interest into the correct plasmid for the desired experiment. This chapter will illustrate each of the steps in amplifying, purifying, and screening plasmid DNA, and will illustrate design of a ligation experiment with an example that includes key troubleshooting steps. At the end of this section, you should be comfortable with basic plasmid manipulations and be able to design your own cloning experiment. We then survey several other cloning techniques of special interest to molecular biophysics, including some approaches to cloning large fragments of DNA and methods of mutating and rearranging genes.

2.2 OBTAINING AND STORING PLASMIDS

Many hundreds of plasmids are available commercially; the primary suppliers are listed at the end of this chapter. Some plasmids are empty except for a resistance gene and a promoter; these are known as *cloning vectors* (**Figure 2.1a**). More complex plasmids are intended for expression of the protein in *E. coli*, in which case they are called *bacterial expression vectors* (**Figure 2.1b**). Expression vectors for eukaryotic cells (yeast, mammalian cells, plants, etc.) contain an entire expression sequence that permits the gene to express in these cells (**Figure 2.1c**). Plasmids for expression of genes in bacteria other than *E. coli* and its relatives also usually use such an expression sequence, and *E. coli* is used as an intermediate because of its ease of use for cloning; this will be discussed further in **Chapter 3**.

It is often possible to purchase a plasmid containing your gene of interest and use it directly in experiments. For example, if you want to express green fluorescent protein–labeled *actin* (AcGFP) in mammalian cells, the plasmid in **Figure 2.1c** will do the trick. However, sometimes the commercial plasmid contains the wrong promoter, the wrong resistance gene, or other undesirable features. In this case, the gene of interest can be removed from its original plasmid and put into a plasmid with the necessary features. For example, if you wanted to express AcGFP in mammalian cells only in the presence of tetracycline, then the constitutive promoter in **Figure 2.1c** would not work. You would cut out the AcGFP gene from this plasmid and place it into a plasmid containing a tetracycline-inducible promoter. This is an example of a *molecular cloning* experiment; we will go through this example in detail later in the chapter.

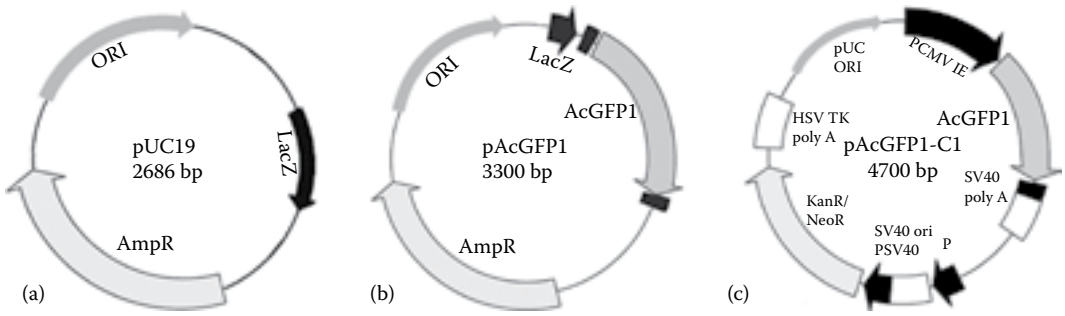


Figure 2.1 Plasmids for cloning and expression. (a) A commonly used basic cloning plasmid vector, pUC19. It has an origin of replication (ORI) and encodes a resistance gene for ampicillin (AmpR) as well as a fragment of the gene for β-galactosidase (LacZ) that can be used as a colorimetric screen for the plasmid in some strains of bacteria. (b) A bacterial expression vector, pAcGFP1, made from a pUC19 base. The AcGFP1 gene is inserted immediately downstream of the LacZ. This means that the bacteria will make a protein that is the fusion of β-galactosidase and AcGFP1, which encodes a green fluorescent protein. (c) Plasmids for expression in mammalian cells are more complex. This example vector encodes the same AcGFP1 fluorescent protein as in (b). However, two things are necessary for expression in mammalian cells. The first is a mammalian *promoter* sequence, the cytomegalovirus immediate early promoter (PCMV IE). This will permit expression in nearly all mammalian cell types at a high level. After the fluorescent protein gene is a polyadenylation sequence (SV40 polyA), which permits correct processing of the mRNA. Instead of AmpR, this plasmid contains a resistance gene that is useful in both bacteria and mammalian cells: KanR/NeoR. (Kanamycin is used in bacteria, neomycin or G418 in mammalian cells.) This gene is preceded by two promoters, a bacterial promoter (P) and PSV40, which permits expression in mammalian cells. Another polyadenylation sequence is needed for correct processing of this second gene.

