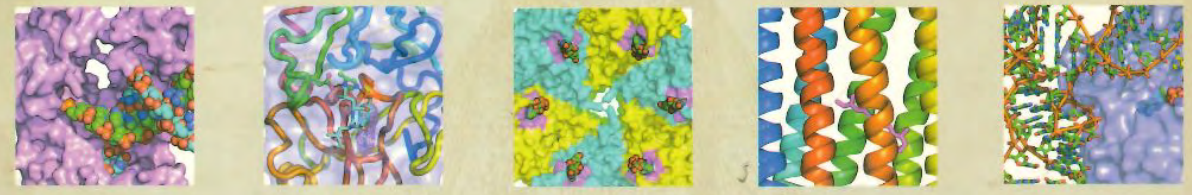
FUNDAMENTALS OF BIOCHEMISTRY

THIRD EDITION

Life at the Molecular Level



Donald Voet Judith G. Voet Charlotte W. Pratt

FOR INSTRUCTORS

WileyPLUS is built around the activities you perform in your class each day. With WileyPLUS you can:

Prepare & Present

Create outstanding class presentations using a wealth of resources such as enhanced art, PowerPoint slides containing text art optimized for presentation, animated figures, Guided Explorations, Interactive Exercises (featuring Jmol rendered 3D molecules), and kinemages. You can even add materials you have created yourself.

"It has been a great help, and I believe it has helped me to achieve a better grade."

Michael Morris,
Columbia Basin College

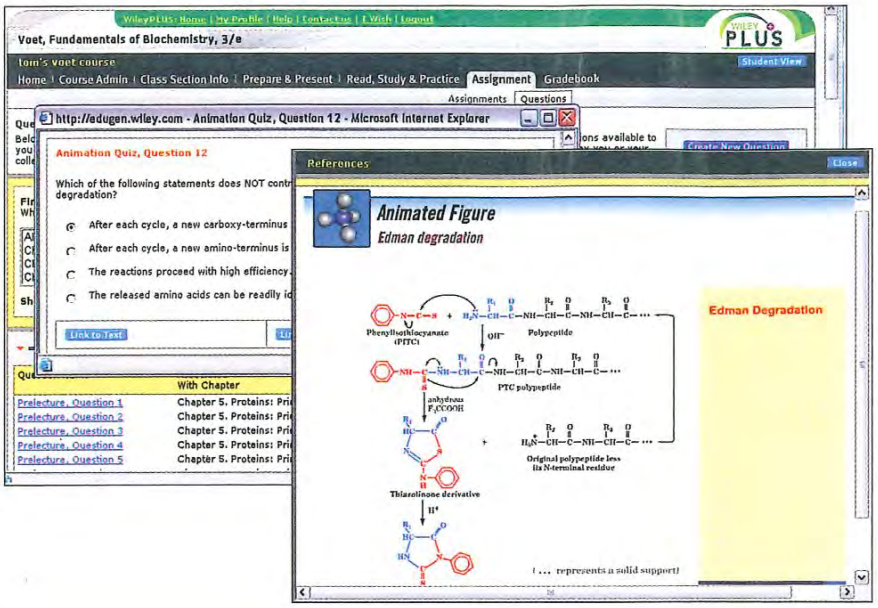
Create Assignments

Automate the assigning and grading of homework or quizzes by using the provided question banks, featuring over 700 conceptual questions, with detailed answer feedback.

Track Student Progress

Keep track of your students' progress and analyze individual and overall class results.

Now Available with WebCT, eCollege, and ANGEL Learning!



FOR STUDENTS

You have the potential to make a difference!

WileyPLUS is a powerful online system packed with features to help you make the most of your potential and get the best grade you can!

With WileyPLUS you get:

- A complete online version of your text and other study resources.
- Problem-solving help, instant grading, and feedback on your homework and quizzes.
- The ability to track your progress and grades throughout the term.

For more information on what *WileyPLUS* can do to help you and your students reach their potential, please visit www.wileyplus.com/experience.

82% of students surveyed said it made them better prepared for tests. *

*Based upon 7,000 responses to student surveys in academic year 2006-2007.

THIRD EDITION

FUNDAMENTALS OF Biochemistry

LIFE AT THE MOLECULAR LEVEL

Donald Voet
University of Pennsylvania

Judith G. Voet
Swarthmore College, Emeritus

Charlotte W. Pratt
Seattle Pacific University

John Wiley & Sons, Inc.

IN MEMORY OF WILLIAM P. JENCKS

scholar, teacher, friend

Vice-President & Executive Publisher	Kaye Pace
Associate Publisher	Petra Recter
Marketing Manager	Amanda Wainer
Assistant Editor	Alyson Rentrop
Senior Production Editor	Sandra Dumas
Production Manager	Dorothy Sinclair
Director of Creative Services	Harry Nolan
Cover Design	Madelyn Lesure
Text Design	Laura C. Ierardi
Photo Department Manager	Hilary Newman
Photo Editors	Hilary Newman, Sheena Goldstein
Illustration Editor	Sigmund Malinowski
Pathways of Discovery Portraits	Wendy Wray
Senior Media Editor	Thomas Kulesa
Production Management Services	Suzanne Ingrao/Ingrao Associates

Background Photo Cover Credit: Lester Lefkowitz/Getty Images

Inset Photo Credits: Based on X-ray structures by (left to right) Thomas Steitz, Yale University; Daniel Koshland, Jr., University of California at Berkeley; Emmanual Skordalakis and James Berger, University of California at Berkeley; Nikolaus Grigorieff and Richard Henderson, MRC Laboratory of Molecular Biology, U.K.; Thomas Steitz, Yale University.

This book was set in 10/12 Times Ten by Aptara and printed and bound by Courier/Kendallville. The cover was printed by Phoenix Color Corporation.

This book is printed on acid free paper. ∞

To order books or for customer service, please call 1-800-CALL WILEY (225-5945).

ISBN-13 978-0470-12930-2

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

About the Authors

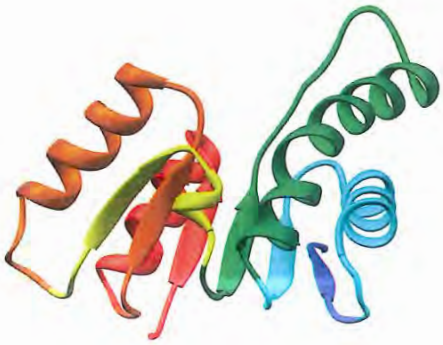
Donald Voet received a B.S. in Chemistry from the California Institute of Technology, a Ph.D. in Chemistry from Harvard University with William Lipscomb, and did postdoctoral research in the Biology Department at MIT with Alexander Rich. Upon completion of his postdoctoral research, Don took up a faculty position in the Chemistry Department at the University of Pennsylvania where, for the past 38 years, he has taught a variety of biochemistry courses as well as general chemistry. His major area of research is the X-ray crystallography of molecules of biological interest. He has been a visiting scholar at Oxford University, the University of California at San Diego, and the Weizmann Institute of Science in Israel. Together with Judith G. Voet, he is Co-Editor-in-Chief of the journal *Biochemistry and Molecular Biology Education.* He is a member of the Education Committee of the International Union of Biochemistry and Molecular Biology. His hobbies include backpacking, scuba diving, skiing, travel, photography, and writing biochemistry textbooks.

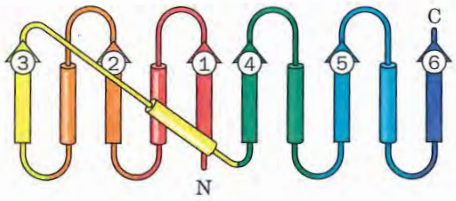
Judith ("Judy") Voet received her B.S. in Chemistry from Antioch College and her Ph.D. in Biochemistry from Brandeis University with Robert H. Abeles. She has done postdoctoral research at the University of Pennsylvania, Haverford College, and the Fox Chase Cancer Center. Her main area of research involves enzyme reaction mechanisms and inhibition. She taught Biochemistry at the University of Delaware before moving to Swarthmore College. She taught there for 26 years, reaching the position of James H. Hammons Professor of Chemistry and Biochemistry before going on "permanent sabbatical leave." She has been a visiting scholar at Oxford University, University of California, San Diego, University of Pennsylvania, and the Weizmann Institute of Science, Israel. She is Co-Editor-in-Chief of the journal *Biochemistry and Molecular Biology Education*. She has been a member of the Education and Professional Development Committee of the American Society for Biochemistry and Molecular Biology as well as the Education Committee of the International Union of Biochemistry and Molecular Biology. Her hobbies include hiking, backpacking, scuba diving, and tap dancing.

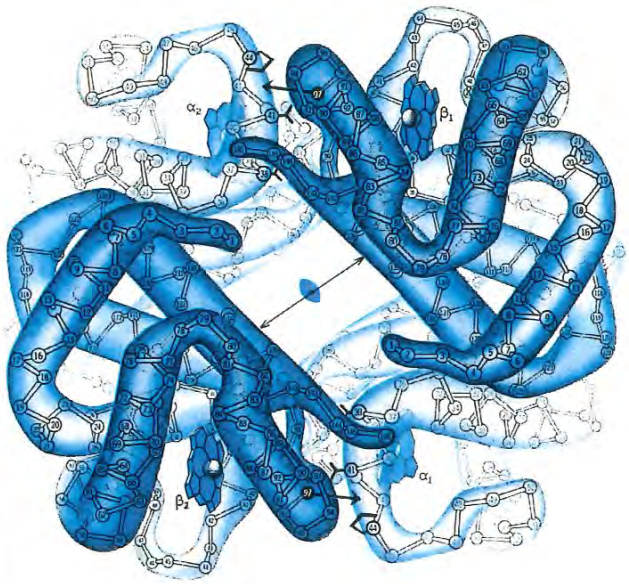
Charlotte Pratt received her B.S. in Biology from the University of Notre Dame and her Ph.D. in Biochemistry from Duke University under the direction of Salvatore Pizzo. Although she originally intended to be a marine biologist, she discovered that Biochemistry offered the most compelling answers to many questions about biological structure–function relationships and the molecular basis for human health and disease. She conducted postdoctoral research in the Center for Thrombosis and Hemostasis at the University of North Carolina at Chapel Hill. She has taught at the University of Washington and currently teaches at Seattle Pacific University. In addition to working as an editor of several biochemistry textbooks, she has co-authored *Essential Biochemistry* and previous editions of *Fundamentals of Biochemistry.*

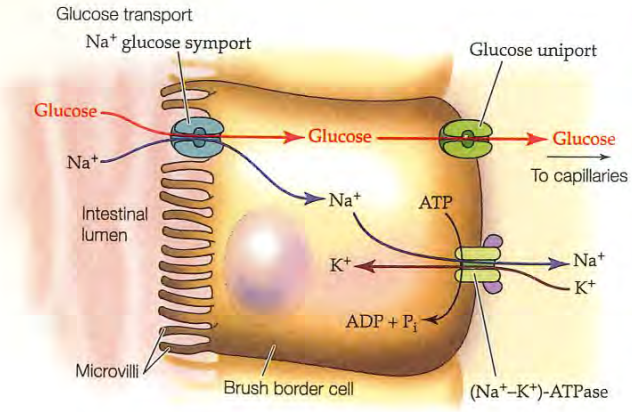
Brief Contents

Contents

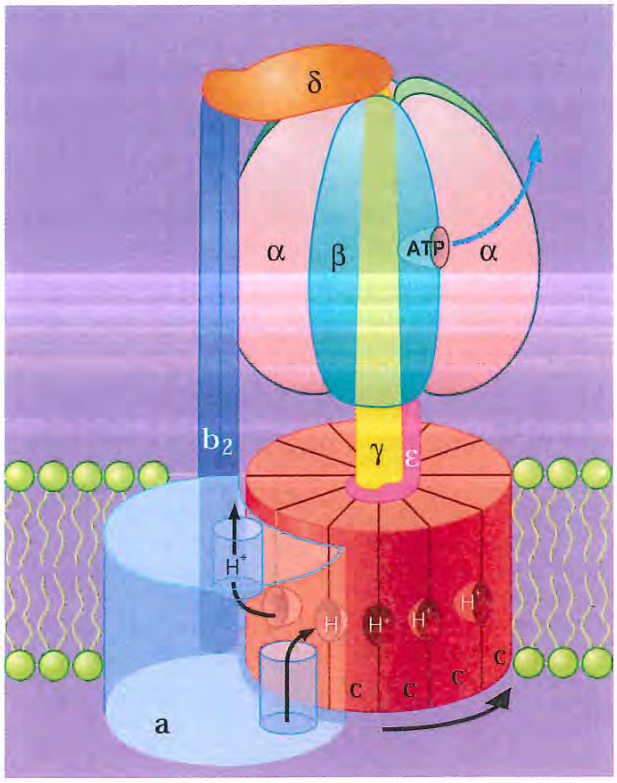


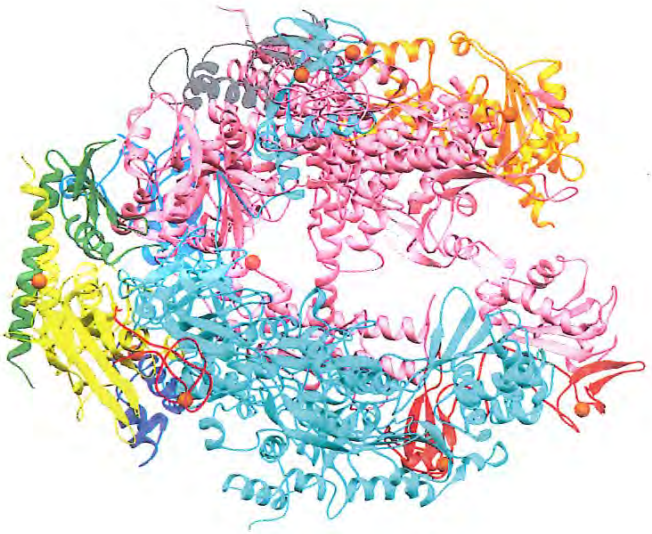


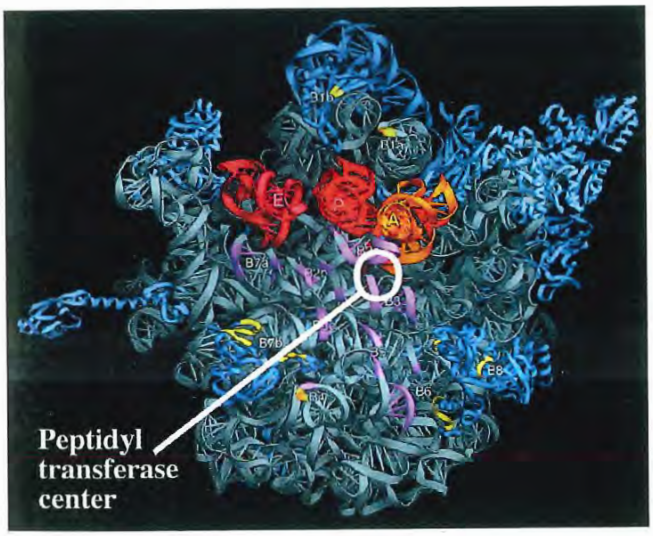


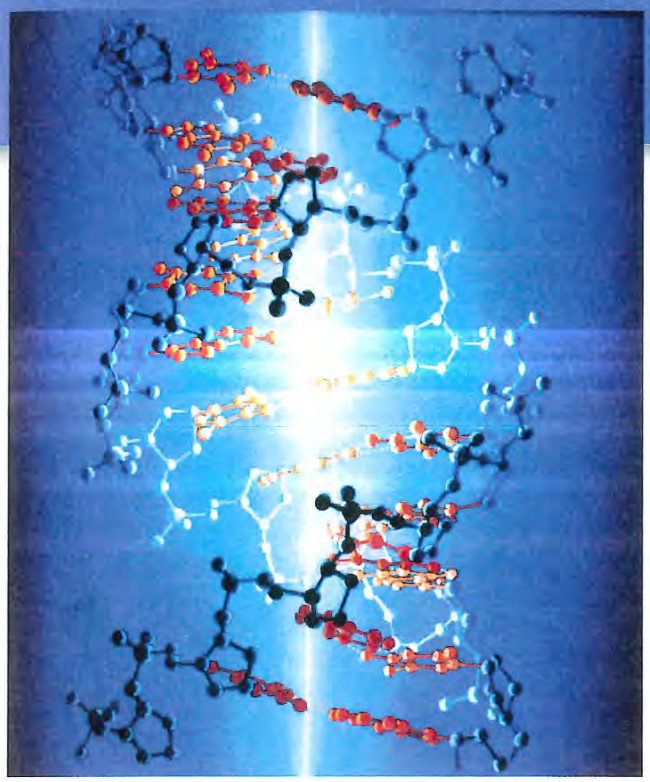


PART III ENZYMES









3 Nucleotides, Nucleic Acids, and Genetic Information

A DNA molecule consists of two strands that wind around a central axis, shown here as a glowing wire. [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.]

MEDIA RESOURCES

(Available at www.wiley.com/college/voet)

Guided Exploration 1. Overview of transcription and translation
Guided Exploration 2. DNA sequence determination by the chain-terminator method
Guided Exploration 3. PCR and site-directed mutagenesis
Interactive Exercise 1. Three-dimensional structure of DNA
Animated Figure 3-26. Construction of a recombinant DNA molecule
Animated Figure 3-27. Cloning with bacteriophage λ
Animated Figure 3-30. Site-directed mutagenesis
Kinemage Exercise 2-1. Structure of DNA
Kinemage Exercise 2-2. Watson–Crick base pairs
Bioinformatics Exercises Chapter 3. Databases for the Storage and "Mining" of Genome Sequences

CHAPTER CONTENTS

Despite obvious differences in lifestyle and macroscopic appearance, organisms exhibit striking similarity at the molecular level. The structures and metabolic activities of all cells rely on a common set of molecules that includes amino acids, carbohydrates, lipids, and nucleotides, as well as their polymeric forms. Each type of compound can be described in terms of its chemical makeup, its interactions with other molecules, and its physiological function. We begin our survey of biomolecules with a discussion of the **nucleotides** and their polymers, the **nucleic acids.**

Nucleotides are involved in nearly every facet of cellular life. Specifically, they participate in oxidation–reduction reactions, energy transfer, intracellular signaling, and biosynthetic reactions. Their polymers, the nucleic acids DNA and RNA, are the primary players in the storage

and decoding of genetic information. Nucleotides and nucleic acids also perform structural and catalytic roles in cells. No other class of molecules participates in such varied functions or in so many functions that are essential for life.

Evolutionists postulate that the appearance of nucleotides permitted the evolution of organisms that could harvest and store energy from their surroundings and, most importantly, could make copies of themselves. Although the chemical and biological details of early life-forms are the subject of speculation, it is incontrovertible that life as we know it is inextricably linked to the chemistry of nucleotides and nucleic acids.

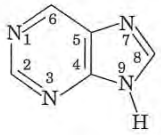
In this chapter, we briefly examine the structures of nucleotides and the nucleic acids DNA and RNA. We also consider how the chemistry of these molecules allows them to carry biological information in the form of a sequence of nucleotides. This information is expressed by the transcription of a segment of DNA to yield RNA, which is then translated to form protein. Because a cell's structure and function ultimately depend on its genetic makeup, we discuss how genomic sequences provide information about evolution, metabolism, and disease. Finally, we consider some of the techniques used in manipulating DNA in the laboratory. In later chapters, we will examine in greater detail the participation of nucleotides and nucleic acids in metabolic processes. Chapter 24 includes additional information about nucleic acid structures, DNA's interactions with proteins, and DNA packaging in cells, as a prelude to several chapters discussing the roles of nucleic acids in the storage and expression of genetic information.

1 Nucleotides

LEARNING OBJECTIVE

- Become familiar with the structures and nomenclature of the eight common nucleotides.

Nucleotides are ubiquitous molecules with considerable structural diversity. *There are eight common varieties of nucleotides, each composed of a nitrogenous base linked to a sugar to which at least one phosphate group is also attached.* The bases of nucleotides are planar, aromatic, heterocyclic molecules that are structural derivatives of either **purine** or **pyrimidine** (although they are not synthesized *in vivo* from either of these organic compounds).

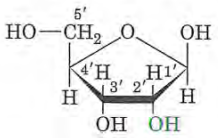


Purine

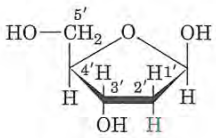
Pyrimidine

The most common purines are **adenine (A)** and **guanine (G),** and the major pyrimidines are **cytosine (C), uracil (U),** and **thymine (T).** The purines form bonds to a five-carbon sugar (a pentose) via their N9 atoms, whereas pyrimidines do so through their N1 atoms (Table 3-1).

In **ribonucleotides,** the pentose is **ribose,** while in **deoxyribonucleotides** (or just **deoxynucleotides**), the sugar is **2′-deoxyribose** (i.e., the carbon at position 2′ lacks a hydroxyl group).