Another very common type of situation is that you wish to express a gene that has been described in the literature but that is not available commercially. In this case, the procedure is to contact the author and request a sample of the plasmid described in his/her published work. Authors in most molecular biology journals are required to make these plasmids available upon request. This sequence may then be used as is, or the gene or pieces of the gene may be removed from the host plasmid and placed into a vector of your choice.

Whether commercial or from an individual laboratory, the plasmid will come in one of three forms:

- As a solution at a given concentration in water or simple buffer
- *Lyophilized* (dry) in a given amount (micrograms)
- As a bacterial *stab* (a culture of *E. coli* containing the plasmid "stabbed" into nutrient *agar*)

Dissolved plasmids should be frozen at –20°C. Lyophilized plasmids may be stored at 4°C until they are resuspended in water; then they should be frozen. Bacterial stabs should be stored at 4°C. In all cases, the plasmid should be amplified before any further experiments are done, to make sure enough is available and to store stocks in a stable form for future use.

2.3 SELECTION OF AN APPROPRIATE *E. COLI* AMPLIFICATION STRAIN; TRANSFORMATION OF *E. COLI* WITH PLASMID

Transformation

If the plasmid comes by itself and not inside cells (i.e., in solution or lyophilized), it will first need to be *transformed* into an appropriate *amplification strain* of *E. coli*. These are procedures that modern molecular biology suppliers have made routine.

Transformation refers to the process of inserting the plasmid into *E. coli* cells, where it can be replicated. *E. coli* will not simply take up plasmid DNA from its environment; it must be made *competent* to do so. Amplification strains of *E. coli* are sold in small vials (aliquots) of *chemically competent* or *electrically competent* cells. Each vial is designed for a single cloning experiment. A small amount of plasmid DNA is placed into the vial, and the cells are exposed to heat (for chemically competent cells) or an electric field (for electrically competent cells). Each of these procedures is thought to result in the opening of minute pores in the *E. coli* membrane that permit the plasmid DNA to enter. No special equipment apart from a heat plate or heated water bath is needed to use chemically competent cells. (The "chemically" refers to how they are made, not how they are transformed.) However, to use electrically competent cells, an *electroporator* is needed. This is a specialized instrument that applies a specific voltage across *electroporation cuvettes*, which are plastic cell holders containing two parallel metal plates (**Figure 2.2**).

Which amplification strain should be chosen? There are a few common ones that can be used for nearly all plasmids. However, if the plasmid is especially large or if it shows signs of instability (see the end of **Section 2.4** for further discussion), a particularly stable strain may be used (**Table 2.1**).

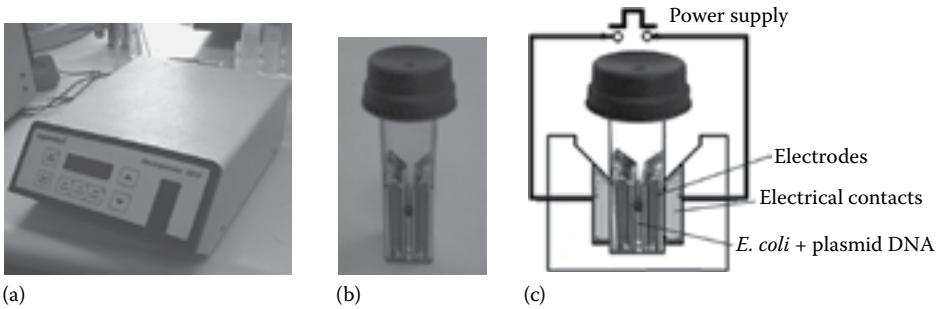


Figure 2.2 Electroporation. (a) Electroporator. (b) Electroporation cuvette (holds about 1 mL, path length usually 1 cm). (c) Schematic of how it works. A critical value of the electric field is required to open pores in the cell (usually 10 kV/cm for bacteria). The solution containing the *E. coli* must be low in electrolytes (salt), or the electrodes will arc, making a loud noise and killing all of the cells.

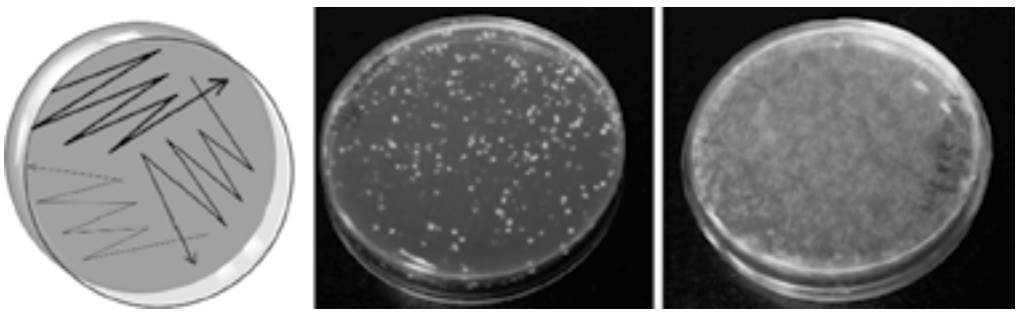
Table 2.1

Commercially Available Expression Strains of *E. coli* and Their Recommended Uses

Strain	Uses	Transformation Efficiency
TOP10	Routine cloning. Very commonly used.	10^9 (chemical) 10^{10} (electro)
DH10B	Routine cloning. Very commonly used.	10^9
DH5α	Routine cloning. Very commonly used.	10^9
Stbl2, Stbl3, Stbl4	For unstable plasmids. Stbl4 can be used for very large plasmids (>200 kb).	10^8–10^9
Mach1	Very fast-growing for rapid turnover times.	10^9
MegaX DH10B Electrocomp	Highest transformation efficiency (tricky ligations, etc.).	$>3 \times 10^{10}$
INV110	Prepares unmethylated DNA.	10^6

Selection

After transformation, the cells are plated onto a *selective plate*. This is a petri dish containing *nutrient agar* plus the antibiotic whose resistance gene is expressed in the plasmid. (Recipes for different nutrient media are given in **Chapter 3**.) Good plating technique (**Figure 2.3a**) ensures the growth of single bacterial *colonies* (**Figure 2.3b**). Each colony is made up of genetically identical cells, so it is good microbiological practice to always work with cells from a single, distinct colony. When too many bacteria are plated at once, this results in a *lawn* in which individual colonies cannot be distinguished (**Figure 2.3c**). It is bad practice to take bacteria from a lawn. Similarly, cells should not be allowed to grow for too long, or the selective antibiotic will begin to degrade, and colonies of nonresistant (thus non-plasmid-containing) bacteria will begin to grow. These are called "satellite" colonies and should be ignored. Note that if the original plasmid was supplied as a stab, the material from the stab is plated directly onto a plate in the same fashion as the transformed bacteria. A plate is stable at 4°C for a week or so unless



(a) (b) (c)

Figure 2.3 Plating bacteria. (a) Method of streaking a droplet of transfected cells or a smear from a stab so as to obtain single colonies. Each successive streak contains a lower concentration of bacteria. (b) Plate showing single, discrete colonies. (c) Bacterial lawn with no distinguishable colonies.

it becomes contaminated with environmental fungi or other bacteria; this can be identified as colonies of an unexpected shape, size, or color on the plate.