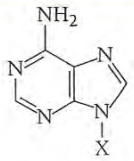
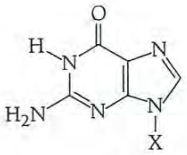
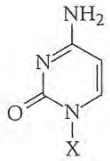
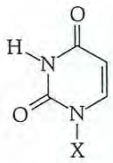
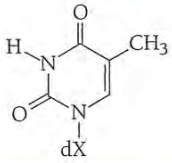


Ribose



Deoxyribose

Table 3-1 Names and Abbreviations of Nucleic Acid Bases, Nucleosides, and Nucleotides

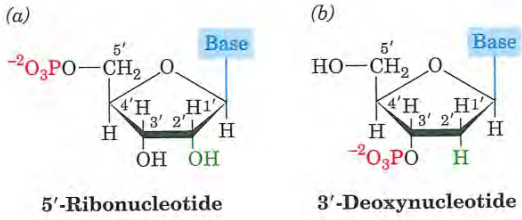
Base Formula	Base (X = H)	Nucleoside (X = ribose[a])	Nucleotide[b] (X = ribose phosphate[a])
	Adenine Ade A	Adenosine Ado A	Adenylic acid Adenosine monophosphate AMP
	Guanine Gua G	Guanosine Guo G	Guanylic acid Guanosine monophosphate GMP
	Cytosine Cyt C	Cytidine Cyd C	Cytidylic acid Cytidine monophosphate CMP
	Uracil Ura U	Uridine Urd U	Uridylic acid Uridine monophosphate UMP
	Thymine Thy T	Deoxythymidine dThd dT	Deoxythymidylic acid Deoxythymidine monophosphate dTMP

[a]The presence of a 2′-deoxyribose unit in place of ribose, as occurs in DNA, is implied by the prefixes "deoxy" or "d." For example, the deoxynucleoside of adenine is deoxyadenosine or dA. However, for thymine-containing residues, which rarely occur in RNA, the prefix is redundant and may be dropped. The presence of a ribose unit may be explicitly implied by the prefix "ribo."

[b]The position of the phosphate group in a nucleotide may be explicitly specified as in, for example, 3′-AMP and 5′-GMP.

Note that the "primed" numbers refer to the atoms of the pentose; "unprimed" numbers refer to the atoms of the nitrogenous base.

In a ribonucleotide or a deoxyribonucleotide, one or more phosphate groups are bonded to atom C3′ or atom C5′ of the pentose to form a 3′-nucleotide or a 5′-nucleotide, respectively (Fig. 3-1). When the phosphate group is absent, the compound is known as a **nucleoside.** A 5′-nucleotide can therefore be called a nucleoside-5′-phosphate. Nucleotides most commonly contain one to three phosphate groups at the C5′ position and are called nucleoside monophosphates, diphosphates, and triphosphates.



■ **Figure 3-1 | Chemical structures of nucleotides.** (*a*) A 5′-ribonucleotide and (*b*) a 3′-deoxynucleotide. The purine or pyrimidine base is linked to C1′ of the pentose and at least one phosphate (*red*) is also attached. A nucleoside consists only of a base and a pentose.

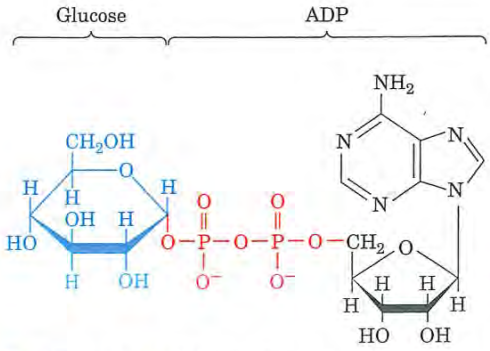
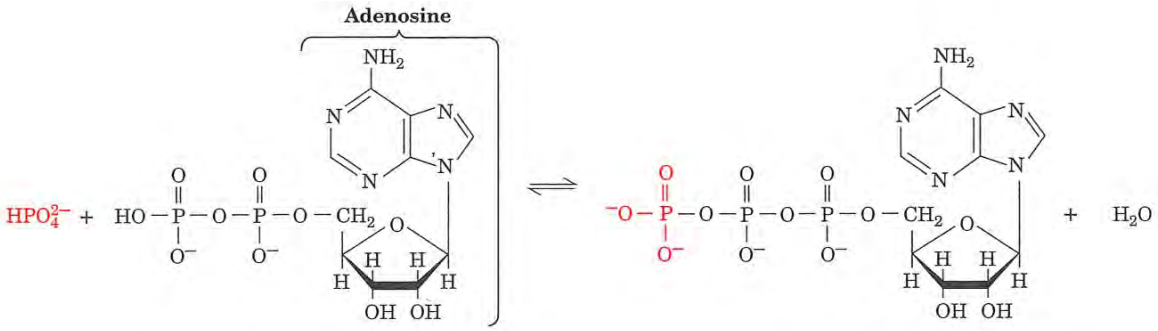


Figure 3-2 | ADP-glucose. In this nucleotide derivative, glucose (*blue*) is attached to adenosine (*black*) by a diphosphate group (*red*).

The structures, names, and abbreviations of the common bases, nucleosides, and nucleotides are given in Table 3-1. Ribonucleotides are components of **RNA (ribonucleic acid),** whereas deoxynucleotides are components of **DNA (deoxyribonucleic acid).** Adenine, guanine, and cytosine occur in both ribonucleotides and deoxynucleotides (accounting for six of the eight common nucleotides), but uracil primarily occurs in ribonucleotides and thymine occurs in deoxynucleotides. Free nucleotides, which are anionic, are almost always associated with the counterion Mg^{2+} in cells.

Nucleotides Participate in Metabolic Reactions. The bulk of the nucleotides in any cell are found in polymeric forms, as either DNA or RNA, whose primary functions are information storage and transfer. However, free nucleotides and nucleotide derivatives perform an enormous variety of metabolic functions not related to the management of genetic information.

Perhaps the best known nucleotide is **adenosine triphosphate (ATP),** a nucleotide containing adenine, ribose, and a triphosphate group. ATP is often mistakenly referred to as an energy-storage molecule, but it is more accurately termed an energy carrier or energy transfer agent. The process of photosynthesis or the breakdown of metabolic fuels such as carbohydrates and fatty acids leads to the formation of ATP from **adenosine diphosphate (ADP):**



ATP diffuses throughout the cell to provide energy for other cellular work, such as biosynthetic reactions, ion transport, and cell movement. The chemical potential energy of ATP is made available when it transfers one (or two) of its phosphate groups to another molecule. This process can be represented by the reverse of the preceding reaction, namely, the hydrolysis of ATP to ADP. (As we shall see in later chapters, the interconversion of ATP and ADP in the cell is not freely reversible, and free phosphate groups are seldom released directly from ATP.) The degree to which ATP participates in routine cellular activities is illustrated by calculations indicating that while the concentration of cellular ATP is relatively moderate (~5 mM), humans typically recycle their own weight of ATP each day.

Nucleotide derivatives participate in a wide variety of metabolic processes. For example, starch synthesis in plants proceeds by repeated additions of glucose units donated by ADP–glucose (Fig. 3-2). Other nucleotide derivatives, as we shall see in later chapters, carry groups that undergo oxidation–reduction reactions. The attached group, which may be a small molecule such as glucose (Fig. 3-2) or even another nucleotide, is typically linked to the nucleotide through a mono- or diphosphate group.

CHECK YOUR UNDERSTANDING

Describe the general structure of a nucleoside and a nucleotide.

Describe the difference between a ribonucleotide and a deoxyribonucleotide.

2 Introduction to Nucleic Acid Structure

Nucleotides can be joined to each other to form the polymers that are familiar to us as RNA and DNA. In this section, we describe the general features of these nucleic acids. Nucleic acid structure is considered further in Chapter 24.

LEARNING OBJECTIVES

- Understand how nucleotides are linked together to form nucleic acids.
- Become familiar with the structural features of the DNA double helix.

A | Nucleic Acids Are Polymers of Nucleotides

The nucleic acids are chains of nucleotides whose phosphates bridge the 3′ and 5′ positions of neighboring ribose units (Fig. 3-3). The phosphates of these **polynucleotides** are acidic, so at physiological pH, nucleic acids are polyanions. The linkage between individual nucleotides is known as a **phosphodiester bond,** so named because the phosphate is esterified to two ribose units. Each nucleotide that has been incorporated into the polynucleotide is known as a **nucleotide residue.** The terminal residue whose C5′

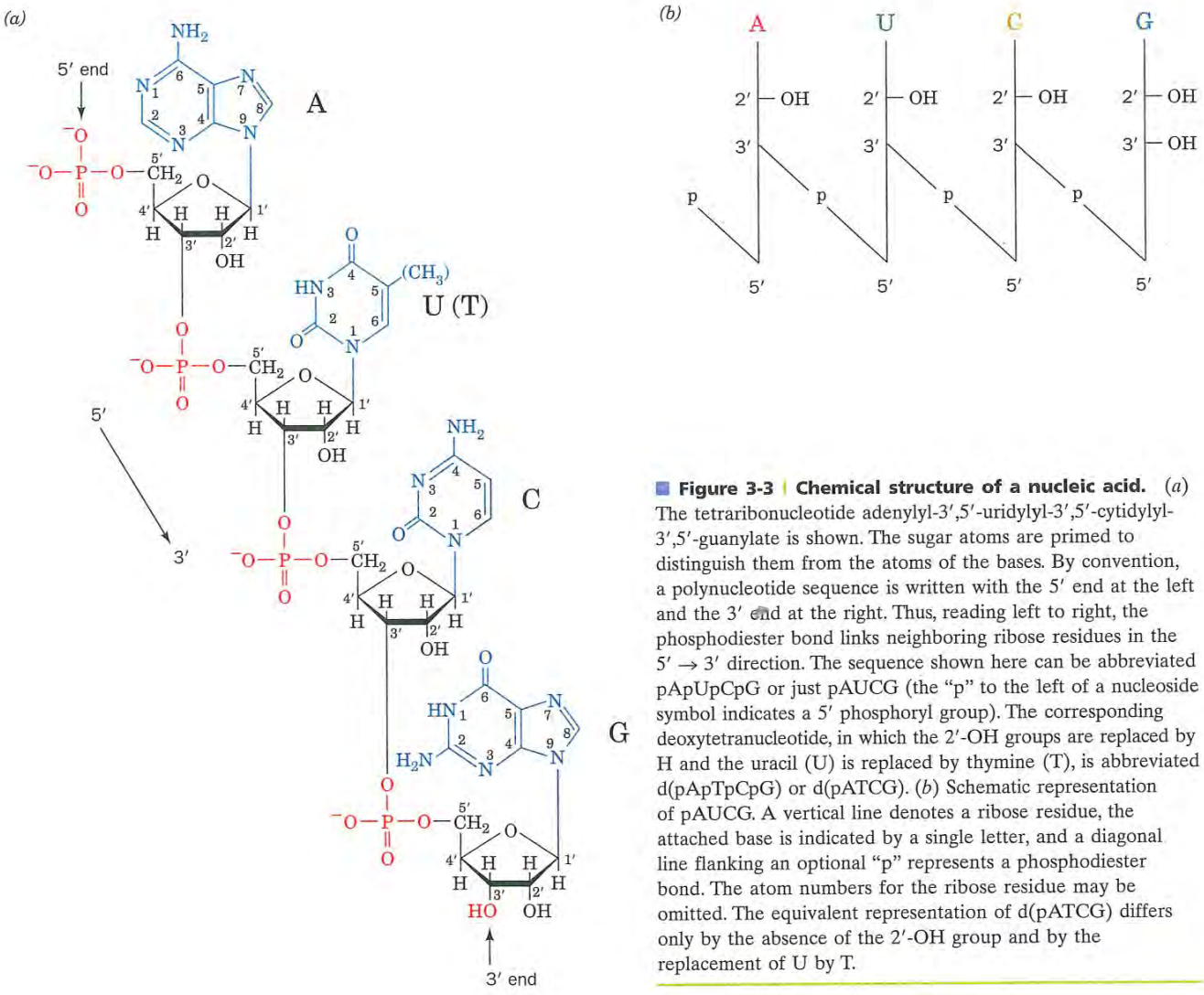


Figure 3-3 | Chemical structure of a nucleic acid. (*a*) The tetraribonucleotide adenylyl-3′,5′-uridylyl-3′,5′-cytidylyl-3′,5′-guanylate is shown. The sugar atoms are primed to distinguish them from the atoms of the bases. By convention, a polynucleotide sequence is written with the 5′ end at the left and the 3′ end at the right. Thus, reading left to right, the phosphodiester bond links neighboring ribose residues in the 5′ → 3′ direction. The sequence shown here can be abbreviated pApUpCpG or just pAUCG (the "p" to the left of a nucleoside symbol indicates a 5′ phosphoryl group). The corresponding deoxytetranucleotide, in which the 2′-OH groups are replaced by H and the uracil (U) is replaced by thymine (T), is abbreviated d(pApTpCpG) or d(pATCG). (*b*) Schematic representation of pAUCG. A vertical line denotes a ribose residue, the attached base is indicated by a single letter, and a diagonal line flanking an optional "p" represents a phosphodiester bond. The atom numbers for the ribose residue may be omitted. The equivalent representation of d(pATCG) differs only by the absence of the 2′-OH group and by the replacement of U by T.

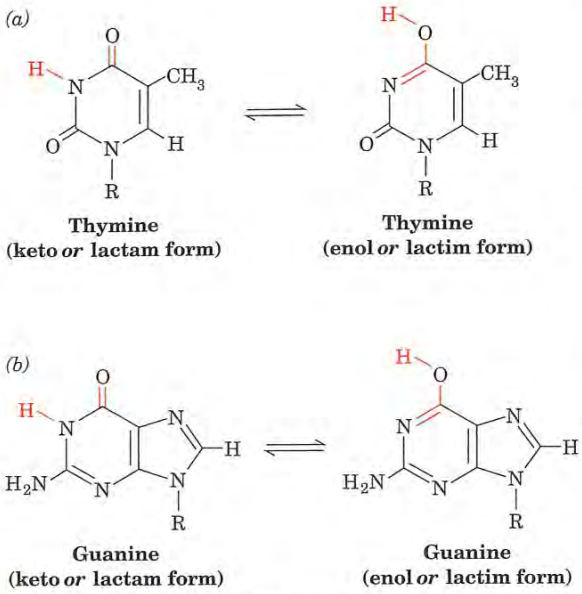


Figure 3-4 | Tautomeric forms of bases. Some of the possible tautomeric forms of (*a*) thymine and (*b*) guanine are shown. Cytosine and adenine can undergo similar proton shifts.

is not linked to another nucleotide is called the **5′ end,** and the terminal residue whose C3′ is not linked to another nucleotide is called the **3′ end.** By convention, the sequence of nucleotide residues in a nucleic acid is written, left to right, from the 5′ end to the 3′ end.

The properties of a polymer such as a nucleic acid may be very different from the properties of the individual units, or **monomers,** before polymerization. As the size of the polymer increases from **dimer, trimer, tetramer,** and so on through **oligomer** (Greek: *oligo,* few), physical properties such as charge and solubility may change. In addition, *a polymer of nonidentical residues has a property that its component monomers lack—namely, it contains information in the form of its sequence of residues.*

Chargaff's Rules Describe the Base Composition of DNA.

Although there appear to be no rules governing the nucleotide composition of typical RNA molecules, DNA has equal numbers of adenine and thymine residues (A = T) and equal numbers of guanine and cytosine residues (G = C). These relationships, known as **Chargaff's rules,** were discovered in the late 1940s by Erwin Chargaff, who devised the first reliable quantitative methods for the compositional analysis of DNA.

DNA's base composition varies widely among different organisms. It ranges from ~25 to 75 mol % G + C in different species of bacteria. However, it is more or less constant among related species; for example, in mammals G + C ranges from 39 to 46%. The significance of Chargaff's rules was not immediately appreciated, but we now know that the structural basis for the rules derives from DNA's double-stranded nature.

B | DNA Forms a Double Helix

The determination of the structure of DNA by James Watson and Francis Crick in 1953 is often said to mark the birth of modern molecular biology. The **Watson–Crick structure** of DNA not only provided a model of what is arguably the central molecule of life, it also suggested the molecular mechanism of heredity. Watson and Crick's accomplishment, which is ranked as one of science's major intellectual achievements, was based in part on two pieces of evidence in addition to Chargaff's rules: the correct tautomeric forms of the bases and indications that DNA is a helical molecule.

The purine and pyrimidine bases of nucleic acids can assume different tautomeric forms (**tautomers** are easily converted isomers that differ only in hydrogen positions; Fig. 3-4). X-Ray, nuclear magnetic resonance (NMR), and spectroscopic investigations have firmly established that the nucleic acid bases are overwhelmingly in the keto tautomeric forms shown in Fig. 3-3. In 1953, however, this was not generally appreciated. Information about the dominant tautomeric forms was provided by Jerry Donohue, an office mate of Watson and Crick and an expert on the X-ray structures of small organic molecules.

Figure 3-5 | An X-ray diffraction photograph of a vertically oriented DNA fiber. This photograph, taken by Rosalind Franklin, provided key evidence for the elucidation of the Watson–Crick structure. The central X-shaped pattern indicates a helix, whereas the heavy black arcs at the top and bottom of the diffraction pattern reveal the spacing of the stacked bases (3.4 Å). [Courtesy of Maurice Wilkins, King's College, London.]

Evidence that DNA is a helical molecule was provided by an X-ray diffraction photograph of a DNA fiber taken by Rosalind Franklin (Fig. 3-5). The appearance of the photograph enabled Crick, an X-ray crystallographer by training, to deduce (a) that DNA is a helical molecule and (b) that its planar aromatic bases form a stack that is parallel to the fiber axis.

The limited structural information, along with Chargaff's rules, provided few clues to the structure of DNA; Watson and Crick's model sprang mostly from their imaginations and model-building studies. Once the Watson–Crick

model had been published, however, its basic simplicity combined with its obvious biological relevance led to its rapid acceptance. Later investigations have confirmed the general accuracy of the Watson–Crick model, although its details have been modified.

The Watson–Crick model of DNA has the following major features:

1. Two polynucleotide chains wind around a common axis to form a **double helix** (Fig. 3-6).
2. The two strands of DNA are **antiparallel** (run in opposite directions), but each forms a right-handed helix. (The difference between a right-handed and a left-handed helix is shown in Fig. 3-7.)

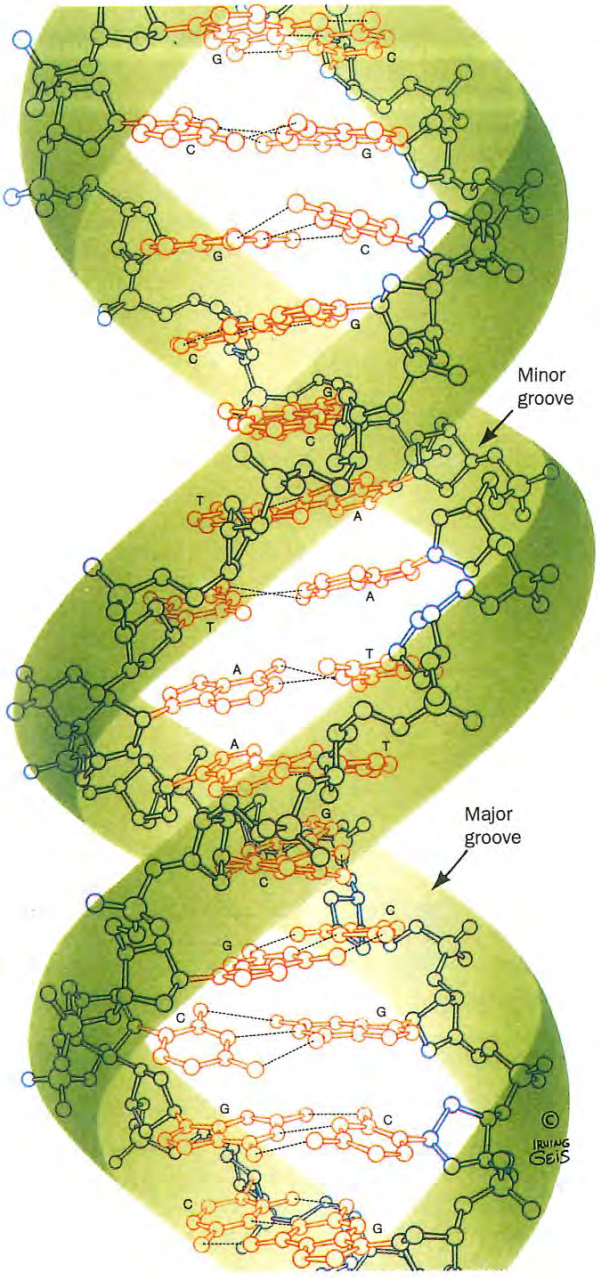


Figure 3-6 | Three-dimensional structure of DNA. The repeating helix is based on the structure of the self-complementary dodecamer d(CGCGAATTCGCG) determined by Richard Dickerson and Horace Drew. The view in this ball-and-stick model is perpendicular to the helix axis. The sugar–phosphate backbones (*blue, with green ribbon outlines*) wind around the periphery of the molecule. The bases (*red*) form hydrogen-bonded pairs that occupy the core. H atoms have been omitted for clarity. The two strands run in opposite directions. [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.] **See Interactive Exercise 1 and Kinemage Exercise 2-1.**

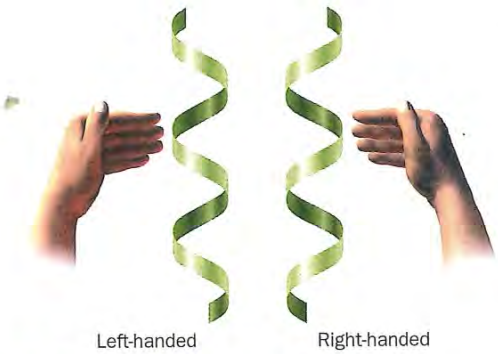


Figure 3-7 | Diagrams of left- and right-handed helices. In each case, the fingers curl in the direction the helix turns when the thumb points in the direction the helix rises. Note that the handedness is retained when the helices are turned upside down.

3. The bases occupy the core of the helix and sugar–phosphate chains run along the periphery, thereby minimizing the repulsions between charged phosphate groups. The surface of the double helix contains two grooves of unequal width: the **major** and **minor grooves.**
4. Each base is hydrogen bonded to a base in the opposite strand to form a planar **base pair.** The Watson–Crick structure can accommodate only two types of base pairs. Each adenine residue must pair with a thymine residue and vice versa, and each guanine residue must pair with a cytosine residue and vice versa (Fig. 3-8). These hydrogen-bonding interactions, a phenomenon known as **complementary base pairing,** result in the specific association of the two chains of the double helix.