Transformation efficiency

The fraction of the transformed competent cells that should be plated to get a good number of colonies, not a lawn, can be estimated by adjusting the amount of DNA used and using the published *transformation efficiency* (T) of the strain. This is defined as the number of colonies obtained per microgram of DNA. Typical DNA concentrations used per reaction are 1–10 ng; more than 10 ng usually does not result in more colonies. Thus, the number of colonies obtained (N) is given by

$$N = T(transformants/\mu g) \times DNA transformed(\mu g) \times fraction plated \qquad (2.1)$$

If your transformation efficiencies are much lower than the published values, your technique is probably faulty. Treat the cells with care!

Competent cells are rather costly, and most labs prepare their own by setting aside one aliquot from a commercial batch and amplifying it. This is a recommended procedure but can be tricky. See **Practical Tips 2.2** for recipes and tips for the preparation of chemically and electrically competent cells. Transformation efficiencies should be determined for each batch by using a standard plasmid, such as pUC19, at a known concentration. For routine cloning experiments, efficiencies of 10^6 transformants/μg are acceptable. However, for *ligation reactions, mutagenesis*, or other low-efficiency operations, values of 10^9–10^{11} are desirable. Some labs use homemade cells for routine operations and commercial cells for difficult cloning procedures.

2.4 PLASMID AMPLIFICATION AND PURIFICATION

Amplification

Once single colonies have been obtained, they may be "picked" with a sterile pipette tip or toothpick and used to inoculate 3–5 mL of nutrient medium containing the selective antibiotic. The cells are grown for 12–16 h, at which point the culture should be visibly turbid. (If not, something has gone wrong, and the transformation should be repeated.) This volume of culture may be used to purify a small amount of DNA for screening purposes. This scale of purification is called a *miniprep*.

For plasmids of known sequence that will be used in experiments, larger amounts are desirable. In this case, the seed culture is diluted 1:100 into 50–100 mL (a *midiprep*) or 150–250 mL (a *maxiprep*) in fresh medium containing antibiotic. The cultures are grown for a further 12–16 h, and the cells are *pelleted* (centrifuged to the bottom of a tube) for DNA extraction.

Purification

Traditional DNA purification methods involve ultracentrifugation on a CsCl density gradient. First the *E. coli* are broken open or *lysed*, and centrifuged to remove the major contaminants such as the cell wall. The lysate is then mixed with CsCl and a fluorescent dye (ethidium bromide [EtBr]) that intercalates into the DNA double helix. After 20 h of centrifugation, the plasmid DNA forms a distinct band that is identified by the dye fluorescence and removed with a needle.

PRACTICAL TIPS 2.2: PREPARATION OF ELECTRICALLY AND CHEMICALLY COMPETENT CELLS

TIPS

- Start with fresh, discrete single colony.
- Make sure there is no detergent on the glassware. (Rinse well with distilled water.)
- Watch your cell densities carefully. Bacterial concentrations are estimated by measuring the optical density at 600 nm (OD_{600}) in a UV–Vis spectrophotometer. Do not let OD_{600} go above log phase (different for different spectrometers; do a growth curve to determine exact value).
- Treat the cells very gently. Centrifuge at the lowest possible speed; do not pipette them up and down.

FOR ELECTRICALLY AND CHEMICALLY COMPETENT CELLS

The Night Before

Inoculate 100 mL of a very rich nutrient medium (e.g., 2x YT, see **Appendix 2**) with 10 µL of commercial competent cells. No antibiotics are added since there is no plasmid! Incubate 37°C for 14–16 h, shaking at 200 rpm.

The Next Morning

Keep *everything* on ice from now on. All pipettes, glassware, etc. should be chilled. You can even do the preparation in a cold room if you have access to one.

(1) Dilute the overnight culture 1:10 in fresh medium (for 1 L total volume). Place in shaker at 37°C and read OD_{600} every 20 min until it reaches mid-log phase (usually ~OD_{600} = 0.6).
(2) Pellet the cells by centrifugation for 20 min at 4°C and approximately 4000 g.

FOR ELECTRICALLY COMPETENT CELLS

(3) Resuspend each pellet in 35–40 mL of sterile cold water and transfer to a 50 mL centrifuge tube. Remove the supernatant, resuspend the pellet in fresh cold water, and repeat.
(4) Repeat these wash steps with 10% glycerol in sterile cold water. On the final wash, pool all of the cells together.
(5) Estimate the pellet volume and resuspend in an equal amount of 10% glycerol (or slightly less if very high levels of competence are desired).
(6) Prepare a slush of dry ice/ethanol.
(7) Distribute the cells into Eppendorf tubes in 50 µL aliquots and quick-freeze them in the slush.
(8) Transfer to a precooled cardboard box and store at –80°C for a year or more.

FOR CHEMICALLY COMPETENT CELLS

(3) Resuspend the pellet in 1/2 the original volume of sterile, cold 100 mM $CaCl_2$. Incubate on ice 20 min.
(4) Pellet the cells, resuspend in 1/10 the original volume of sterile, cold 100 mM $CaCl_2$. Incubate on ice 60 min.
(5) Add cold, sterile glycerol to a final concentration of 15%.
(6) Prepare a slush of dry ice/ethanol.
(7) Distribute the cells into Eppendorf tubes in 50 µL aliquots and quick-freeze them in the slush.
(8) Transfer to a precooled cardboard box and store at –80°C for a year or more.

The genomic DNA bands at a lower density than plasmid DNA. Although this method is time consuming and the reagents are toxic, it can be repeated to yield extremely pure plasmid DNA, is less costly than commercial kits for large-scale preparations, and is still widely used by those who need large amounts (hundreds of micrograms to milligrams) of pure plasmid DNA. Some references given at the end of the chapter provide protocols for these methods.

For routine applications, most laboratories use commercial plasmid purification kits that are based upon *exchange chromatography*. A resin or membrane is provided in a column along with a selection of buffers. The composition of the initial buffers favors binding of the DNA to the resin, usually based upon the molecule's negative charge. Once it is bound, it can be washed to remove impurities. The last buffer favors DNA dissociation, and the pure product is collected. Anion or cation exchange chromatography techniques are ubiquitous in biochemistry for purification of molecules, and we will not discuss them in detail here except to refer to basic textbooks and papers cited at the end of the chapter. The size of the column varies according to the amount of DNA to be purified; its binding capacity can be as high as 10 mg of DNA for a *gigaprep*. Most commercial kits yield plasmid DNA at least as pure as a single round of CsCl centrifugation, which is sufficient for nearly all applications. Some kits also contain an extra wash step for removal of the *E. coli* lipopolysaccharide (LPS), or *endotoxin*. These endotoxin-free or "endo-free" kits are recommended when the DNA is going to be used in eukaryotic cells, since LPS is toxic.

Measuring concentration and purity of extracted DNA

Plasmid DNA is eluted from the purification columns in a supercoiled form. Its concentration and purity can be determined using ultraviolet–visible (UV–Vis) absorbance spectroscopy; nucleic acids absorb most strongly at 260 nm (A_{260}), while proteins absorb at 280 nm and nonspecific turbidity can be measured further toward the visible (usually A_{320}) (**Figure 2.4**). Concentration is calculated using Beer's law:

$$c = \frac{A}{\varepsilon \ell}, \tag{2.2}$$

where A is A_{260}(sample) – A_{260}(blank) or $A_{260} - A_{320}$, ℓ is the cuvette path length (usually 1 cm), and ε is the extinction coefficient. The concentration of the sample

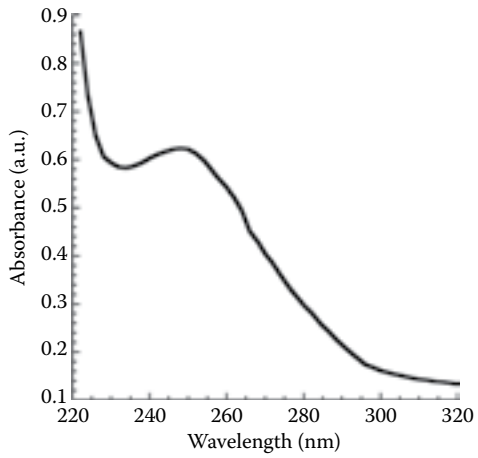


Figure 2.4 Absorbance spectrum of purified plasmid DNA.