The Watson–Crick structure can accommodate any sequence of bases on one polynucleotide strand if the opposite strand has the complementary

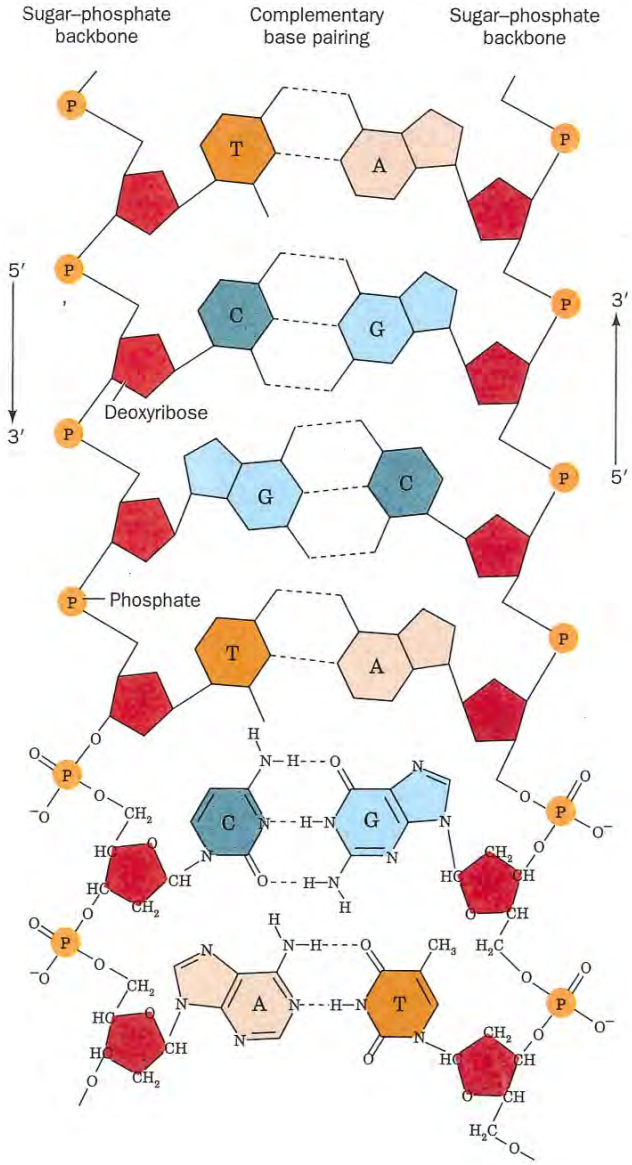
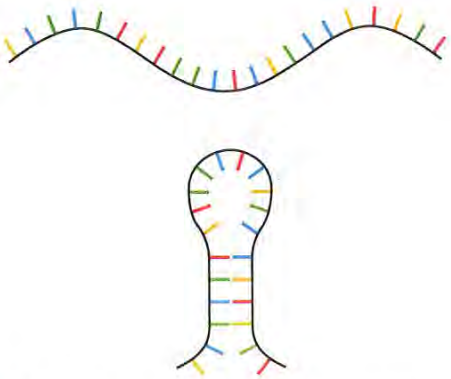


Figure 3-8 | Complementary strands of DNA. Two polynucleotide chains associate by base pairing to form double-stranded DNA. A pairs with T, and G pairs with C by forming specific hydrogen bonds. **See Kinemage Exercise 2-2.**

base sequence. This immediately accounts for Chargaff's rules. More importantly, it suggests that *each DNA strand can act as a* ***template*** *for the synthesis of its complementary strand and hence that hereditary information is encoded in the sequence of bases on either strand.*

Most DNA Molecules Are Large. The extremely large size of DNA molecules is in keeping with their role as the repository of a cell's genetic information. Of course, an organism's **genome,** its unique DNA content, may be allocated among several **chromosomes** (Greek: *chromos,* color + *soma,* body), each of which contains a separate DNA molecule. Note that many organisms are **diploid;** that is, they contain two equivalent sets of chromosomes, one from each parent. Their content of unique **(haploid)** DNA is half their total DNA. For example, humans are diploid organisms that carry 46 chromosomes per cell; their haploid number is therefore 23.

Because of their great lengths, DNA molecules are described in terms of the number of base pairs **(bp)** or thousands of base pairs **(kilobase pairs,** or **kb).** Naturally occurring DNAs vary in length from ~5 kb in small DNA-containing viruses to well over 250,000 kb in the largest mammalian chromosomes. Although DNA molecules are long and relatively stiff, they are not completely rigid. We shall see in Chapter 24 that the DNA double helix forms coils and loops when it is packaged inside the cell. Furthermore, depending on the nucleotide sequence, DNA may adopt slightly different helical conformations. Finally, in the presence of other cellular components, the DNA may bend sharply or the two strands may partially unwind.



■ **Figure 3-9 | Formation of a stem–loop structure.** Base pairing between complementary sequences within an RNA strand allows the polynucleotide to fold back on itself.

C | RNA Is a Single-Stranded Nucleic Acid

Single-stranded DNA is rare, occurring mainly as the hereditary material of certain viruses. In contrast, RNA occurs primarily as single strands, which usually form compact structures rather than loose extended chains (double-stranded RNA is the hereditary material of certain viruses). An RNA strand—which is identical to a DNA strand except for the presence of 2′-OH groups and the substitution of uracil for thymine—can base-pair with a complementary strand of RNA or DNA. As expected, A pairs with U (or T in DNA), and G with C. Base pairing often occurs intramolecularly, giving rise to **stem–loop** structures (Fig. 3-9) or, when loops interact with each other, to more complex structures.

The intricate structures that can potentially be adopted by single-stranded RNA molecules provide additional evidence that RNA can do more than just store and transmit genetic information. Numerous investigations have found that certain RNA molecules can specifically bind small organic molecules and can catalyze reactions involving those molecules. These findings provide substantial support for theories that *many of the processes essential for life began through the chemical versatility of small polynucleotides* (a situation known as the **RNA world**). We will further explore RNA structure and function in Section 24-2C.

■ **CHECK YOUR UNDERSTANDING**

Describe the double-helical structure of DNA. List the structural differences between DNA and RNA.

3 Overview of Nucleic Acid Function

LEARNING OBJECTIVE

- Understand that genetic information is contained in the sequence of nucleotides in DNA and can be expressed through its transcription into RNA, which is then translated into protein.

DNA is the carrier of genetic information in all cells and in many viruses. Yet a period of over 75 years passed from the time the laws of inheritance were discovered by Gregor Mendel until the biological role of DNA was elucidated. Even now, many details of how genetic information is expressed and transmitted to future generations are still unclear.

Figure 3-10 | Transformed pneumococci. The large colonies are virulent pneumococci that resulted from the transformation of nonpathogenic pneumococci (smaller colonies) by DNA extracted from the virulent strain. We now know that this DNA contained a gene that was defective in the nonpathogenic strain. [From Avery, O.T., MacLeod, C.M., and McCarty, M., *J. Exp. Med.* **79,** 153 (1944). Copyright © 1944 by Rockefeller University Press.]

Mendel's work with garden peas led him to postulate that an individual plant contains a pair of factors (which we now call **genes**), one inherited from each parent. But Mendel's theory of inheritance, reported in 1866, was almost universally ignored by his contemporaries, whose knowledge of anatomy and physiology provided no basis for its understanding. Eventually, genes were hypothesized to be part of chromosomes, and the pace of genetic research greatly accelerated.

A | DNA Carries Genetic Information

Until the 1940s, it was generally assumed that genes were made of protein, since proteins were the only biochemical entities that, at the time, seemed complex enough to serve as agents of inheritance. Nucleic acids, which had first been isolated in 1869 by Friedrich Miescher, were believed to have monotonously repeating nucleotide sequences and were therefore unlikely candidates for transmitting genetic information.

It took the efforts of Oswald Avery, Colin MacLeod, and Maclyn McCarty to demonstrate that DNA carries genetic information. Their experiments, completed in 1944, showed that DNA—not protein—extracted from a virulent (pathogenic) strain of the bacterium *Diplococcus pneumoniae* was the substance that **transformed** (permanently changed) a nonpathogenic strain of the organism to the virulent strain (Fig. 3-10). Avery's discovery was initially greeted with skepticism, but it influenced Erwin Chargaff, whose rules (Section 3-2A) led to subsequent models of the structure and function of DNA.

The double-stranded, or duplex, nature of DNA facilitates its **replication.** When a cell divides, each DNA strand acts as a template for the assembly of its complementary strand (Fig. 3-11). Consequently, every progeny cell contains a complete DNA molecule (or a complete set of DNA molecules in organisms whose genomes contain more than one chromosome). Each DNA molecule consists of one parental strand and one daughter strand. Daughter strands are synthesized by the stepwise polymerization of nucleotides that specifically pair with bases on the parental strands. The mechanism of replication, while straightforward in principle, is exceedingly complex in the cell, requiring a multitude of cellular factors to proceed with fidelity and efficiency, as we shall see in Chapter 25.

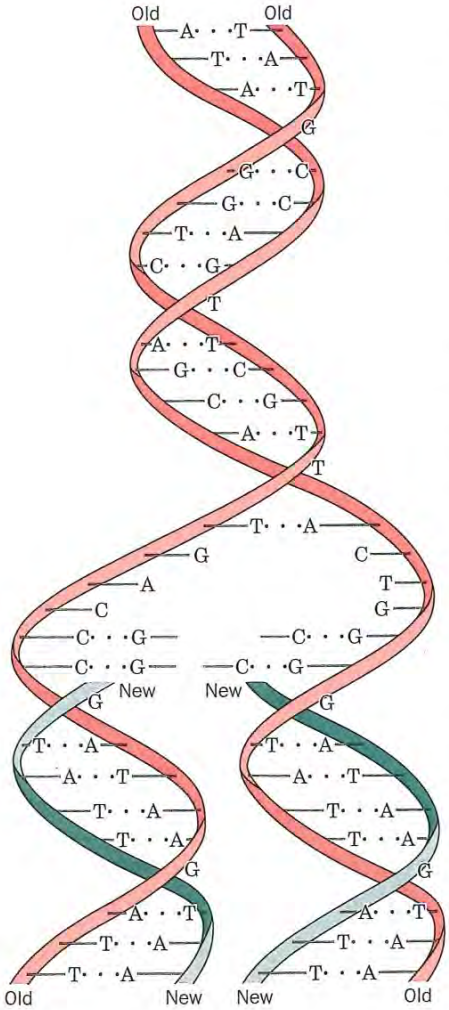


Figure 3-11 | DNA replication. Each strand of parental DNA (*red*) acts as a template for the synthesis of a complementary daughter strand (*green*). Thus, the resulting double-stranded molecules are identical.

B | Genes Direct Protein Synthesis

The question of how sequences of nucleotides control the characteristics of organisms took some time to be answered. In experiments with the mold *Neurospora crassa* in the 1940s, George Beadle and Edward Tatum found that *there is a specific connection between genes and enzymes, the one gene–one enzyme theory.* Beadle and Tatum showed that mutant varieties of *Neurospora* that were generated by irradiation with X-rays required additional nutrients in order to grow. Presumably, the offspring of the radiation-damaged cells lacked the specific enzymes necessary to synthesize those nutrients.

The link between DNA and enzymes (nearly all of which are proteins) is RNA. *The DNA of a gene is **transcribed** to produce an RNA molecule that is complementary to the DNA. The RNA sequence is then **translated** into the corresponding sequence of amino acids to form a protein* (Fig. 3-12). These transfers of biological information are summarized in the so-called **central dogma of molecular biology** formulated by Crick in 1958 (Fig. 3-13).

Just as the daughter strands of DNA are synthesized from free deoxynucleoside triphosphates that pair with bases in the parent DNA strand, RNA strands are synthesized from free ribonucleoside triphosphates that pair with the complementary bases in one DNA strand of a gene (transcription is described in greater detail in Chapter 26). The RNA that corresponds to a protein-coding gene (called **messenger RNA,** or **mRNA**) makes its way to a **ribosome,** an organelle that is itself composed largely of RNA **(ribosomal RNA,** or **rRNA).** At the ribosome, each set of three nucleotides in the mRNA pairs with three complementary nucleotides in a small RNA molecule called a **transfer RNA,** or **tRNA** (Fig. 3-14). Attached to each tRNA molecule is its corresponding amino acid. The ribosome catalyzes the joining of amino acids, which are the monomeric units of proteins (protein synthesis is described in detail in Chapter 27). Amino acids are added to the growing protein chain according to the order in which the tRNA molecules bind to the mRNA. Since the nucleotide sequence of the mRNA in turn reflects the sequences of nucleotides in the gene, DNA directs the synthesis of proteins. It follows that alterations to the genetic material of an organism **(mutations)** may manifest themselves as proteins with altered structures and functions.

Using techniques that are described in the following sections and in other parts of this book, researchers can compile a catalog of all the

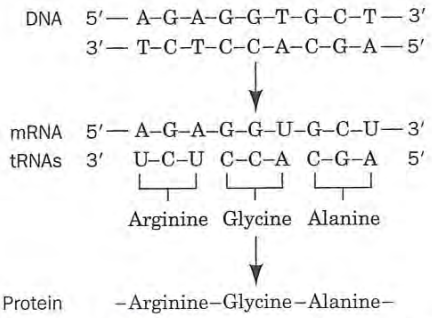


Figure 3-12 | Transcription and translation. One strand of DNA directs the synthesis of messenger RNA (mRNA). The base sequence of the transcribed RNA is complementary to that of the DNA strand. The message is translated when transfer RNA (tRNA) molecules align with the mRNA by complementary base pairing between three-nucleotide segments known as **codons.** Each tRNA carries a specific amino acid. These amino acids are covalently joined to form a protein. Thus, the sequence of bases in DNA specifies the sequence of amino acids in a protein.

See Guided Exploration 1
Overview of transcription and translation.

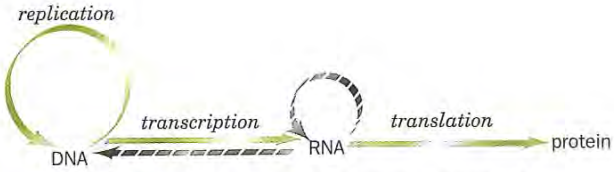


Figure 3-13 | The central dogma of molecular biology. Solid arrows indicate the types of information transfers that occur in all cells: DNA directs its own replication to produce new DNA molecules; DNA is transcribed into RNA; RNA is translated into protein. Dashed lines represent information transfers that occur only in certain organisms.

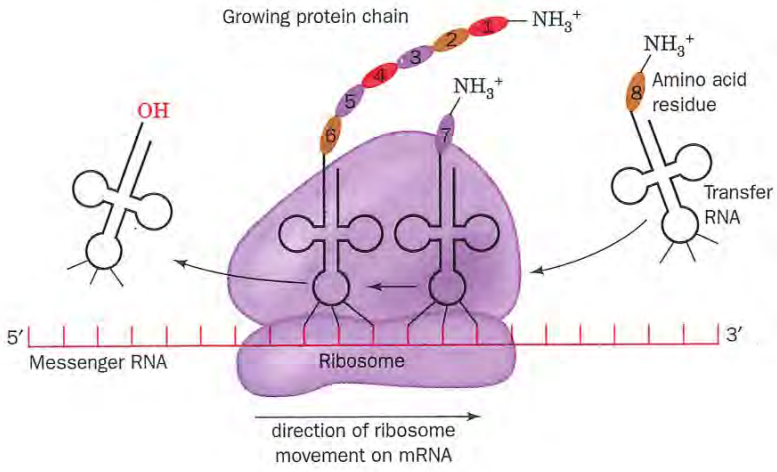


Figure 3-14 | Translation. tRNA molecules with their attached amino acids bind to complementary three-nucleotide sequences (codons) on mRNA. The ribosome facilitates the alignment of the tRNA and the mRNA, and it catalyzes the joining of amino acids to produce a protein chain. When a new amino acid is added, the preceding tRNA is ejected, and the ribosome proceeds along the mRNA.

information encoded in an organism's DNA. The study of the genome's size, organization, and gene content is known as **genomics.** By analogy, **transcriptomics** refers to the study of gene expression, which focuses on the set of mRNA molecules, or **transcriptome**, that is transcribed from DNA under any particular set of circumstances. Finally, **proteomics** is the study of the proteins (the **proteome**) produced as a result of transcription and translation. Although an organism's genome remains essentially unchanged throughout its lifetime, its transcriptome and proteome may vary significantly among different types of tissues, developmental stages, and environmental conditions.

CHECK YOUR UNDERSTANDING

Summarize the central dogma of molecular biology.

4 Nucleic Acid Sequencing

LEARNING OBJECTIVES

- Understand why restriction endonucleases are useful for generating DNA fragments.
- Understand the steps required to sequence DNA by the chain-terminator method.
- Understand what kinds of information have been provided by sequencing the human genome.
- Understand that changes in DNA allow evolution to occur.

Much of our current understanding of protein structure and function rests squarely on information gleaned not from the proteins themselves, but indirectly from their genes. *The ability to determine the sequence of nucleotides in nucleic acids has made it possible to deduce the amino acid sequences of their encoded proteins and, to some extent, the structures and functions of those proteins. Nucleic acid sequencing has also revealed information about the regulation of genes.* Certain portions of genes that are not actually transcribed into RNA nevertheless influence how often a gene is transcribed and translated, that is, **expressed.** Moreover, efforts to elucidate the sequences in hitherto unmapped regions of DNA have led to the discovery of new genes and new regulatory elements. *Once in hand, a nucleic acid sequence can be duplicated, modified, and expressed, making it possible to study proteins that could not otherwise be obtained in useful quantities.* In this section, we describe how nucleic acids are sequenced and what information the sequences may reveal. In the following section, we discuss the manipulation of purified nucleic acid sequences for various purposes.

The overall strategy for sequencing any polymer of nonidentical units is

1. Cleave the polymer into specific fragments that are small enough to be fully sequenced.
2. Determine the sequence of residues in each fragment.
3. Determine the order of the fragments in the original polymer by repeating the preceding steps using a degradation procedure that yields a set of fragments that overlap the cleavage points in the first step.

The first efforts to sequence RNA used nonspecific enzymes to generate relatively small fragments whose nucleotide composition was then determined by partial digestion with an enzyme that selectively removed nucleotides from one end or the other (Fig. 3-15). Sequencing RNA in this manner was tedious and time-consuming. Using such methods, it took Robert Holley seven years to determine the sequence of a 76-residue tRNA molecule.

After 1975, dramatic progress was made in nucleic acid sequencing technology. The advances were made possible by the discovery of enzymes that could cleave DNA at specific sites and by the development of rapid sequencing techniques for DNA. The advent of modern molecular cloning techniques (Section 3-5) also made it possible to produce sufficient quantities of specific DNA to be sequenced. These cloning techniques are necessary because most specific DNA sequences are normally present in a genome in only a single copy.

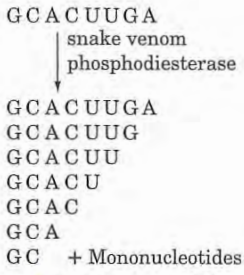


Figure 3-15 | Determining the sequence of an oligonucleotide using nonspecific enzymes. The oligonucleotide is partially digested with snake venom phosphodiesterase, which breaks the phosphodiester bonds between nucleotide residues, starting at the 3′ end of the oligonucleotide. The result is a mixture of fragments of all lengths, which are then separated. Comparing the base composition of a pair of fragments that differ in length by one nucleotide establishes the identity of the 3′-terminal nucleotide in the larger fragment. Analysis of each pair of fragments reveals the sequence of the original oligonucleotide.

A | Restriction Endonucleases Cleave DNA at Specific Sequences

Many bacteria are able to resist infection by **bacteriophages** (viruses that are specific for bacteria) by virtue of a **restriction–modification system.** The bacterium modifies certain nucleotides in specific sequences of its own DNA by adding a methyl ($—CH_3$) group in a reaction catalyzed by a **modification methylase.** A **restriction endonuclease,** which recognizes the same nucleotide sequence as does the methylase, cleaves any DNA that has not been modified on at least one of its two strands. (An **endonuclease** cleaves a nucleic acid within the polynucleotide strand; an **exonuclease** cleaves a nucleic acid by removing one of its terminal residues.) This system destroys foreign (phage) DNA containing a recognition site that has not been modified by methylation. The host DNA is always at least half methylated, because although the daughter strand is not methylated until shortly after it is synthesized, the parental strand to which it is paired is already modified (and thus protects both strands of the DNA from cleavage by the restriction enzyme).

Type II restriction endonucleases are particularly useful in the laboratory. These enzymes cleave DNA within the four- to eight-base sequence that is recognized by their corresponding modification methylase. (Type I and Type III restriction endonucleases cleave DNA at sites other than their recognition sequences.) Over 3000 Type II restriction enzymes with over 200 different recognition sequences have been characterized. Some of the more widely used restriction enzymes are listed in Table 3-2. A

Table 3-2 Recognition and Cleavage Sites of Some Restriction Enzymes

Enzyme	Recognition Sequence[a]	Microorganism
*Alu*I	AG↓CT	*Arthrobacter luteus*
*Bam*HI	G↓GATCC	*Bacillus amyloliquefaciens* H
*Bgl*I	GCCNNNNN↓NGGC	*Bacillus globigii*
*Bgl*II	A↓GATCT	*Bacillus globigii*
*Eco*RI	G↓AATTC	*Escherichia coli* RY13
*Eco*RII	↓CC(A/T)GG	*Escherichia coli* R245
*Eco*RV	GAT↓ATC	*Escherichia coli* J62 pLG74
*Hae*II	RGCGC↓Y	*Haemophilus aegyptius*
*Hae*III	GG↓CC	*Haemophilus aegyptius*
*Hin*dIII	A↓AGCTT	*Haemophilus influenzae* R_d
*Hpa*II	C↓CGG	*Haemophilus parainfluenzae*
*Msp*I	C↓CGG	*Moraxella* species
*Pst*I	CTGCA↓G	*Providencia stuartii* 164
*Pvu*II	CAG↓CTG	*Proteus vulgaris*
*Sal*I	G↓TCGAC	*Streptomyces albus* G
*Taq*I	T↓CGA	*Thermus aquaticus*
*Xho*I	C↓TCGAG	*Xanthomonas holcicola*

[a]The recognition sequence is abbreviated so that only one strand, reading 5′ to 3′, is given. The cleavage site is represented by an arrow (↓). R, Y, and N represent a purine nucleotide, a pyrimidine nucleotide, and any nucleotide, respectively.