should be adjusted so that A_{260} is between 0.05 and 1.0, as more concentrated samples fall outside the linear concentration-versus-absorbance range, and more dilute samples are prone to error. For most maxipreps and minipreps, this means that the sample should be diluted 50–100 times. Values of ε are determined by the length and sequence of the DNA but can be approximated as 6600 M^{-1} cm^{-1} for double-stranded DNA (or RNA) and 8500 M^{-1} cm^{-1} for single-stranded DNA (or RNA). (Note that absorbance techniques do not work well with quick and dirty miniprep kits, whose products are too full of contaminants to yield reliable values.)

The purity of the DNA can be estimated by the ratio A_{260}/A_{280}, since the major contaminants are usually proteins, which absorb strongly at 280 nm. A value of 2.0 indicates pure nucleic acid; a good value for a very clean midiprep or maxiprep is 1.8. Anything less than 1.6 is cause for concern. Again, this technique should not be used for most kit-based minipreps. This also cannot identify nucleic acid contaminants such as genomic DNA or cellular RNA. Finally, it is important to note that a preparation can contain pure plasmid DNA and still not be what was wanted. Some plasmids, particularly large ones, have a tendency to *recombine*, and entire large pieces of the sequence may be missing. It is also possible for bacteria to "expel" toxic genes, retaining only the antibiotic resistance. For these reasons, it is always necessary to screen any amplified plasmid by the technique of *restriction mapping*.

2.5 PLASMID RESTRICTION MAPPING AND AGAROSE GEL ELECTROPHORESIS

Restriction enzymes

Restriction mapping is one of the most useful tools in molecular biology. It resulted from the identification and exploitation of the way in which DNA is exchanged and

Table 2.2

Some Restriction Enzymes, Illustrating Types of Ends Produced, Recognition Sequences, and Different Types of Degeneracy

Enzyme Name (Organism)	Recognition Sequence and Cut Pattern	Notes
Sma I (*Serratia marcescens*)	5′—CCC↓GGG—3′ 3′—GGG↑CCC—5′	Isoschizomer of Xma I, produces blunt end
Xma I (*Xanthomonas malvacearum*)	5′—C↓CCGGG—3′ 3′—GGGCC↑C—5′	Isoschizomer of Sma I, produces sticky end
Bam HI (*Bacillus amyloliquefaciens* H)	5′—G↓GATCC—3′ 3′—CCTAG↑G—5′	Compatible sticky ends with Bg lII
Bg lII (*Bacillus globigii*)	5′—A↓GATCT—3′ 3′—TCTAG↑A—5′	Compatible sticky ends with Bam HI
Bse YI (*Bacillus* sp. 2521)	5′—C↓CCAGC—3′ 3′—GGGTC↑G—5′	Nonpalindromic
Bpm I (*Bacillus pumilus*)	5′—CTCGAG$(N)_{16}$↓—3′ 3′—GACCTC$(N)_{14}$↑—5′	Cuts downstream of recognition site
Not I (*Nocardia otitidis-caviarum*)	5′—GC↓GGCCGC—3′ 3′—CGCCGG↑CG—5′	8-base-pair recognition site

Note: The name of the enzyme is an abbreviation of the scientific name of the organism from which it was isolated; if more than one from the same species is used, it is numbered II, III, etc. Note that Sma I and Xma I recognize the exact same site; they are thus called isoschizomers of each other. However, Sma I produces a blunt end, while Xma I yields sticky ends. Bam HI and Bg lII are not isoschizomers of each other but produce compatible sticky ends. Also note that most restriction sites are palindromic (they read the same way 5′ to 3′ as 3′ to 5′), but Bse YI is an exception to this rule. Not all enzymes cut exactly where they recognize; "N" refers to any nucleotide.