Source: Roberts, R.J. and Macelis, D., REBASE—the restriction enzyme database, http://rebase.neb.com.

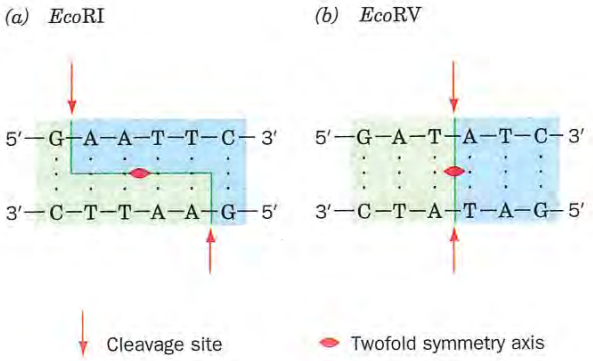


Figure 3-16 | Restriction sites. The recognition sequences for Type II restriction endonucleases are palindromes, sequences with a twofold axis of symmetry. (*a*) Recognition site for *Eco*RI, which generates DNA fragments with sticky ends. (*b*) Recognition site for *Eco*RV, which generates blunt-ended fragments.

restriction enzyme is named by the first letter of the genus and the first two letters of the species of the bacterium that produced it, followed by its serotype or strain designation, if any, and a roman numeral if the bacterium contains more than one type of restriction enzyme. For example, *Eco*RI is produced by *E. coli* strain RY13.

Interestingly, most Type II restriction endonucleases recognize and cleave palindromic DNA sequences. A **palindrome** is a word or phrase that reads the same forward or backward. Two examples are "refer" and "Madam, I'm Adam." In a palindromic DNA segment, the sequence of nucleotides is the same in each strand, and the segment is said to have twofold symmetry (Fig. 3-16). Most restriction enzymes cleave the two strands of DNA at positions that are staggered, producing DNA fragments with complementary single-strand extensions. Restriction fragments with such **sticky ends** can associate by base pairing with other restriction fragments generated by the same restriction enzyme. Some restriction endonucleases cleave the two strands of DNA at the symmetry axis to yield restriction fragments with fully base-paired **blunt ends.**

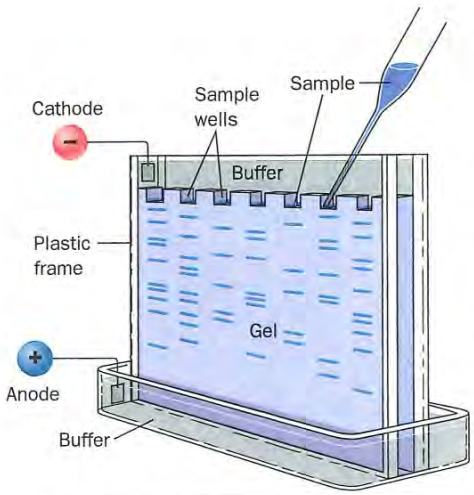


Figure 3-17 | Apparatus for gel electrophoresis. Samples are applied in slots at the top of the gel and electrophoresed in parallel lanes. Negatively charged molecules such as DNA migrate through the gel matrix toward the anode in response to an applied electric field. Because smaller molecules move faster, the molecules in each lane are separated according to size. Following electrophoresis, the separated molecules may be visualized by staining, fluorescence, or a radiographic technique.

B | Electrophoresis Separates Nucleic Acids According to Size

Treating a DNA molecule with a restriction endonuclease produces a series of precisely defined fragments that can be separated according to size. **Gel electrophoresis** is commonly used for the separation. In principle, a charged molecule moves in an electric field with a velocity proportional to its overall charge density, size, and shape. For molecules with a relatively homogeneous composition (such as nucleic acids), shape and charge density are constant, so the velocity depends primarily on size. Electrophoresis is carried out in a gel-like matrix, usually made from **agarose** (carbohydrate polymers that form a loose mesh) or **polyacrylamide** (a more rigid cross-linked synthetic polymer). The gel is typically held between two glass plates (Fig. 3-17). The molecules to be separated are applied to one end of the gel, and the molecules move through the pores in the matrix under the influence of an electric field. Smaller molecules move more rapidly through the gel and therefore migrate farther in a given time.

Following electrophoresis, the separated molecules may be visualized in the gel by an appropriate technique, such as addition of a stain that binds tightly to the DNA or by radioactive labeling. Depending on the dimensions of the gel and the visualization technique used, samples containing less than a nanogram of material can be separated and detected by gel electrophoresis. Several samples can be electrophoresed simultaneously. For example, the fragments obtained by digesting a DNA sample with

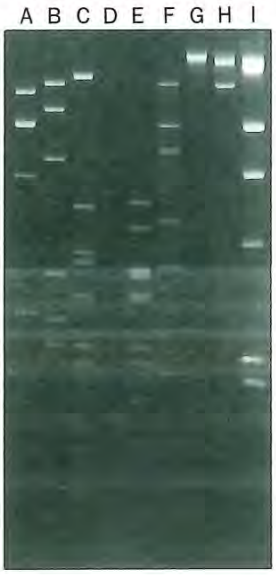


Figure 3-18 | Electrophoretogram of restriction digests. The plasmid pAgK84 has been digested with (A) *Bam*HI, (B) *Pst*I, (C) *Bgl*II, (D) *Hae*III, (E) *Hin*cII, (F) *Sac*I, (G) *Xba*I, and (H) *Hpa*I. Lane I contains bacteriophage λ digested with *Hin*dIII as a standard since these fragments have known sizes. The restriction fragments in each lane are made visible by fluorescence against a black background. [From Slota, J.E. and Farrand, S.F., *Plasmid* **8,** 180 (1982). Copyright © 1982 by Academic Press.]

different restriction endonucleases can be visualized side by side (Fig. 3-18). The sizes of the various fragments can be determined by comparing their electrophoretic mobilities to the mobilities of fragments of known size.

C | DNA Is Sequenced by the Chain-Terminator Method

Here we discuss the most commonly used procedure for sequencing DNA, the **chain-terminator method,** which was devised by Frederick Sanger. The first step in this procedure is to obtain single polynucleotide strands. Complementary DNA strands can be separated by heating, which breaks the hydrogen bonds between bases. Next, polynucleotide fragments that terminate at positions corresponding to each of the four nucleotides are generated. Finally, the fragments are separated and detected.

The Chain-Terminator Method Uses DNA Polymerase. The chain-terminator method (also called the **dideoxy method**) uses an *E. coli* enzyme to make complementary copies of the single-stranded DNA being sequenced. The enzyme is a fragment of **DNA polymerase I,** one of the enzymes that participates in replication of bacterial DNA (Section 25-2A). Using the single DNA strand as a template, DNA polymerase I assembles the four deoxynucleoside triphosphates **(dNTPs),** dATP, dCTP, dGTP, and dTTP, into a complementary polynucleotide chain that it elongates in the 5′→3′ direction (Fig. 3-19).

DNA polymerase I can sequentially add deoxynucleotides only to the 3′ end of a polynucleotide. Hence, replication is initiated in the presence of a

See Guided Exploration 2
DNA sequence determination by the chain-terminator method.

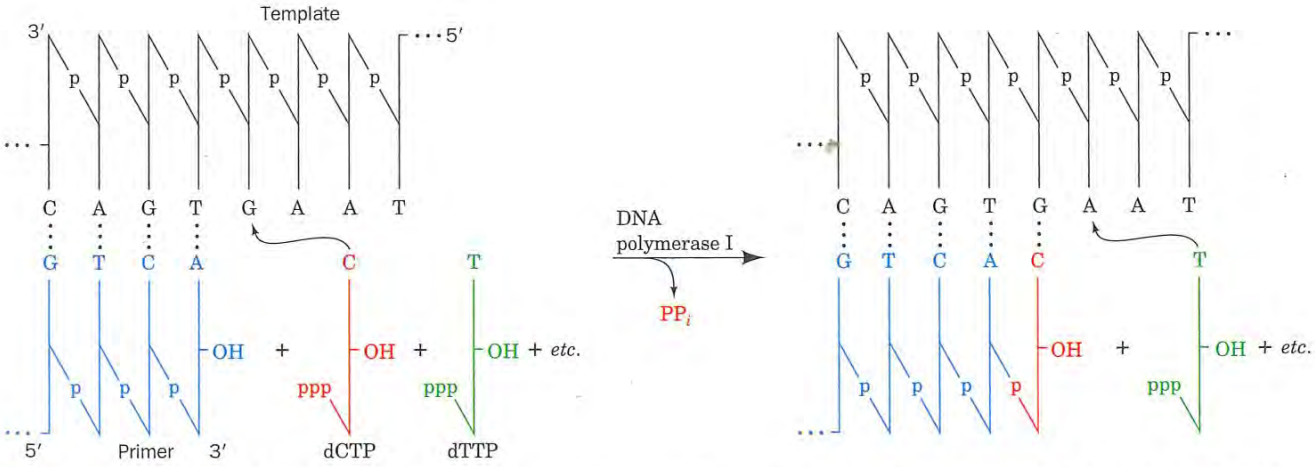


Figure 3-19 | Action of DNA polymerase I. Using a single DNA strand as a template, the enzyme elongates the primer by stepwise addition of complementary nucleotides. Incoming nucleotides pair with bases on the template strand and are joined to the growing polynucleotide strand in the 5′ → 3′ direction. The polymerase-catalyzed reaction requires a free 3′-OH group on the growing strand. **Pyrophosphate** ($P_2O_7^{4-}$; PP_i) is released with each nucleotide addition.

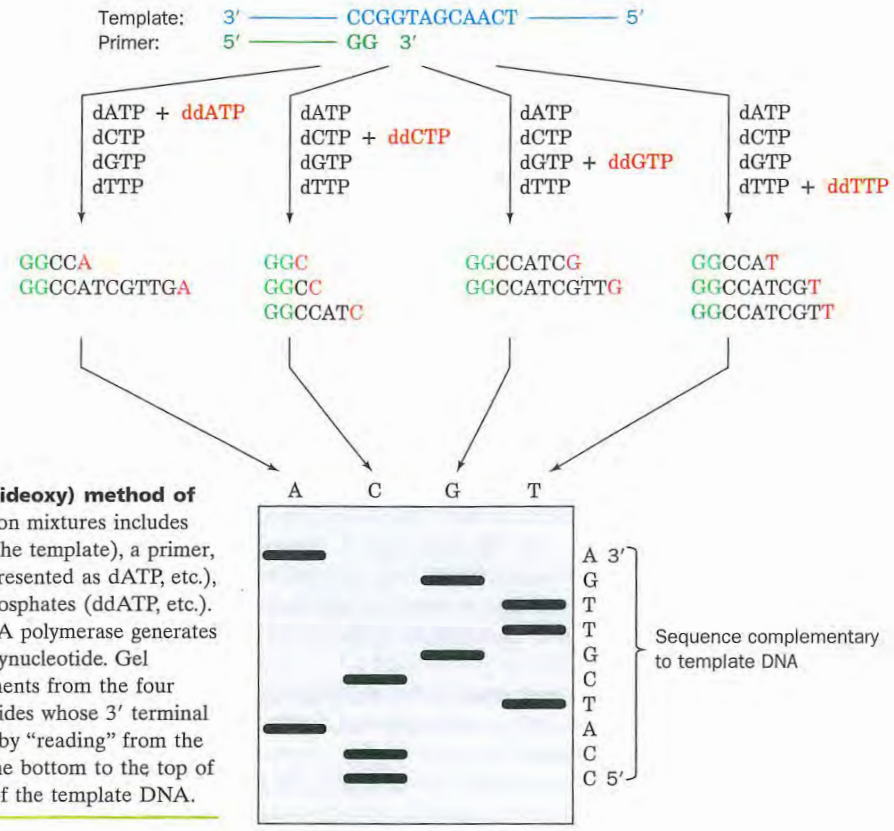
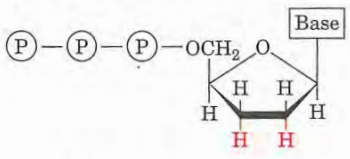


Figure 3-20 | The chain-terminator (dideoxy) method of DNA sequencing. Each of the four reaction mixtures includes the single-stranded DNA to be sequenced (the template), a primer, the four deoxynucleoside triphosphates (represented as dATP, etc.), and one of the four dideoxynucleoside triphosphates (ddATP, etc.). Extension of the primer by the action of DNA polymerase generates stretches of DNA terminating with a dideoxynucleotide. Gel electrophoresis in parallel lanes of the fragments from the four reaction mixtures yields a set of polynucleotides whose 3′ terminal residues are known. The sequence obtained by "reading" from the smallest fragment to the largest (i.e., from the bottom to the top of the gel) is complementary to the sequence of the template DNA.

short polynucleotide (a **primer**) that is complementary to the 3′ end of the template DNA and thus becomes the 5′ end of the new strand. The primer base-pairs with the template strand, and nucleotides are sequentially added to the 3′ end of the primer. If the DNA being sequenced is a restriction fragment, as it usually is, it begins and ends with a restriction site. The primer can therefore be a short DNA segment with the sequence of this restriction site.

DNA Synthesis Terminates after Specific Bases. In the chain-terminator technique (Fig. 3-20), the DNA to be sequenced is incubated with DNA polymerase I, a suitable primer, and the four dNTP **substrates** (reactants in enzymatic reactions) for the polymerization reaction. The reaction mixture also includes a "tagged" compound, either one of the dNTPs or the primer. The tag, which may be a radioactive isotope (e.g., ^{32}P) or a fluorescent label, permits the products of the polymerase reaction to be easily detected.

The key component of the reaction mixture is a small amount of a **2′,3′-dideoxynucleoside triphosphate (ddNTP),**



2′,3′-Dideoxynucleoside triphosphate

which lacks the 3′-OH group of deoxynucleotides. *When the dideoxy analog is incorporated into the growing polynucleotide in place of the corresponding normal nucleotide, chain growth is terminated because addition of the next nucleotide requires a free 3′-OH.* By using only a small amount of the ddNTP, a series of truncated chains is generated, each of which ends with the dideoxy analog at one of the positions occupied by the corresponding base.

Relatively modest sequencing tasks use four reaction mixtures, each with a different ddNTP, and the reaction products are electrophoresed in parallel lanes. The lengths of the truncated chains indicate the positions where the dideoxynucleotide was incorporated. Thus, the sequence of the replicated strand can be directly read from the gel (Fig. 3-21). The gel must have sufficient resolving power to separate fragments that differ in length by only one nucleotide. Two sets of gels, one run for a longer time than the other, can be used to obtain the sequence of up to 800 bases of DNA. Note that the sequence obtained by the chain-terminator method is complementary to the DNA strand being sequenced.

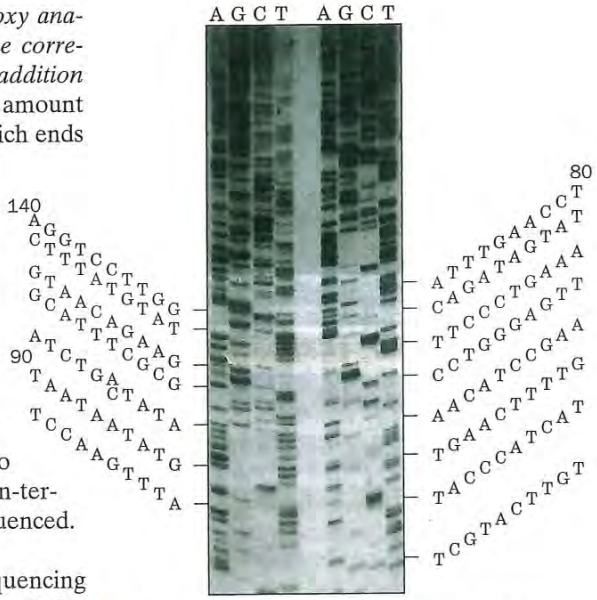


Figure 3-21 | An autoradiogram of a sequencing gel. The positions of radioactive DNA fragments produced by the chain-terminator method were visualized by laying X-ray film over the gel after electrophoresis. A second loading of the gel (the four lanes at right) was made 90 min after the initial loading in order to obtain the sequences of the smaller fragments. The deduced sequence of 140 nucleotides is written along the side. [From Hindley, J., DNA sequencing, *in* Work, T.S. and Burdon, R.H. (Eds.), *Laboratory Techniques in Biochemistry and Molecular Biology,* Vol. 10, p. 82, Elsevier (1983). Used by permission.]

Large Sequencing Projects Are Automated. Large-scale sequencing operations are accelerated by automation. In a variation of the chain-terminator method, the primers used in the four chain-extension reactions are each linked to a different fluorescent dye. The separately reacted mixtures are combined and subjected to gel electrophoresis in a single lane. As each fragment exits the bottom of the gel, its terminal base is identified by its characteristic fluorescence (Fig. 3-22) with an error rate of ~1%.

In the most advanced systems, the sequencing gel is contained in an array of up to 96 capillary tubes (rather than in a slab-shaped apparatus), sample preparation and sample loading are performed by robotic systems, and electrophoresis and data analysis are fully automated. These systems can simultaneously sequence 96 DNA samples averaging ~600 bases each with a turnaround time of ~2.5 hr and hence can identify up to 550,000 bases per day—all with only ~15 min of human attention (a skilled

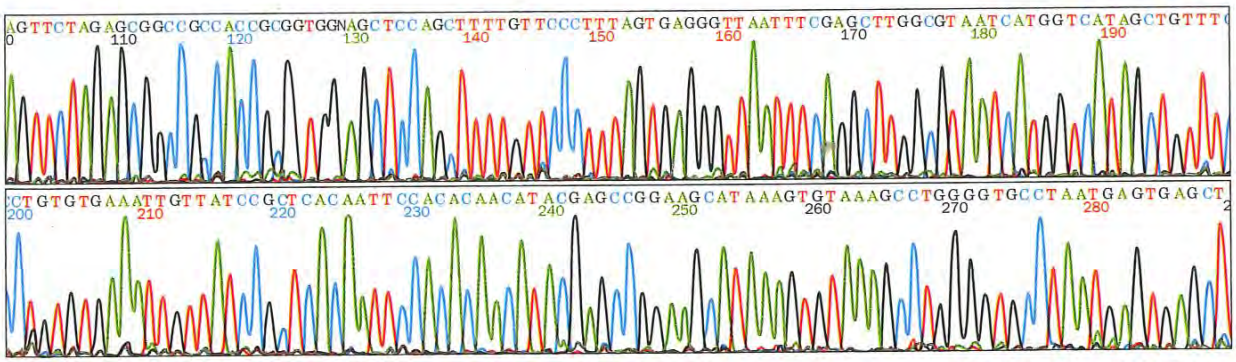


Figure 3-22 | Automated DNA sequencing. In this variant of the technique, a different fluorescent dye is attached to the primer in each of the four reaction mixtures in the chain-terminator procedure. The four reaction mixtures are combined for electrophoresis. Each of the four colored curves therefore represents the electrophoretic pattern of fragments containing one of the dideoxynucleotides: Green, red, black, and blue peaks correspond to fragments ending in ddATP, ddTTP, ddGTP, and ddCTP, respectively. The 3′-terminal base of each oligonucleotide, identified by the fluorescence of its gel band, is indicated by a single letter (A, T, G, or C). This portion of the readout corresponds to nucleotides 100–290 of the DNA segment being sequenced. [Courtesy of Mark Adams, The Institute for Genomic Research, Rockville, Maryland.]

BOX 3-1 PATHWAYS OF DISCOVERY

Francis Collins and the Gene for Cystic Fibrosis

Francis S. Collins (1950–)

By the mid-twentieth century, the molecular basis of several human diseases was appreciated. For example, sickle-cell anemia (Section 7-1E) was known to be caused by an abnormal hemoglobin protein. Studies of sickle-cell hemoglobin eventually revealed the underlying genetic defect, a mutation in a hemoglobin gene. It therefore seemed possible to trace other diseases to defective genes. But for many genetic diseases, even those with well-characterized symptoms, no defective protein had yet been identified. One such disease was cystic fibrosis, which is characterized mainly by the secretion of thick mucus that obstructs the airways and creates an ideal environment for bacterial growth. Cystic fibrosis is the most common inherited disease in individuals of northern European descent, striking about 1 in 2500 newborns and leading to death by early adulthood due to irreversible lung damage. It was believed that identifying the molecular defect in cystic fibrosis would lead to better understanding of the disease and to the ability to design more effective treatments.

Enter Francis Collins, who began his career by earning a doctorate in physical chemistry but then enrolled in medical school to take part in the molecular biology revolution. As a physician-scientist, Collins developed methods for analyzing large stretches of DNA in order to home in on specific genes, including the one that, when mutated, causes cystic fibrosis. By analyzing the DNA of individuals with the disease (who had two copies of the defective gene) and of family members who were asymptomatic carriers (with one normal and one defective copy of the gene), Collins and his team localized the cystic fibrosis gene to the long arm of chromosome 7. They gradually closed in on a DNA segment that appears to be present in a number of mammalian species, which suggests that the segment contains an essential gene. The cystic fibrosis gene was finally identified in 1989. Collins had demonstrated the feasibility of identifying a genetic defect in the absence of other molecular information.