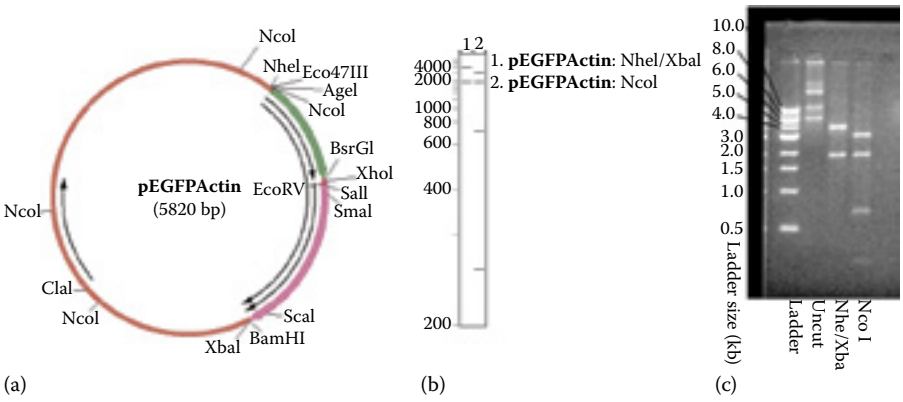


Figure 2.5 Restriction analysis by mapping and agarose gel electrophoresis. (a) Plasmid map of the plasmid pEGFP–actin, showing the gene coding regions and many of the restriction sites. Different computer programs are available to find these restriction sites based upon the plasmid sequence published by the manufacturer or on public databases. (b) Predicted appearance of the gel fragments if this plasmid were to be cut with (1) the enzymes Nhe I and Xba I or (2) Nco I. Even though Nco I is not a good enzyme for cloning, because it cuts the plasmid in too many places, it is useful for screening as it provides a good characteristic pattern with four different-sized fragments. (c) Actual appearance of the gel with plasmid cut as in (b). The leftmost lane contains the ladder; the brightest band (3.0 kb) contains 125 ng of DNA, while the other bands each contain ~40 ng. The next lane contains uncut plasmid, showing an unpredictable pattern. The third and fourth lanes are the restriction digests corresponding to (b).

eliminated in nature. Plasmids can only replicate in bacteria if they are circular; if cut open, they are simply degraded. A series of enzymes have evolved in bacteria to protect them against invading foreign DNA by recognizing and cleaving specific DNA sequences. These are known as restriction endonucleases or *restriction enzymes*, and are isolated and purified by molecular biology suppliers and sold in catalogs according to their target sequence. Most restriction enzymes identify and cut a 4- or 6-base-pair sequence; a few recognize 8- or 10-base-pair sequences and are especially useful when cloning using large plasmids. Sometimes, the recognized sequence can be degenerate (**Table 2.2**). Over a hundred of these enzymes are commonly used.

These specific sequences recognized by restriction enzymes, known as *restriction sites*, are a key element in plasmids used for cloning. The occurrence of restriction sites for the most common enzymes is used to create a unique *restriction map* of each plasmid (**Figure 2.5a**).

Screening purified DNA

The plasmid map can be used in several ways. For screening, a small amount of plasmid (20–100 ng) is cut with one or more enzymes, resulting in two or more fragments of specific sizes. These fragments are then separated by a technique called *agarose gel electrophoresis*, which is based upon two principles: the negative charge of DNA and the ability of 0.5–2% solutions of agarose to gel into a hydrogel with pore sizes that limit the diffusion of linear DNA molecules according to size. Thus, if DNA is placed into the gel in an electrolyte solution and an electric field is applied, the fragments will migrate toward the cathode as a function of size. The precise physics behind the migration has not been determined (see **Advanced Topic 2.1**), but linear DNA of molecular weight MW migrates at a speed proportional to $\log^{-1}(MW)$ (**Figure 2.5b**). The DNA is visualized using a fluorescent *intercalating* dye, usually EtBr (**Figure 2.5c**). Commercial *ladders* are available