Once the cystic fibrosis gene was in hand, it was a relatively straightforward process to deduce the probable structure and function of the encoded protein, which turned out to be a membrane channel for chloride ions. When functioning normally, the protein helps regulate the ionic composition and viscosity of extracellular secretions. Discovery of the cystic fibrosis gene also made it possible to design tests to identify carriers so that they could take advantage of genetic counseling.

Throughout Collins' work on the cystic fibrosis gene and during subsequent hunts for the genes that cause neurofibromatosis and Huntington's disease, he was mindful of the ethical implications of the new science of molecular genetics. Collins has been a strong advocate for protecting the privacy of genetic information. At the same time, he recognizes the potential therapeutic use of such information. In his tenure as director of the human genome project, he was committed to making the results freely and immediately accessible, as a service to researchers and the individuals who might benefit from new therapies based on molecular genetics.

Riordan, J.R., Rommens, J.M., Kerem, B.-S., Alon, N., Rozmahel, R., Grzelczak, Z., Zielensky, J., Lok, S., Plavsic, N., Chou, J.-L., Drumm, M.L., Iannuzzi, M.C., Collins, F.S., and Tsui, L.-C., Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA, *Science* **245,** 1066–1073 (1989).

operator can identify only ~25,000 bases per year using the above-described manual methods). Sequencing the 3.2-billion-bp human genome required hundreds of such advanced sequencing systems.

Databases Store Nucleotide Sequences. The results of sequencing projects large and small are customarily deposited in online databases such as GenBank (see the Bioinformatics Exercises). Over 150 billion nucleotides in 80 million sequences have been recorded as of late 2006.

Nucleic acid sequencing has become so routine that directly determining a protein's amino acid sequence (Section 5-3) is generally far more time-consuming than determining the base sequence of its corresponding gene. In fact, nucleic acid sequencing is invaluable for studying genes whose products have not yet been identified. If the gene can be sequenced, the probable function of its protein product may be deduced by comparing the base sequence to those of genes whose products are already characterized (see Box 3-1).

D | Entire Genomes Have Been Sequenced

The advent of large-scale sequencing techniques brought to fruition the dream of sequencing entire genomes. However, the major technical hurdle in sequencing all the DNA in an organism's genome is not the DNA sequencing itself but, rather, assembling the tens of thousands to tens of millions of sequenced segments (depending on the size of the genome) into contiguous blocks and assigning them to their correct chromosomal positions. To do so required the development of automated sequencing protocols and mathematically sophisticated computer algorithms.

The first complete genome sequence to be determined, that of the bacterium *Haemophilus influenzae,* was reported in 1995 by Craig Venter. By mid-2007, the complete genome sequences of over 500 prokaryotes had been reported (with many more being determined) as well as those of dozens of eukaryotes, including humans, human pathogens, plants, and laboratory organisms (Table 3-3).

The determination of the ~3.2-billion-nucleotide human genome sequence was a gargantuan undertaking involving hundreds of scientists working in two groups, one led by Venter and the other by Francis Collins (Box 3-1), Eric Lander, and John Sulston. After over a decade of intense

Table 3-3 Some Sequenced Genomes

Organism	Genome Size (kb)	Number of Chromosomes
Mycoplasma genitalium (human parasite)	580	1
Rickettsia prowazekii (putative relative of mitochondria)	1,112	1
Haemophilus influenzae (human pathogen)	1,830	1
Escherichia coli (human symbiont)	4,639	1
Saccharomyces cerevisiae (baker's yeast)	11,700	16
Plasmodium falciparum (protozoan that causes malaria)	30,000	14
Caenorhabditis elegans (nematode)	97,000	6
Arabidopsis thaliana (dicotyledonous plant)	117,000	5
Drosophila melanogaster (fruit fly)	137,000	4
Oryza sativa (rice)	390,000	12
Danio rerio (zebra fish)	1,700,000	25
Gallus gallus (chicken)	1,200,000	40
Mus musculus (mouse)	2,500,000	20
Homo sapiens	3,200,000	23

effort, the "rough draft" of the human genome sequence was reported in early 2001 and the "finished" sequence was reported in mid-2003. This stunning achievement promises to revolutionize the way both biochemistry and medicine are viewed and practiced, although it is likely to require many years of further effort before its full significance is understood. Nevertheless, numerous important conclusions can already be drawn, including these:

1. About half the human genome consists of repeating sequences of various types.
2. Up to 60% of the genome is transcribed to RNA.
3. Only 1.1% to 1.4% of the genome (~2% of the transcribed RNA) encodes protein.
4. The human genome appears to contain only ~23,000 protein-encoding genes [also known as **open reading frames (ORFs)**] rather than the 50,000 to 140,000 ORFs that had previously been predicted. This compares with the ~6000 ORFs in yeast, ~13,000 in *Drosophila,* ~18,000 in *C. elegans,* and ~26,000 in *Arabidopsis* (although these numbers will almost certainly change as our ability to recognize ORFs improves).
5. Only a small fraction of human proteins are unique to vertebrates; most occur in other if not all life-forms.
6. Two randomly selected human genomes differ, on average, by only 1 nucleotide per 1250; that is, any two people are likely to be >99.9% genetically identical.

The obviously greater complexity of humans (vertebrates) relative to invertebrate forms of life is unlikely to be due to the not-much-larger numbers of ORFs that vertebrates encode. Rather, it appears that vertebrate proteins themselves are more complex than those of invertebrates; that is, vertebrate proteins tend to have more domains (modules) than invertebrate proteins, and these modules are more often selectively expressed through **alternative gene splicing** (a phenomenon in which a given gene transcript can be processed in multiple ways so as to yield different proteins when translated; Section 26-3A). Thus, many vertebrate genes encode several different although similar proteins.

E | Evolution Results from Sequence Mutations

One of the richest rewards of nucleic acid sequencing technology is the information it provides about the mechanisms of evolution. The chemical and physical properties of DNA, such as its regular three-dimensional shape and the elegant process of replication, may leave the impression that genetic information is relatively static. In fact, *DNA is a dynamic molecule, subject to changes that alter genetic information.* For example, the mispairing of bases during DNA replication can introduce errors known as **point mutations** in the daughter strand. Mutations also result from DNA damage by chemicals or radiation. More extensive alterations in genetic information are caused by faulty **recombination** (exchange of DNA between chromosomes) and the **transposition** of genes within or between chromosomes and, in some cases, from one organism to another. All these alterations to DNA provide the raw material for natural selection. When a mutated gene is transcribed and the messenger RNA is subsequently translated, the resulting protein may have properties that confer some advantage to the individual. As a beneficial change is passed from generation to

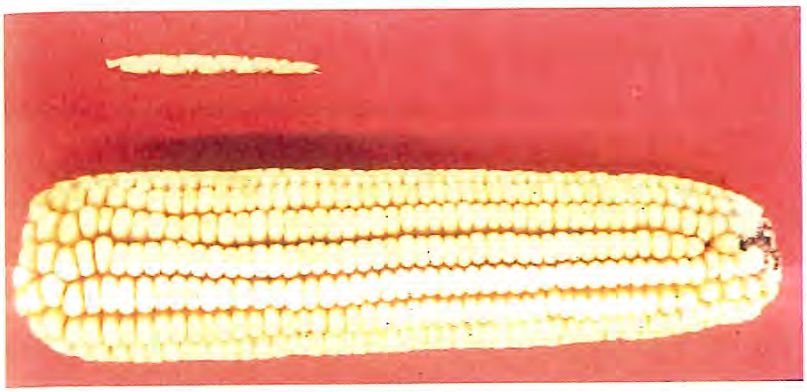


Figure 3-23 | Maize and teosinte. Despite the large differences in phenotype—maize (*bottom*) has hundreds of easily chewed kernels whereas teosinte (*top*) has only a few hard, inedible kernels—the plants differ in only a few genes. The ancestor of maize is believed to be a mutant form of teosinte in which the kernels were more exposed. [John Doebley/Visuals Unlimited.]

generation, it may become part of the standard genetic makeup of the species. Of course, many changes occur as a species evolves, not all of them simple and not all of them gradual.

Phylogenetic relationships can be revealed by comparing the sequences of similar genes in different organisms. The number of nucleotide differences between the corresponding genes in two species roughly indicates the degree to which the species have diverged through evolution. The regrouping of prokaryotes into archaea and bacteria (Section 1-2C) according to rRNA sequences present in all organisms illustrates the impact of sequence analysis.

Nucleic acid sequencing also reveals that species differing in **phenotype** (physical characteristics) are nonetheless remarkably similar at the molecular level. For example, humans and chimpanzees share 98–99% of their DNA. Studies of corn (maize) and its putative ancestor, teosinte, suggest that the plants differ in only a handful of genes governing kernel development (teosinte kernels are encased by an inedible shell; Fig. 3-23).

Small mutations in DNA are apparently responsible for relatively large evolutionary leaps. This is perhaps not so surprising when the nature of genetic information is considered. A mutation in a gene segment that does not encode protein might interfere with the binding of cellular factors that influence the timing of transcription. A mutation in a gene encoding an RNA might interfere with the binding of factors that affect the efficiency of translation. Even a minor rearrangement of genes could disrupt an entire developmental process, resulting in the appearance of a novel species. Notwithstanding the high probability that most sudden changes would lead to diminished individual fitness or the inability to reproduce, the capacity for sudden changes in genetic information is consistent with the fossil record. Ironically, the discontinuities in the fossil record that are probably caused in part by sudden genetic changes once fueled the adversaries of Charles Darwin's theory of evolution by natural selection.

CHECK YOUR UNDERSTANDING

Explain how the restriction–modification system operates.
Summarize the steps in the chain-terminator procedure for sequencing DNA.
What proportion of the human genome is transcribed? Translated?
Explain how evolution can result from a mutation in DNA.

5 Manipulating DNA

LEARNING OBJECTIVES

- Understand how recombinant DNA molecules are constructed and propagated.
- Understand that a DNA library is a collection of cloned DNA segments that can be screened to find a particular gene.
- Understand that the polymerase chain reaction copies and thereby amplifies a defined segment of DNA.
- Understand that recombinant DNA technology can be used to manipulate genes for protein expression or for the production of transgenic organisms.

Along with nucleic acid sequencing, techniques for manipulating DNA *in vitro* and *in vivo* (in the test tube and in living systems) have produced dramatic advances in biochemistry, cell biology, and genetics. In many cases, this **recombinant DNA technology** has made it possible to purify specific DNA sequences and to prepare them in quantities sufficient for study. Consider the problem of isolating a unique 1000-bp length of chromosomal DNA from *E. coli*. A 10-L culture of cells grown at a density of

$\sim 10^{10}$ cells · mL^{-1} contains only ~0.1 mg of the desired DNA, which would be all but impossible to separate from the rest of the DNA using classical separation techniques (Sections 5-2 and 24-3). *Recombinant DNA technology, also called* ***molecular cloning*** *or* ***genetic engineering,*** *makes it possible to isolate, amplify, and modify specific DNA sequences.*

A | Cloned DNA Is an Amplified Copy

The following approach is used to obtain and amplify a segment of DNA:

1. A fragment of DNA of the appropriate size is generated by a restriction enzyme, by PCR (Section 3-5C), or by chemical synthesis.
2. The fragment is incorporated into another DNA molecule known as a **vector,** which contains the sequences necessary to direct DNA replication.
3. The vector—with the DNA of interest—is introduced into cells, where it is replicated.
4. Cells containing the desired DNA are identified, or selected.

Cloning refers to the production of multiple identical organisms derived from a single ancestor. The term **clone** refers to the collection of cells that contain the vector carrying the DNA of interest or to the DNA itself. In a suitable host organism, such as *E. coli* or yeast, large amounts of the inserted DNA can be produced.

Cloned DNA can be purified and sequenced (Section 3-4). Alternatively, if a cloned gene is flanked by the properly positioned regulatory sequences for RNA and protein synthesis, the host may also produce large quantities of the RNA and protein specified by that gene. Thus, cloning provides materials (nucleic acids and proteins) for other studies and also provides a means for studying gene expression under controlled conditions.

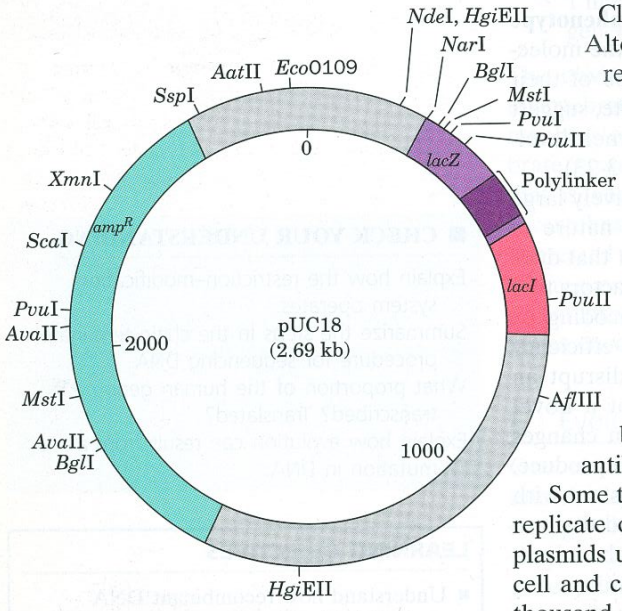


Figure 3-24 | The plasmid pUC18. As shown in this diagram, the circular plasmid contains multiple restriction sites, including a **polylinker** sequence that contains 13 restriction sites that are not present elsewhere on the plasmid. The three genes expressed by the plasmid are amp^R, which confers resistance to the antibiotic **ampicillin;** *lacZ,* which encodes the enzyme **β-galactosidase;** and *lacI,* which encodes a factor that controls transcription of *lacZ* (as described in Section 28-2A).

Cloning Vectors Carry Foreign DNA. A variety of small, autonomously replicating DNA molecules are used as cloning vectors. **Plasmids** are circular DNA molecules of 1 to 200 kb found in bacteria or yeast cells. Plasmids can be considered molecular parasites, but in many instances they benefit their host by providing functions, such as resistance to antibiotics, that the host lacks.

Some types of plasmids are present in one or a few copies per cell and replicate only when the bacterial chromosome replicates. However, the plasmids used for cloning are typically present in hundreds of copies per cell and can be induced to replicate until the cell contains two or three thousand copies (representing about half of the cell's total DNA). The plasmids that have been constructed for laboratory use are relatively small, replicate easily, carry genes specifying resistance to one or more antibiotics, and contain a number of conveniently located restriction endonuclease sites into which foreign DNA can be inserted. Plasmid vectors can be used to clone DNA segments of no more than ~10 kb. The *E. coli* plasmid designated **pUC18** (Fig. 3-24) is a representative cloning vector ("pUC" stands for plasmid-Universal Cloning).

Bacteriophage λ (Fig. 3-25) is an alternative cloning vector that can accommodate DNA inserts up to 16 kb. The central third of the 48.5-kb phage genome is not required for infection and can therefore be replaced

by foreign DNAs of similar size. The resulting **recombinant,** or **chimera** (named after the mythological monster with a lion's head, goat's body, and serpent's tail), is packaged into phage particles that can then be introduced into the host cells. One advantage of using phage vectors is that the recombinant DNA is produced in large amounts in easily purified form. **Baculoviruses,** which infect insect cells, are similarly used for cloning in cultures of insect cells.

Much larger DNA segments—up to several hundred kilobase pairs—can be cloned in large vectors known as **bacterial artificial chromosomes (BACs)** or **yeast artificial chromosomes (YACs).** YACs are linear DNA molecules that contain all the chromosomal structures required for normal replication and segregation during yeast cell division. BACs, which replicate in *E. coli,* are derived from circular plasmids that normally replicate long regions of DNA and are maintained at the level of approximately one copy per cell (properties similar to those of actual chromosomes).

Ligase Joins Two DNA Segments. A DNA segment to be cloned is often obtained through the action of restriction endonucleases. Most restriction enzymes cleave DNA to yield sticky ends (Section 3-4A). Therefore, as Janet Mertz and Ron Davis first demonstrated in 1972, *a restriction fragment can be inserted into a cut made in a cloning vector by the same restriction enzyme* (Fig. 3-26). The complementary ends of the two DNAs form base pairs **(anneal)** and the sugar–phosphate backbones are covalently **ligated,** or spliced together, through the action of an enzyme named **DNA ligase.** (A ligase produced by a bacteriophage can also join blunt-ended restriction fragments.) A great advantage of using a restriction enzyme to construct a recombinant DNA molecule is that the DNA insert can later be precisely excised from the cloned vector by cleaving it with the same restriction enzyme.

Selection Detects the Presence of a Cloned DNA. The expression of a chimeric plasmid in a bacterial host was first demonstrated in 1973 by Herbert Boyer and Stanley Cohen. A host bacterium can take up a plasmid when the two are mixed together, but the vector becomes permanently established in its bacterial host (transformation) with an efficiency of only ~0.1%. However, a single transformed cell can multiply without limit, producing large quantities of recombinant DNA. Bacterial cells are typically plated on a semisolid growth medium at a low enough density that discrete colonies, each arising from a single cell, are visible.

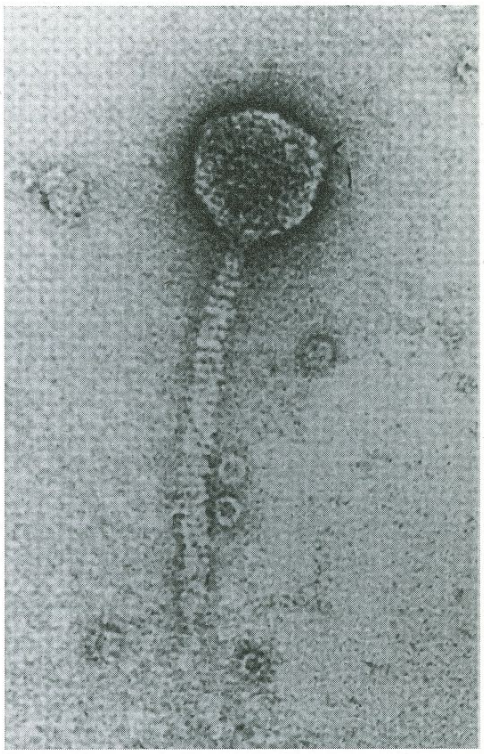


Figure 3-25 | Bacteriophage λ. During phage infection, DNA contained in the "head" of the phage particle enters the bacterial cell, where it is replicated ~100 times and packaged to form progeny phage. [Electron micrograph courtesy of A.F. Howatson. From Lewin, B., *Gene Expression,* Vol. 3, Fig. 5.23, Wiley (1977).]

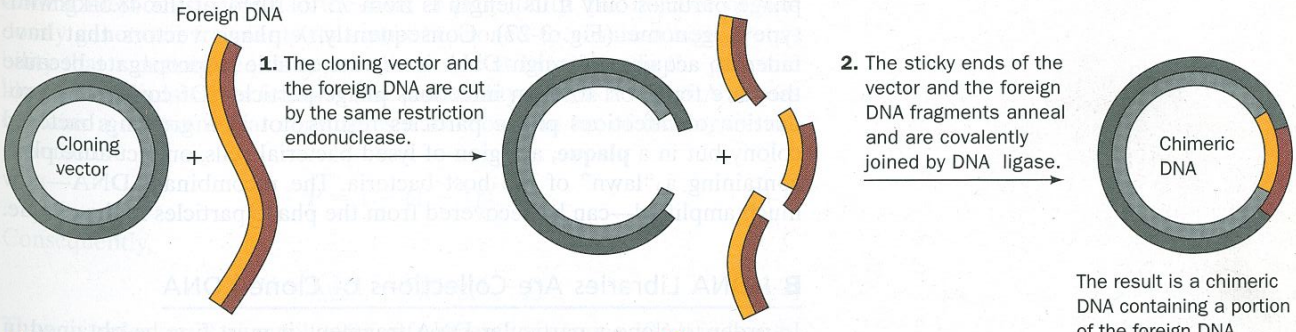
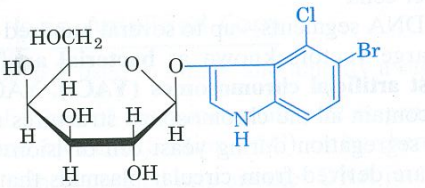
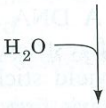


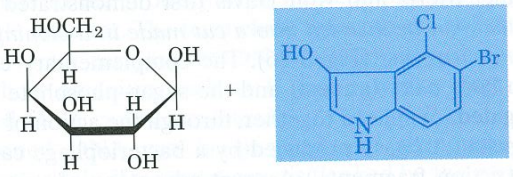
Figure 3-26 | Construction of a recombinant DNA molecule. See the Animated Figures.

It is essential to select only those host organisms that have been transformed and that contain a properly constructed vector. In the case of plasmid transformation, selection can be accomplished through the use of antibiotics and/or chromogenic (color-producing) substances. For example, the *lacZ* gene in the pUC18 plasmid (see Fig. 3-24) encodes the enzyme β-galactosidase, which cleaves the colorless compound **X-gal** to a blue product:



5-Bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal)
(*colorless*)





β-D-Galactose **5-Bromo-4-chloro-3-hydroxyindole**
(*blue*)

Cells of *E. coli* that have been transformed by an unmodified pUC18 plasmid form blue colonies. However, if the plasmid contains a foreign DNA insert in its polylinker region, the colonies are colorless because the insert interrupts the protein-coding sequence of the *lacZ* gene and no functional β-galactosidase is produced. Bacteria that have failed to take up any plasmid are also colorless due to the absence of β-galactosidase, but these cells can be excluded by adding the antibiotic ampicillin to the growth medium (the plasmid includes the gene amp^R, which confers ampicillin resistance). Thus, successfully transformed cells form colorless colonies in the presence of ampicillin. Genes such as amp^R are known as **selectable markers.**

Genetically engineered bacteriophage λ vectors contain restriction sites that flank the dispensable central third of the phage genome. This segment can be replaced by foreign DNA, but the chimeric DNA is packaged in phage particles only if its length is from 75 to 105% of the 48.5-kb wild-type λ genome (Fig. 3-27). Consequently, λ phage vectors that have failed to acquire a foreign DNA insert are unable to propagate because they are too short to form infectious phage particles. Of course, the production of infectious phage particles results not in a growing bacterial colony but in a **plaque,** a region of lysed bacterial cells, on a culture plate containing a "lawn" of the host bacteria. The recombinant DNA—now much amplified—can be recovered from the phage particles in the plaque.

B | DNA Libraries Are Collections of Cloned DNA

In order to clone a particular DNA fragment, it must first be obtained in relatively pure form. The magnitude of this task can be appreciated by considering that, for example, a 1-kb fragment of human DNA represents

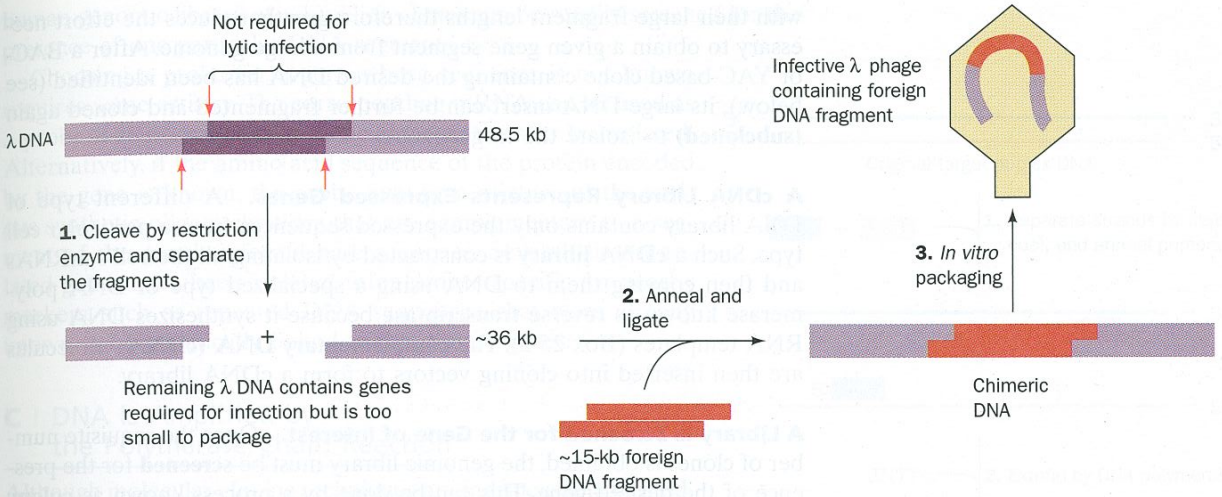


Figure 3-27 | Cloning with bacteriophage λ. Removal of a nonessential portion of the phage genome allows a segment of foreign DNA to be inserted. The DNA insert can be packaged into an infectious phage particle only if the insert DNA has the appropriate size. **See the Animated Figures.**

only 0.000031% of the 3.2 billion-bp human genome. Of course, identifying a particular DNA fragment requires knowing something about its nucleotide sequence or its protein product. In practice, it is usually more difficult to identify a particular DNA fragment from an organism and then clone it than it is to clone all the organism's DNA that might contain the DNA of interest and then identify the clones containing the desired sequence.

A Genomic Library Includes All of an Organism's DNA. The cloned set of all DNA fragments from a particular organism is known as its **genomic library.** Genomic libraries are generated by a procedure known as **shotgun cloning.** The chromosomal DNA of the organism is isolated, cleaved to fragments of clonable size, and inserted into a cloning vector. The DNA is usually fragmented by partial rather than exhaustive restriction digestion so that the genomic library contains intact representatives of all the organism's genes, including those that contain restriction sites. DNA in solution can also be mechanically fragmented **(sheared)** by rapid stirring.

Given the large size of the genome relative to a gene, the shotgun cloning method is subject to the laws of probability. The number of randomly generated fragments that must be cloned to ensure a high probability that a desired sequence is represented at least once in the genomic library is calculated as follows: The probability P that a set of N clones contains a fragment that constitutes a fraction f, in bp, of the organism's genome is

$$P = 1 - (1 - f)^N \qquad [3\text{-}1]$$

Consequently,

$$N = \log(1 - P)/\log(1 - f) \qquad [3\text{-}2]$$

Thus, in order for P to equal 0.99 for fragments averaging 10 kb in length, $N = 2162$ for the 4600-kb *E. coli* chromosome and 63,000 for the 137,000-kb *Drosophila* genome. The use of BAC- or YAC-based genomic libraries

with their large fragment lengths therefore greatly reduces the effort necessary to obtain a given gene segment from a large genome. After a BAC- or YAC-based clone containing the desired DNA has been identified (see below), its large DNA insert can be further fragmented and cloned again **(subcloned)** to isolate the target DNA.

A cDNA Library Represents Expressed Genes. A different type of DNA library contains only the expressed sequences from a particular cell type. Such a **cDNA library** is constructed by isolating all the cell's mRNAs and then copying them to DNA using a specialized type of DNA polymerase known as **reverse transcriptase** because it synthesizes DNA using RNA templates (Box 25-2). The **complementary DNA (cDNA)** molecules are then inserted into cloning vectors to form a cDNA library.

A Library Is Screened for the Gene of Interest. Once the requisite number of clones is obtained, the genomic library must be **screened** for the presence of the desired gene. This can be done by a process known as **colony** or ***in situ* hybridization** (Latin: *in situ,* in position; Fig. 3-28). The cloned yeast colonies, bacterial colonies, or phage plaques to be tested are transferred, by **replica plating,** from a master plate to a nitrocellulose filter (replica plating is also used to transfer colonies to plates containing different growth media). Next, the filter is treated with NaOH, which lyses the cells or phages and separates the DNA into single strands, which preferentially bind to the nitrocellulose. The filter is then dried to fix the DNA in place and incubated with a labeled **probe.** The probe is a short segment of DNA or RNA whose sequence is complementary to a portion of the DNA of interest. After washing away unbound probe, the presence of the probe on the nitrocellulose is detected by a technique appropriate for the label used (e.g., exposure to X-ray film for a radioactive probe, a process known as **autoradiography,** or illumination with an appropriate wavelength for a fluorescent probe). Only those colonies or plaques containing the desired gene bind the probe and are thereby detected. The corresponding clones can then be retrieved from the master plate. Using this technique, a

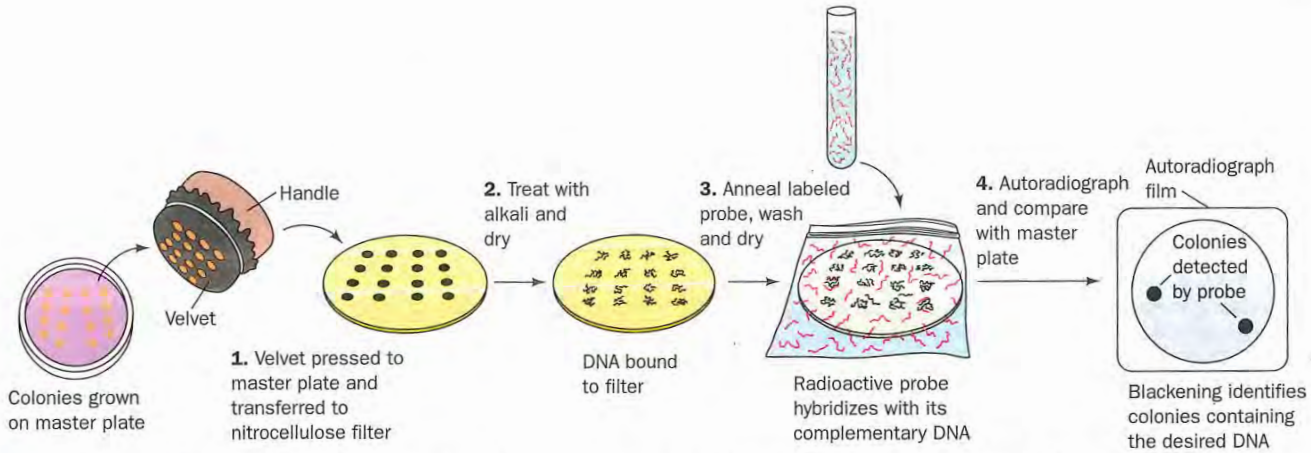


Figure 3-28 | Colony (*in situ*) hybridization. Colonies are transferred from a "master" culture plate by replica plating. Clones containing the DNA of interest are identified by the ability to bind a specific probe. Here, binding is detected by laying X-ray film over the dried filter. Since the colonies on the master plate and on the filter have the same spatial distribution, positive colonies are easily retrieved.

human genomic library of ~1 million clones can be readily screened for the presence of one particular DNA segment.

Choosing a probe for a gene whose sequence is not known requires some artistry. The corresponding mRNA can be used as a probe if it is available in sufficient quantities to be isolated. Alternatively, if the amino acid sequence of the protein encoded by the gene is known, the probe may be a mixture of the various synthetic oligonucleotides that are complementary to a segment of the gene's inferred base sequence. Several disease-related genes have been isolated using probes specific for nearby markers, such as repeated DNA sequences, that were already known to be genetically linked to the disease genes.

C | DNA Is Amplified by the Polymerase Chain Reaction

Although molecular cloning techniques are indispensable to modern biochemical research, the **polymerase chain reaction (PCR)** is often a faster and more convenient method for amplifying a specific DNA. Segments of up to 6 kb can be amplified by this technique, which was devised by Kary Mullis in 1985. *In PCR, a DNA sample is separated into single strands and incubated with DNA polymerase, dNTPs, and two oligonucleotide primers whose sequences flank the DNA segment of interest. The primers direct the DNA polymerase to synthesize complementary strands of the target DNA* (Fig. 3-29). Multiple cycles of this process, each doubling the amount of the target DNA, geometrically amplify the DNA starting from as little as a single gene copy. In each cycle, the two strands of the duplex DNA are separated by heating, the primers are annealed to their complementary segments on the DNA, and the DNA polymerase directs the synthesis of the complementary strands. The use of a heat-stable DNA polymerase, such as ***Taq* polymerase** isolated from *Thermus aquaticus*, a bacterium that thrives at 75°C, eliminates the need to add fresh enzyme after each round of heating (heat inactivates most enzymes). Hence, in the presence of sufficient quantities of primers and dNTPs, PCR is carried out simply by cyclically varying the temperature.

Twenty cycles of PCR increase the amount of the target sequence around a millionfold ($\sim 2^{20}$) with high specificity. Indeed, PCR can amplify a target DNA present only once in a sample of 10^5 cells, so this method can be used without prior DNA purification. The amplified DNA can then be sequenced or cloned.

PCR amplification has become an indispensable tool. Clinically, it is used to diagnose infectious diseases and to detect rare pathological events such as mutations leading to cancer. Forensically, the DNA from a single hair or sperm can be am-

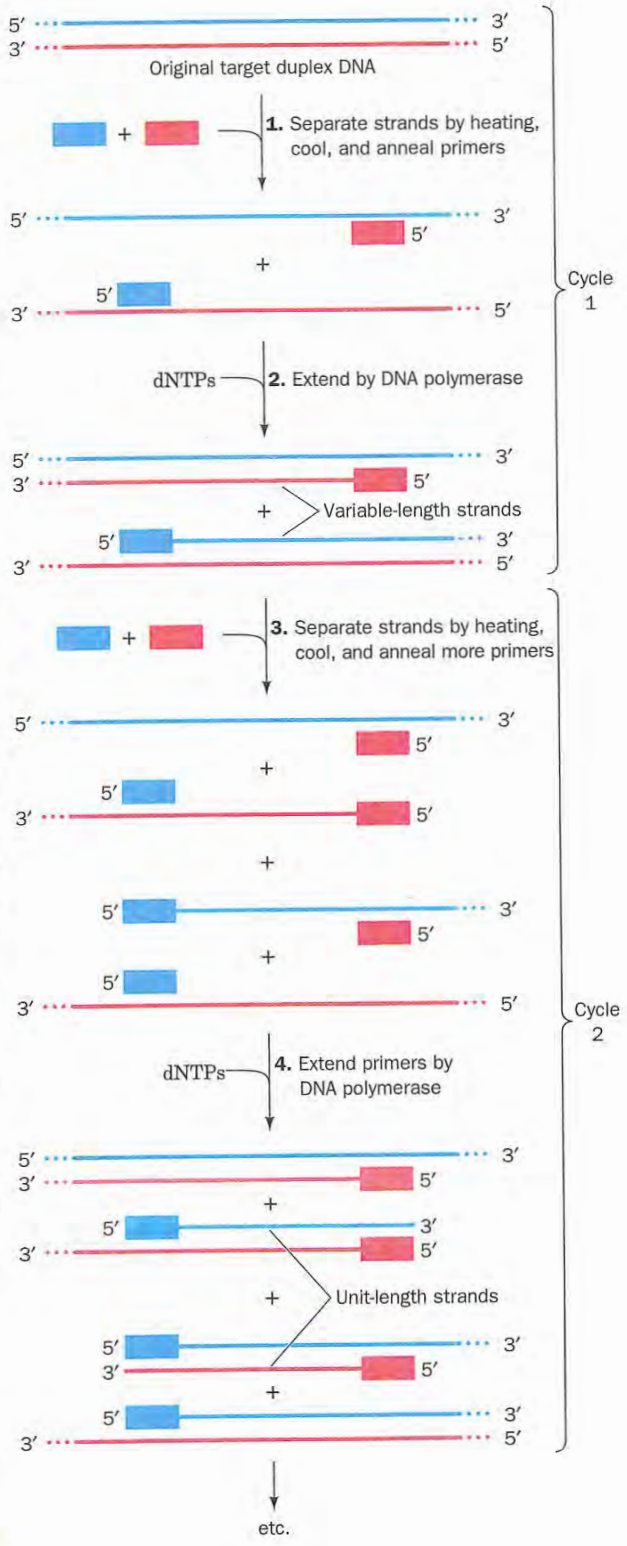
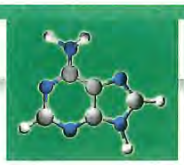


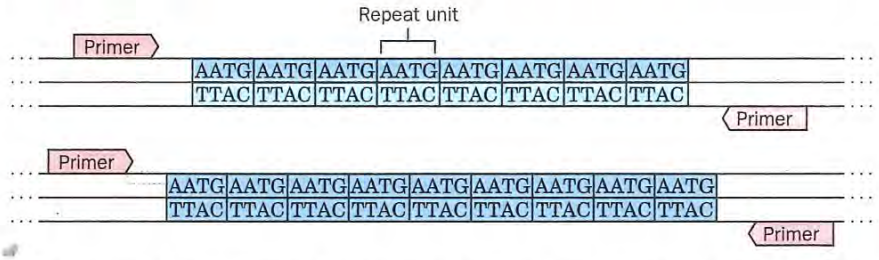
Figure 3-29 | The polymerase chain reaction (PCR). In each cycle of the reaction, the strands of the duplex DNA are separated by heating, the reaction mixture is cooled to allow primers to anneal to complementary sequences on each strand, and DNA polymerase extends the primers. The number of "unit-length" strands doubles with every cycle after the second cycle. By choosing primers specific for each end of a gene, the gene can be amplified over a millionfold.



BOX 3-2 PERSPECTIVES IN BIOCHEMISTRY

DNA Fingerprinting

Forensic DNA testing takes advantage of DNA sequence variations or **polymorphisms** that occur among individuals. Many genetic polymorphisms have no functional consequences because they occur in regions of the DNA that contain many repetitions but do not encode genes (although if they are located near a "disease" gene, they can be used to track and identify the gene). Modern **DNA fingerprinting** methods examine these noncoding repetitive DNA sequences in samples that have been amplified by PCR.

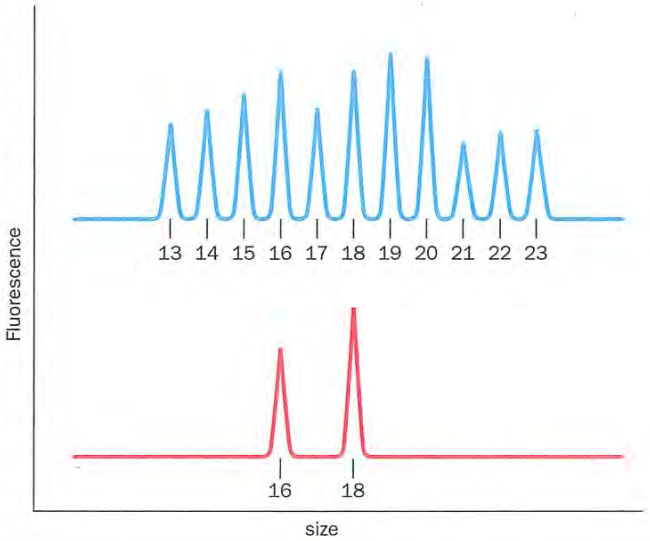


Tandemly repeated DNA sequences occur throughout the human genome and include **short tandem repeats (STRs),** which contain variable numbers of repeating segments of two to seven base pairs. The most popular STR sites for forensic use contain tetranucleotide repeats. The number of repeats at any one site on the DNA varies between individuals, even within a family. Each different number of repeats at a site is called an **allele,** and each individual can have two alleles, one from each parent. Since PCR is the first step of the fingerprinting process, only a tiny amount (~1 ng) of DNA is needed. The region of DNA containing the STR is amplified by PCR using primers that are complementary to the unique (nonrepeating) sequences flanking the repeats.

The amplified products are separated by electrophoresis and detected by the fluorescent tag on their primers. An STR allele is small enough (<500 bp) that DNA fragments differing by a four-base repeat can be readily differentiated. The allele designation for each STR site is generally the number of times a repeated unit is present. STR sites that have been selected for forensic use generally have 7 to 30 different alleles.

In the example shown here, the upper trace shows the fluorescence of the electrophoretogram of reference standards (the set of all possible alleles, each identified by the number of repeat units, from 13 to 23). The lower trace corresponds to the sample being tested, which contains two alleles, one with 16 repeats and one with 18 repeats. Several STR sites can be analyzed simultaneously by using the appropriate primers and tagging them with different fluorescence dyes.

The probability of two individuals having matching DNA fingerprints depends on the number of STR sites examined and the number of alleles at each site. For example, if a pair of alleles at one site occurs in the population with a frequency of 10% (1/10), and a pair of alleles at a second site occurs with a frequency of 5% (1/20), then the probability that the DNA fingerprints from two individuals would match at both sites is 1 in 200 (1/10 × 1/20; the probabilities of independent events are multiplied). By examining multiple STR sites, the probability of obtaining matching fingerprints by chance becomes exceedingly small.



See Guided Exploration 3
PCR and site-directed mutagenesis.

plified by PCR so that it can be used to identify the donor (Box 3-2). Traditional ABO blood-type analysis requires a coin-sized drop of blood; PCR is effective on pinhead-sized samples of biological fluids. Courts now

consider DNA sequences as unambiguous identifiers of individuals, as are fingerprints, because the chance of two individuals sharing extended sequences of DNA is typically one in a million or more. In a few cases, PCR has dramatically restored justice to convicts who were released from prison on the basis of PCR results that proved their innocence—even many years after the crime-scene evidence had been collected.

D | Recombinant DNA Technology Has Numerous Practical Applications

The ability to manipulate DNA sequences allows genes to be altered and expressed in order to obtain proteins with improved functional properties or to correct genetic defects.

Cloned Genes Can Be Expressed. The production of large quantities of scarce or novel proteins is relatively straightforward only for bacterial proteins: A cloned gene must be inserted into an **expression vector,** a plasmid that contains properly positioned transcriptional and translational control sequences. The production of a protein of interest may reach 30% of the host's total cellular protein. Such genetically engineered organisms are called **overproducers.** Bacterial cells often sequester large amounts of useless and possibly toxic (to the bacterium) protein as insoluble inclusions, which sometimes simplifies the task of purifying the protein.

Bacteria can produce eukaryotic proteins only if the recombinant DNA that carries the protein-coding sequence also includes bacterial transcriptional and translational control sequences. Synthesis of eukaryotic proteins in bacteria also presents other problems. For example, many eukaryotic genes are large and contain stretches of nucleotides **(introns)** that are transcribed and excised before translation (Section 26-3A); bacteria lack the machinery to excise the introns. In addition, many eukaryotic proteins are posttranslationally modified by the addition of carbohydrates or by other reactions. These problems can be overcome by using expression vectors that propagate in eukaryotic hosts, such as yeast or cultured insect or animal cells.

Table 3-4 lists some recombinant proteins produced for medical and agricultural use. In many cases, purification of these proteins directly from human or animal tissues is unfeasible on ethical or practical grounds.

Table 3-4 Some Proteins Produced by Genetic Engineering

Protein	Use
Human insulin	Treatment of diabetes
Human growth hormone	Treatment of some endocrine disorders
Erythropoietin	Stimulation of red blood cell production
Colony-stimulating factors	Production and activation of white blood cells
Coagulation factors IX and X	Treatment of blood clotting disorders (hemophilia)
Tissue-type plasminogen activator	Lysis of blood clots after heart attack and stroke
Bovine growth hormone	Production of milk in cows
Hepatitis B surface antigen	Vaccination against hepatitis B

Expression systems permit large-scale, efficient preparation of the proteins while minimizing the risk of contamination by viruses or other pathogens from tissue samples.

Site-Directed Mutagenesis Alters a Gene's Nucleotide Sequence. After isolating a gene, it is possible to modify the nucleotide sequence to alter the amino acid sequence of the encoded protein. **Site-directed mutagenesis,** a technique pioneered by Michael Smith, *mimics the natural process of evolution and allows predictions about the structural and functional roles of particular amino acids in a protein to be rigorously tested in the laboratory.*

Synthetic oligonucleotides are required to specifically alter genes through site-directed mutagenesis. An oligonucleotide whose sequence is identical to a portion of the gene of interest except for the desired base changes is used to direct replication of the gene. The oligonucleotide hybridizes to the corresponding **wild-type** (naturally occurring) sequence if there are no more than a few mismatched base pairs. Extension of the oligonucleotide, called a primer, by DNA polymerase yields the desired altered gene (Fig. 3-30). The altered gene can then be inserted into an appropriate vector. A mutagenized primer can also be used to generate altered genes by PCR.

Transgenic Organisms Contain Foreign Genes. For many purposes it is preferable to tailor an intact organism rather than just a protein—true genetic engineering. Multicellular organisms expressing a gene from another organism are said to be **transgenic,** and the transplanted foreign gene is called a **transgene.**

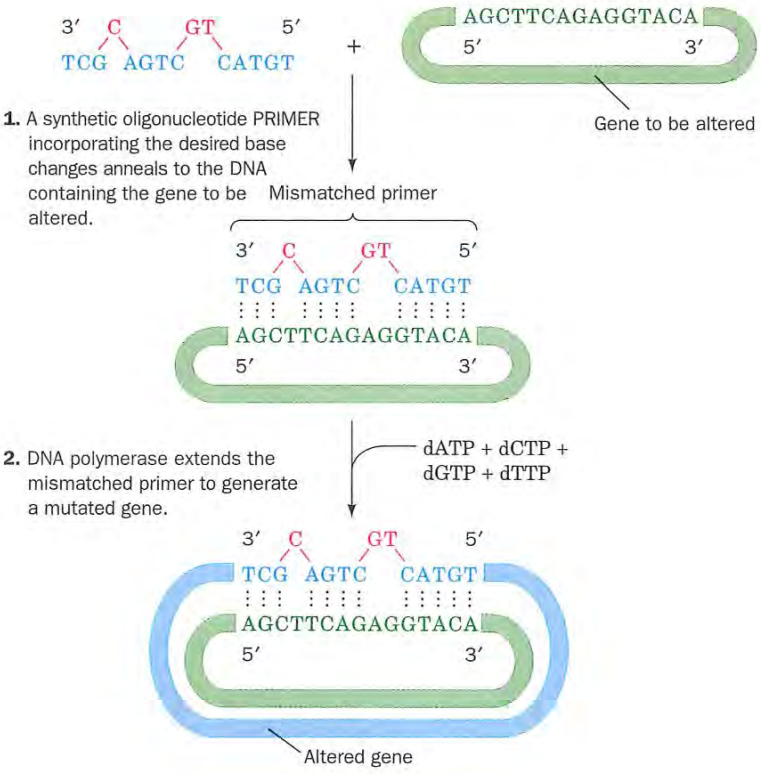


Figure 3-30 | Site-directed mutagenesis. The altered gene can be inserted into a suitable cloning vector to be amplified, expressed, or used to generate a mutant organism. **See the Animated Figures.**

For the change to be permanent, that is, heritable, a transgene must be stably integrated into the organism's germ cells. For mice, this is accomplished by microinjecting cloned DNA encoding the desired altered characteristics into a fertilized egg and implanting it into the uterus of a foster mother. A well-known example of a transgenic mouse contains extra copies of a growth hormone gene (Fig. 3-31).

Transgenic farm animals have also been developed. Ideally, the genes of such animals could be tailored to allow the animals to grow faster on less food or to be resistant to particular diseases. Some transgenic farm animals have been engineered to secrete medically useful proteins into their milk. Harvesting such a substance from milk is much more cost-effective than producing the same substance in bacterial cultures.

One of the most successful transgenic organisms is corn (maize) that has been genetically modified to produce a protein that is toxic to plant-eating insects (but harmless to vertebrates). The toxin is synthesized by the soil microbe *Bacillus thuringiensis*. The toxin gene has been cloned into corn in order to confer protection against the European corn borer, a commercially significant pest that spends much of its life cycle inside the corn plant, where it is largely inaccessible to chemical insecticides. The use of "Bt corn," which is now widely planted in the United States, has greatly reduced the need for such toxic substances.

Transgenic plants have also been engineered for better nutrition. For example, researchers have developed a strain of rice with foreign genes that encode enzymes necessary to synthesize **β-carotene** (an orange pigment that is the precursor of **vitamin A**) and a gene for the iron-storage protein **ferritin.** The genetically modified rice, which is named "golden rice" (Fig. 3-32), should help alleviate vitamin A deficiencies (which afflict some 400 million people) and iron deficiencies (an estimated 30% of the world's population suffers from iron deficiency). Other transgenic plants include freeze-tolerant strawberries and slow-ripening tomatoes.

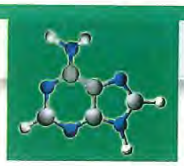
There is presently a widely held popular suspicion, particularly in Europe, that genetically modified or "GM" foods will somehow be harmful. However, extensive research, as well as considerable consumer experience, has failed to reveal any deleterious effects caused by GM foods (see Box 3-3).

Transgenic organisms have greatly enhanced our understanding of gene expression. Animals that have been engineered to contain a defective gene or that lack a gene entirely (a so-called **gene knockout**) also serve as experimental models for human diseases.

Figure 3-31 | Transgenic mouse. The gigantic mouse on the left was grown from a fertilized ovum that had been microinjected with DNA containing the rat growth hormone gene. He is nearly twice the weight of his normal littermate on the right. [Courtesy of Ralph Brinster, University of Pennsylvania.]

Genetic Defects Can Be Corrected. **Gene therapy** is the transfer of new genetic material to the cells of an individual in order to produce a therapeutic effect. Although the potential benefits of this as yet rudimentary technology are enormous, there are numerous practical obstacles to overcome. For example, the retroviral vectors (RNA-containing viruses) commonly used to directly introduce genes into humans can provoke a fatal immune response.

Figure 3-32 | Golden rice. The white grains on the right are the wild type. The grains on the left have been engineered to store up to three times more iron and to synthesize β-carotene, which gives them their yellow color. [Courtesy of Ingo Potrykus.]



BOX 3-3 PERSPECTIVES IN BIOCHEMISTRY

Ethical Aspects of Recombinant DNA Technology

In the early 1970s, when genetic engineering was first discussed, little was known about the safety of the proposed experiments. After considerable debate, during which there was a moratorium on such experiments, regulations for recombinant DNA research were drawn up. The rules prohibit obviously dangerous experiments (e.g., introducing the gene for diphtheria toxin into *E. coli,* which would convert this human symbiont into a deadly pathogen). Other precautions limit the risk of accidentally releasing potentially harmful organisms into the environment. For example, many vectors must be cloned in host organisms with special nutrient requirements. These organisms are unlikely to survive outside the laboratory.

The proven value of recombinant DNA technology has silenced nearly all its early opponents. Certainly, it would not have been possible to study some pathogens, such as the virus that causes AIDS, without cloning. The lack of recombinant-induced genetic catastrophes so far does not guarantee that recombinant organisms won't ever adversely affect the environment. Nevertheless, the techniques used by genetic engineers mimic those used in nature—that is, mutation and selection—so natural and man-made organisms are fundamentally similar. In any case, people have been breeding plants and animals for several millennia already, and for many of the same purposes that guide experiments with recombinant DNA.

There are other ethical considerations to be faced as new genetic engineering techniques become available. Bacterially produced human growth hormone is now routinely prescribed to increase the stature of abnormally short children. However, should athletes be permitted to use this protein, as some reportedly have, to increase their size and strength? Few would dispute the use of gene therapy, if it can be developed, to cure such genetic defects as sickle-cell anemia (Section 7-1E) and Lesch–Nyhan syndrome (Section 23-1D). If, however, it becomes possible to alter complex (i.e., multigene) traits such as athletic ability and intelligence, which changes would be considered desirable and who would decide whether to make them? Should gene therapy be used only to correct an individual's defects, or should it also be used to alter genes in the individual's germ cells so that succeeding generations would not inherit the defect? If it becomes easy to determine an individual's genetic makeup, should this information be used in evaluating applicants for educational and employment opportunities or for health insurance? These conundrums have led to the creation of a branch of philosophy, named **bioethics,** designed to deal with them.

CHECK YOUR UNDERSTANDING

What are the roles of the vector and DNA ligase in cloning DNA?

Explain how cells containing recombinant DNA are selected.

What is a DNA library and how can it be screened for a particular gene?

Describe the steps required to amplify a DNA segment by PCR.

Explain how site-directed mutagenesis can be used to produce an altered protein in bacterial cells.

What is the difference between manipulating a gene for gene therapy and for producing a transgenic organism?

The only documented success of gene therapy in humans has occurred in children with a form of **severe combined immunodeficiency disease (SCID)** known as **SCID-X1,** which without treatment would have required their isolation in a sterile environment to prevent fatal infection. SCID-X1 is caused by a defect in the gene encoding **γc cytokine receptor,** whose action is essential for proper immune system function. Bone marrow cells (the precursors of white blood cells) were removed from the bodies of SCID-X1 victims, incubated with a vector containing a normal γc cytokine receptor gene, and returned to their bodies. The transgenic bone marrow cells restored immune system function. However, because the viral vector integrates into the genome at random, the location of the transgene may affect the expression of other genes, triggering cancer. At least two children have developed leukemia (a white blood cell cancer) as a result of gene therapy for SCID-X1.

SUMMARY

1. Nucleotides consist of a purine or pyrimidine base linked to ribose to which at least one phosphate group is attached. RNA is made of ribonucleotides; DNA is made of deoxynucleotides (which contain 2′-deoxyribose).
2. In DNA, two antiparallel chains of nucleotides linked by phosphodiester bonds form a double helix. Bases in opposite strands pair: A with T, and G with C.

3. Single-stranded nucleic acids, such as RNA, can adopt stem–loop structures.
4. DNA carries genetic information in its sequence of nucleotides. When DNA is replicated, each strand acts as a template for the synthesis of a complementary strand.
5. According to the central dogma of molecular biology, one strand of the DNA of a gene is transcribed into mRNA. The RNA is then translated into protein by the ordered addition of amino acids that are bound to tRNA molecules that base-pair with the mRNA at the ribosome.
6. Restriction endonucleases that recognize certain sequences of DNA are used to specifically cleave DNA molecules.
7. Gel electrophoresis is used to separate and measure the sizes of DNA fragments.
8. In the chain-terminator method of DNA sequencing, the sequence of nucleotides in a DNA strand is determined by enzymatically synthesizing complementary polynucleotides that terminate with a dideoxy analog of each of the four nucleotides. Polynucleotide fragments of increasing size are separated by electrophoresis to reconstruct the original sequence.
9. Mutations and other changes to DNA are the basis for the evolution of organisms.
10. In molecular cloning, a fragment of foreign DNA is inserted into a vector for amplification in a host cell. Transformed cells can be identified by selectable markers.
11. Genomic libraries contain all the DNA of an organism. Clones harboring particular DNA sequences are identified by screening procedures.
12. The polymerase chain reaction amplifies selected sequences of DNA.
13. Recombinant DNA methods are used to produce wild-type or selectively mutagenized proteins in cells or entire organisms.

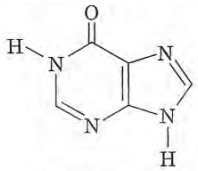
KEY TERMS

nucleotide **39**
nucleic acid **39**
nucleoside **41**
RNA **42**
DNA **42**
polynucleotide **43**
phosphodiester bond **43**
nucleotide residue **43**
5′ end **44**
3′ end **44**
monomer **44**
dimer **44**
trimer **44**
tetramer **44**
oligomer **44**
Chargaff's rules **44**
tautomer **44**
double helix **45**
antiparallel **45**
major groove **46**
minor groove **46**
complementary base pairing **46**
genome **47**
chromosome **47**
diploid **47**
haploid **47**
bp **47**
kb **47**
stem–loop **47**
gene **48**
transformation **48**
replication **48**
transcription **49**
translation **49**
central dogma of molecular biology **49**
mRNA **49**
ribosome **49**
rRNA **49**
tRNA **49**
genomics **50**
transcriptomics **50**
proteomics **50**
gene expression **50**
bacteriophage **51**
restriction–modification system **51**
modification methylase **51**
restriction endonuclease **51**
endonuclease **51**
exonuclease **51**
palindrome **52**
sticky ends **52**
blunt ends **52**
gel electrophoresis **52**
chain-terminator procedure **53**
dNTP **53**
primer **54**
ddNTP **54**
ORF **58**
alternative gene splicing **58**
point mutation **58**
recombination **58**
transposition **58**
phenotype **59**
recombinant DNA technology **59**
vector **60**
cloning **60**
clone **60**
plasmid **60**
recombinant DNA **61**
BAC **61**
YAC **61**
anneal **61**
ligation **61**
selectable marker **62**
plaque **62**
DNA library **62**
genomic library **63**
shotgun cloning **63**
reverse transcriptase **64**
cDNA **64**
screening **64**
colony (*in situ*) hybridization **64**
replica plating **64**
probe **64**
autoradiography **64**
PCR **65**
polymorphism **66**
DNA fingerprinting **66**
STR **66**
allele **66**
expression vector **67**
overproducer **67**
intron **67**
site-directed mutagenesis **68**
wild type **68**
transgenic organism **68**
transgene **68**
gene knockout **69**
gene therapy **69**

PROBLEMS

1. Kinases are enzymes that transfer a phosphoryl group from a nucleoside triphosphate. Which of the following are valid kinase-catalyzed reactions?
   (a) ATP + GDP → ADP + GTP
   (b) ATP + GMP → AMP + GTP
   (c) ADP + CMP → AMP + CDP
   (d) AMP + ATP → 2 ADP
2. A diploid organism with a 45,000-kb haploid genome contains 21% G residues. Calculate the number of A, C, G, and T residues in the DNA of each cell in this organism.

3. A segment of DNA containing 20 base pairs includes 7 guanine residues. How many adenine residues are in the segment? How many uracil residues are in the segment?
4. Draw the tautomeric forms of (a) adenine and (b) cytosine.
5. The adenine derivative hypoxanthine can base-pair with both cytosine and adenine. Show the structures of these base pairs.



Hypoxanthine

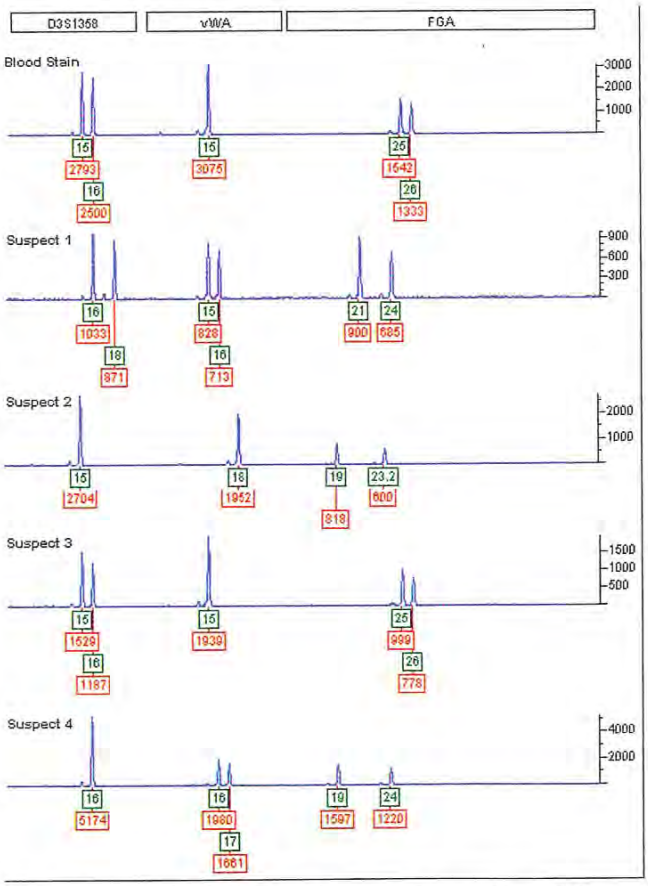
6. Explain why the strands of a DNA molecule can be separated more easily at pH > 11.
7. How many different amino acids could theoretically be encoded by nucleic acids containing four different nucleotides if (a) each nucleotide coded for one amino acid; (b) consecutive sequences of two nucleotides coded for one amino acid; (c) consecutive sequences of three nucleotides coded for one amino acid; (d) consecutive sequences of four nucleotides coded for one amino acid?
8. The recognition sequence for the restriction enzyme *TaqI* is T↓CGA. Indicate the products of the reaction of *TaqI* with the DNA sequence shown.

5′–ACGTCGAATC–3′
3′–TGCAGCTTAG–5′

9. Using the data in Table 3-2, identify restriction enzymes that (a) produce blunt ends; (b) recognize and cleave the same sequence (called **isoschizomers**); (c) produce identical sticky ends.
10. Describe the outcome of a chain-terminator sequencing procedure in which (a) too little ddNTP is added; (b) too much ddNTP is added; (c) too few primers are present; (d) too many primers are present.
11. Calculate the number of clones required to obtain with a probability of 0.99 a specific 5-kb fragment from *C. elegans* (Table 3-3).
12. Describe how to select recombinant clones if a foreign DNA is inserted into the polylinker site of pUC18 and then introduced into *E. coli* cells.
13. Describe the possible outcome of a PCR experiment in which (a) one of the primers is inadvertently omitted from the reaction mixture; (b) one of the primers is complementary to several sites in the starting DNA sample; (c) there is a single-stranded break in the target DNA sequence, which is present in only one copy in the starting sample; (d) there is a double-stranded break in the target DNA sequence, which is present in only one copy in the starting sample.
14. Write the sequences of the two 12-residue primers that could be used to amplify the following DNA segment by PCR.

ATAGGCATAGGCCCATATGGCATAAGG-
CTTTATAATATGCGATAGGCGCTGGTCAG

15. (a) Why is a genomic library larger than a cDNA library for a given organism?
    (b) Why do cDNA libraries derived from different cell types within the same organism differ from each other?
16. A blood stain from a crime scene and blood samples from four suspects were analyzed by PCR using fluorescent primers associated with three STR loci: D3S1358, vWA, and FGA. The resulting electrophoretograms are shown below. The numbers beneath each peak identify the allele (upper box) and the height of the peak in relative fluorescence units (lower box).
    (a) Since everyone has two copies of each chromosome and therefore two alleles of each gene, what accounts for the appearance of only one allele at some loci?
    (b) Which suspect is a possible source of the blood?
    (c) Could the suspect be identified using just one of the three STR loci?
    (d) What can you conclude about the amount of DNA obtained from Suspect 1 compared to Suspect 4?



[From Thompson, W.C., Ford, S., Doom, T., Raymer, M., and Krane, D.E., Evaluating forensic DNA evidence: Essential elements of a competent defense review, *The Champion* **27,** 16–25 (2003).]

BIOINFORMATICS EXERCISES

Bioinformatics Exercises are available at www.wiley.com/college/voet.

Chapter 3
Databases for the Storage and "Mining" of Genome Sequences

1. **Finding Databases.** Locate databases for genome sequences and explore the meaning of terms related to them.
2. **The Institute for Genomic Research.** Explore a prokaryotic genome and find listings for eukaryotic genomes.
3. **Analyzing a DNA Sequence.** Given a DNA sequence, identify its open reading frame and translate it into a protein sequence.
4. **Sequence Homology.** Perform a BLAST search for homologs of a protein sequence.
5. **Plasmids and Cloning.** Predict the sizes of the fragments produced by the action of various restriction enzymes on plasmids.

REFERENCES

DNA Structure and Function

Bloomfield, V.A., Crothers, D.M., and Tinoco, I., Jr., *Nucleic Acids. Structures, Properties, and Functions,* University Science Books (2000).

Dickerson, R.E., DNA structure from A to Z, *Methods Enzymol.* **211,** 67–111 (1992). [Describes the various crystallographic forms of DNA.]

Thieffry, D., Forty years under the central dogma, *Trends Biochem. Sci.* **23,** 312–316 (1998). [Traces the origins, acceptance, and shortcomings of the idea that nucleic acids contain biological information.]

Watson, J.D. and Crick, F.H.C., Molecular structure of nucleic acids, *Nature* **171,** 737–738 (1953); *and* Genetical implications of the structure of deoxyribonucleic acid, *Nature* **171,** 964–967 (1953). [The seminal papers that are widely held to mark the origin of modern molecular biology.]

DNA Sequencing

Galperin, M.Y., The molecular biology database collection: 2007 update, *Nucleic Acids Res.* **35,** Database issue D3–D4 (2007). [This article cites 968 databases covering various aspects of molecular biology, biochemistry, and genetics. Additional articles in the same issue provide more information on individual databases. Freely available at http://nar.oxfordjournals.org.]

Graham, C.A. and Hill, A.J.M. (Eds.), *DNA Sequencing Protocols* (2nd ed.), Humana Press (2001).

Higgins, D. and Taylor, W. (Eds.), *Bioinformatics. Sequence, Structure and Databanks,* Oxford University Press (2000).

International Human Genome Sequencing Consortium, initial sequencing and analysis of the human genome, *Nature* **409,** 860–921 (2001); *and* Venter, J.C., *et al.,* the sequence of the human genome, *Science* **291,** 1304–1351 (2001). [These and other papers in the same issues of *Nature* and *Science* describe the data that constitute the draft sequence of the human genome and discuss how this information can be used in understanding biological function, evolution, and human health.]

International Human Genome Sequencing Consortium, finishing the euchromatic sequence of the human genome, *Nature* **431,** 931–945 (2004). [Describes the most up-to-date version of the human genome sequence.]

Recombinant DNA Technology

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, K., *Short Protocols in Molecular Biology* (5th ed.), Wiley (2002).

Nicholl, D.S.T., *An Introduction to Genetic Engineering* (2nd ed.), Cambridge University Press (2002).

Pingoud, A., Fuxreiter, M., Pingoud, V., and Wende, W., Type II restriction endonucleases: structure and mechanism, *Cell. Mol. Life Sci.* **62,** 685–707 (2005). [Includes an overview of different types of restriction enzymes.]

Sambrook, J., and Russell, D. *Molecular Cloning* (3rd ed.), Cold Spring Harbor Laboratory (2001). [A three-volume "bible" of laboratory protocols with accompanying background explanations.]

4

Amino Acids

In addition to the well-known taste sensations of sweet, sour, salty, and bitter is umami, the taste sensation elicited by monosodium glutamate (MSG), an amino acid commonly used as a flavor enhancer. [Jackson Vereen/Foodpix/PictureArts Corp.]

CHAPTER CONTENTS

LEARNING OBJECTIVES

- Know the overall structure of an amino acid and the structures of the 20 different R groups.
- Understand how peptide bonds link amino acid residues in a polypeptide.
- Understand that amino acids include ionizable groups whose p*K* values vary when the amino acid is part of a polypeptide.

When scientists first turned their attention to nutrition, early in the nineteenth century, they quickly discovered that natural products containing nitrogen were essential for the survival of animals. In 1839, the Dutch chemist G. J. Mulder coined the term **protein** (Greek: *proteios,* primary) for this class of compounds. The physiological chemists of that time did not realize that proteins were actually composed of smaller components, amino acids, although the first amino acids had been isolated in 1830. In fact, for many years, it was believed that substances from plants—including proteins—were incorporated whole into animal tissues. This misconception was laid to rest when the process of digestion came to light. After it became clear that ingested proteins were broken down to smaller compounds containing amino acids, scientists began to consider the nutritive qualities of those compounds (Box 4-1).

Modern studies of proteins and amino acids owe a great deal to nineteenth and early twentieth century experiments. We now understand that nitrogen-containing amino acids are essential for life and that they are the building blocks of proteins. The central role of amino acids in biochemistry is perhaps not surprising: Several amino acids are among the organic compounds believed to have appeared early in the earth's history (Section 1-1A). Amino acids, as ancient and ubiquitous molecules, have been co-opted by evolution for a variety of purposes in living systems. We begin this chapter by discussing the structures and chemical properties of the common amino acids, including their stereochemistry, and end with a brief summary of the structures and functions of some related compounds.

1 Amino Acid Structure

The analyses of a vast number of proteins from almost every conceivable source have shown that *all proteins are composed of 20 "standard" amino acids.* Not every protein contains all 20 types of amino acids, but most proteins contain most if not all of the 20 types.

The common amino acids are known as **α-amino acids** because they have a primary amino group ($—NH_2$) as a substituent of the **α carbon** atom, the carbon next to the carboxylic acid group (—COOH; Fig. 4-1).

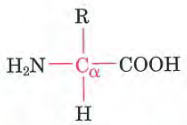
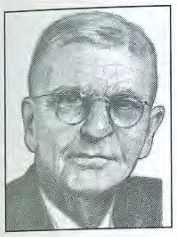


Figure 4-1 | General structure of an α-amino acid. The R groups differentiate the 20 standard amino acids.

BOX 4-1 **PATHWAYS OF DISCOVERY**

William C. Rose and the Discovery of Threonine



William C. Rose (1887–1985)

Identifying the amino acid constituents of proteins was a scientific challenge that grew out of studies of animal nutrition. At the start of the twentieth century, physiological chemists (the term *biochemist* was not yet used) recognized that not all foods provided adequate nutrition. For example, rats fed the corn protein zein as their only source of nitrogen failed to grow unless the amino acids tryptophan and lysine were added to their diet. Knowledge of metabolism at that time was mostly limited to information gleaned from studies in which intake of particular foods in experimental subjects (including humans) was linked to the urinary excretion of various compounds. Results of such studies were consistent with the idea that compounds could be transformed into other compounds, but clearly, nutrients were not wholly interchangeable.

At the University of Illinois, William C. Rose focused his research on nutritional studies to decipher the metabolic relationships of nitrogenous compounds. Among other things, his studies of rat growth and nutrition helped show that purines and pyrimidines were derived from amino acids but that those compounds could not replace dietary amino acids.

In order to examine the nutritional requirements for individual amino acids, Rose hydrolyzed proteins to obtain their component amino acids and then selectively removed certain amino acids. In one of his first experiments, he removed arginine and histidine from a hydrolysate of the milk protein casein. Rats fed on this preparation lost weight unless the amino acid histidine was added back to the food. However, adding back arginine did not compensate for the apparent requirement for histidine. These results prompted Rose to investigate the requirements for all the amino acids. Using similar experimental approaches, Rose demonstrated that cysteine, histidine, and tryptophan could not be replaced by other amino acids.

From preparations based on hydrolyzed proteins, Rose moved to mixtures of pure amino acids. Thirteen of the 19 known amino acids could be purified, and the other six synthesized. However, rats fed these 19 amino acids as their sole source of dietary nitrogen lost weight. Although one possible explanation was that the proportions of the pure amino acids were not optimal, Rose concluded that there must be an additional essential amino acid, present in naturally occurring proteins and their hydrolysates but not in his amino acid mixtures.

After several years of effort, Rose obtained and identified the missing amino acid. In work published in 1935, Rose showed that adding this amino acid to the other 19 could support rat growth. Thus, the twentieth and last amino acid, threonine, was discovered.

Experiments extending over the next 20 years revealed that 10 of the 20 amino acids found in proteins are nutritionally essential, so that removal of one of these causes growth failure and eventually death in experimental animals. The other 10 amino acids were considered "dispensable" since animals could synthesize adequate amounts of them.

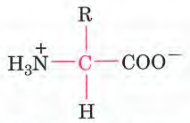
Rose's subsequent work included verifying the amino acid requirements of humans, using graduate students as subjects. Knowing which amino acids were required for normal health—and in what amounts—made it possible to evaluate the potential nutritive value of different types of food proteins. Eventually, these findings helped guide the formulations used for intravenous feeding.

McCoy, R.H., Meyer, C.E., and Rose, W.C., Feeding experiments with mixtures of highly purified amino acids. VIII. Isolation and identification of a new essential amino acid, *J. Biol. Chem.* **112,** 283–302 (1935). [Freely available at http://www.jbc.org.]

The sole exception is proline, which has a secondary amino group (—NH—), although for uniformity we shall refer to proline as an α-amino acid. The 20 standard amino acids differ in the structures of their side chains **(R groups).** Table 4-1 displays the names and complete chemical structures of the 20 standard amino acids.

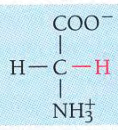
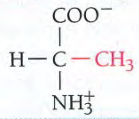
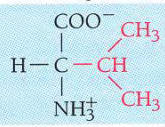
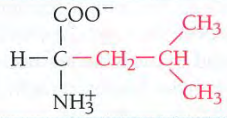
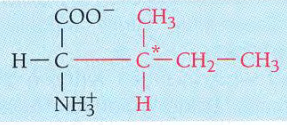
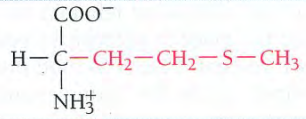
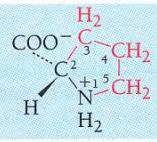
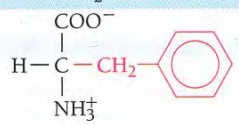
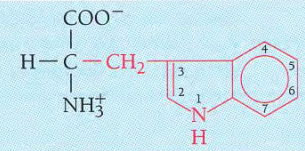
A | Amino Acids Are Dipolar Ions

The amino and carboxylic acid groups of amino acids readily ionize. The pK values of the carboxylic acid groups (represented by pK_1 in Table 4-1) lie in a small range around 2.2, while the pK values of the α-amino groups (pK_2) are near 9.4. *At physiological pH (~7.4), the amino groups are protonated and the carboxylic acid groups are in their conjugate base (carboxylate) form* (Fig. 4-2). An amino acid can therefore act as both an acid and a base.



■ **Figure 4-2 | A dipolar amino acid.** At physiological pH, the amino group is protonated and the carboxylic acid group is unprotonated.

Table 4-1 Covalent Structures and Abbreviations of the "Standard" Amino Acids of Proteins, Their Occurrence, and the p*K* Values of Their Ionizable Groups

Name, Three-letter Symbol, and One-letter Symbol	Structural Formula[a]	Residue Mass (D)[b]	Average Occurrence in Proteins (%)[c]	pK_1 α-COOH[d]	pK_2 α-NH_3^+[d]	pK_R Side Chain[d]
Amino acids with nonpolar side chains						
Glycine Gly G		57.0	7.2	2.35	9.78	
Alanine Ala A		71.1	7.8	2.35	9.87	
Valine Val V		99.1	6.6	2.29	9.74	
Leucine Leu L		113.2	9.1	2.33	9.74	
Isoleucine Ile I		113.2	5.3	2.32	9.76	
Methionine Met M		131.2	2.2	2.13	9.28	
Proline Pro P		97.1	5.2	1.95	10.64	
Phenylalanine Phe F		147.2	3.9	2.20	9.31	
Tryptophan Trp W		186.2	1.4	2.46	9.41	

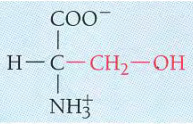
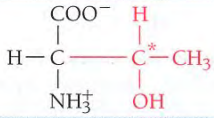
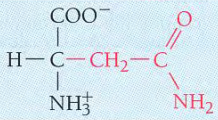
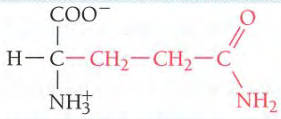
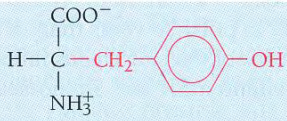
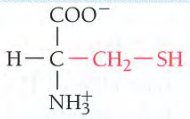
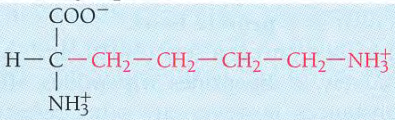
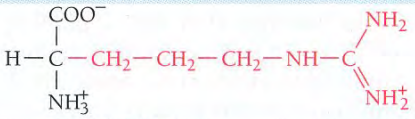
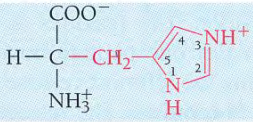
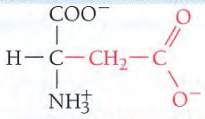
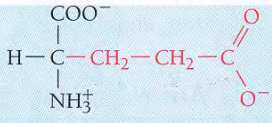
[a]The ionic forms shown are those predominating at pH 7.0 (except for that of histidine[f]) although residue mass is given for the neutral compound. The C_α atoms, as well as those atoms marked with an asterisk, are chiral centers with configurations as indicated according to Fischer projection formulas (Section 4-2). The standard organic numbering system is provided for heterocycles.

[b]The residue masses are given for the neutral residues. For the molecular masses of the parent amino acids, add 18.0 D, the molecular mass of H_2O, to the residue masses. For side chain masses, subtract 56.0 D, the formula mass of a peptide group, from the residue masses.

[c]Calculated from a database of nonredundant proteins containing 300,688 residues as compiled by Doolittle, R.F., *in* Fasman, G.D. (Ed.), *Predictions of Protein Structure and the Principles of Protein Conformation,* Plenum Press (1989).

[d]Data from Dawson, R.M.C., Elliott, D.C., Elliott, W.H., and Jones, K.M., *Data for Biochemical Research* (3rd ed.), pp. 1–31, Oxford Science Publications (1986).

Table 4-1 (continued)

Name, Three-letter Symbol, and One-letter Symbol	Structural Formula[a]	Residue Mass (D)[b]	Average Occurrence in Proteins (%)[c]	pK_1 α-COOH[d]	pK_2 α-NH_3^+[d]	pK_R Side Chain[d]
Amino acids with uncharged polar side chains						
Serine Ser S		87.1	6.8	2.19	9.21	
Threonine Thr T		101.1	5.9	2.09	9.10	
Asparagine[e] Asn N		114.1	4.3	2.14	8.72	
Glutamine[e] Gln Q		128.1	4.3	2.17	9.13	
Tyrosine Tyr Y		163.2	3.2	2.20	9.21	10.46 (phenol)
Cysteine Cys C		103.1	1.9	1.92	10.70	8.37 (sulfhydryl)
Amino acids with charged polar side chains						
Lysine Lys K		128.2	5.9	2.16	9.06	10.54 (ε-NH_3^+)
Arginine Arg R		156.2	5.1	1.82	8.99	12.48 (guanidino)
Histidine[f] His H		137.1	2.3	1.80	9.33	6.04 (imidazole)
Aspartic acid[e] Asp D		115.1	5.3	1.99	9.90	3.90 (β-COOH)
Glutamic acid[e] Glu E		129.1	6.3	2.10	9.47	4.07 (γ-COOH)

[a]The three- and one-letter symbols for asparagine *or* aspartic acid are Asx and B, whereas for glutamine *or* glutamic acid they are Glx and Z. The one-letter symbol for an undetermined or "nonstandard" amino acid is X.

[f]Both neutral and protonated forms of histidine are present at pH 7.0, since its pK_R is close to 7.0.

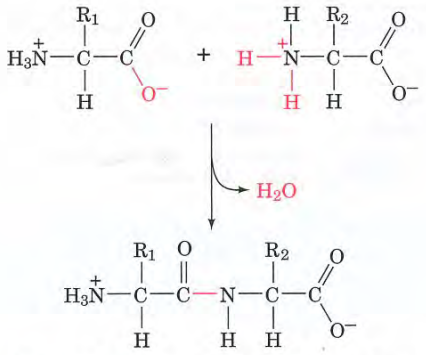


Figure 4-3 | Condensation of two amino acids. Formation of a CO—NH bond with the elimination of a water molecule produces a dipeptide. The peptide bond is shown in red. The residue with a free amino group is the N-terminus of the peptide, and the residue with a free carboxylate group is the C-terminus.

Table 4-1 also lists the pK values for the seven side chains that contain ionizable groups (pK_R).

Molecules such as amino acids, which bear charged groups of opposite polarity, are known as **dipolar ions** or **zwitterions.** Amino acids, like other ionic compounds, are more soluble in polar solvents than in nonpolar solvents. As we shall see, the ionic properties of the side chains influence the physical and chemical properties of free amino acids and amino acids in proteins.

B | Peptide Bonds Link Amino Acids

Amino acids can be polymerized to form chains. This process can be represented as a **condensation reaction** (bond formation with the elimination of a water molecule), as shown in Fig. 4-3. The resulting CO—NH linkage, an amide linkage, is known as a **peptide bond.**

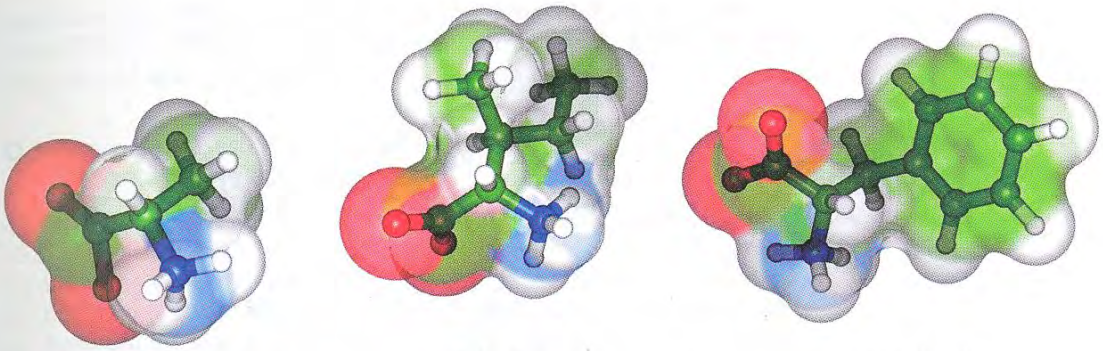
Polymers composed of two, three, a few (3–10), and many amino acid units are known, respectively, as **dipeptides**, **tripeptides**, **oligopeptides**, and **polypeptides.** These substances, however, are often referred to simply as "peptides." After they are incorporated into a peptide, the individual amino acids (the monomeric units) are referred to as amino acid **residues.**

Polypeptides are linear polymers rather than branched chains; that is, each amino acid residue participates in two peptide bonds and is linked to its neighbors in a head-to-tail fashion. The residues at the two ends of the polypeptide each participate in just one peptide bond. The residue with a free amino group (by convention, the leftmost residue, as shown in Fig. 4-3) is called the **amino terminus** or **N-terminus.** The residue with a free carboxylate group (at the right) is called the **carboxyl terminus** or **C-terminus.**

Proteins are molecules that contain one or more polypeptide chains. *Variations in the length and the amino acid sequence of polypeptides are major contributors to the diversity in the shapes and biological functions of proteins,* as we shall see in succeeding chapters.

C | Amino Acid Side Chains Are Nonpolar, Polar, or Charged

The most useful way to classify the 20 standard amino acids is by the polarities of their side chains. According to the most common classification scheme, there are three major types of amino acids: (1) those with



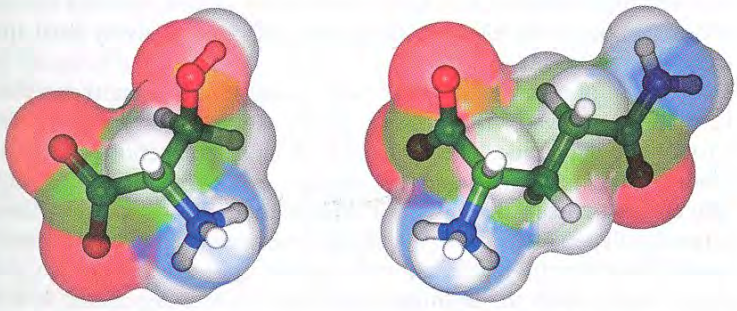
Alanine Isoleucine Phenylalanine

Figure 4-4 | Some amino acids with nonpolar side chains. The amino acids are shown as ball-and-stick models embedded in transparent space-filling models. The atoms are colored according to type with C green, H white, N blue, and O red.

nonpolar R groups, (2) those with uncharged polar R groups, and (3) those with charged polar R groups.

The Nonpolar Amino Acid Side Chains Have a Variety of Shapes and Sizes. Nine amino acids are classified as having nonpolar side chains. The three-dimensional shapes of some of these amino acids are shown in Fig. 4-4. **Glycine** has the smallest possible side chain, an H atom. **Alanine, valine, leucine,** and **isoleucine** have aliphatic hydrocarbon side chains ranging in size from a methyl group for alanine to isomeric butyl groups for leucine and isoleucine. **Methionine** has a thioether side chain that resembles an *n*-butyl group in many of its physical properties (C and S have nearly equal electronegativities, and S is about the size of a methylene group). **Proline** has a cyclic pyrrolidine side group. **Phenylalanine** (with its phenyl moiety) and **tryptophan** (with its indole group) contain aromatic side groups, which are characterized by bulk as well as nonpolarity.

Uncharged Polar Side Chains Have Hydroxyl, Amide, or Thiol Groups. Six amino acids are commonly classified as having uncharged polar side chains (Table 4-1 and Fig. 4-5). **Serine** and **threonine** bear hydroxylic R groups of different sizes. **Asparagine** and **glutamine** have amide-bearing side chains of different sizes. **Tyrosine** has a phenolic group (and, like phenylalanine and tryptophan, is aromatic). **Cysteine** is unique among the



Serine Glutamine

Figure 4-5 | Some amino acids with uncharged polar side chains. Atoms are represented and colored as in Fig. 4-4. Note the presence of electronegative atoms on the side chains.

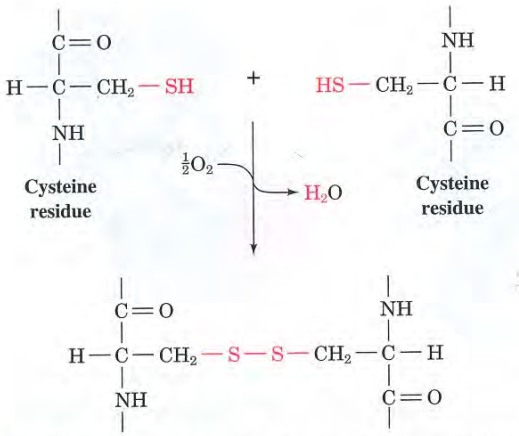


Figure 4-6 | Disulfide-bonded cysteine residues. The disulfide bond forms when the two thiol groups are oxidized.

20 amino acids in that it has a thiol group that can form a disulfide bond with another cysteine through the oxidation of the two thiol groups (Fig. 4-6).

Charged Polar Side Chains Are Positively or Negatively Charged. Five amino acids have charged side chains (Table 4-1 and Fig. 4-7). The side chains of the basic amino acids are positively charged at physiological pH values. **Lysine** has a butylammonium side chain, and **arginine** bears a guanidino group. As shown in Table 4-1, **histidine** carries an imidazolium moiety. Note that only histidine, with a pK_R of 6.04, readily ionizes within the physiological pH range. Consequently, both the neutral and cationic forms occur in proteins. In fact, the protonation–deprotonation of histidine side chains is a feature of numerous enzymatic reaction mechanisms.

The side chains of the acidic amino acids, **aspartic acid** and **glutamic acid,** are negatively charged above pH 3; in their ionized state, they are often referred to as **aspartate** and **glutamate.** Asparagine and glutamine are, respectively, the amides of aspartic acid and glutamic acid.

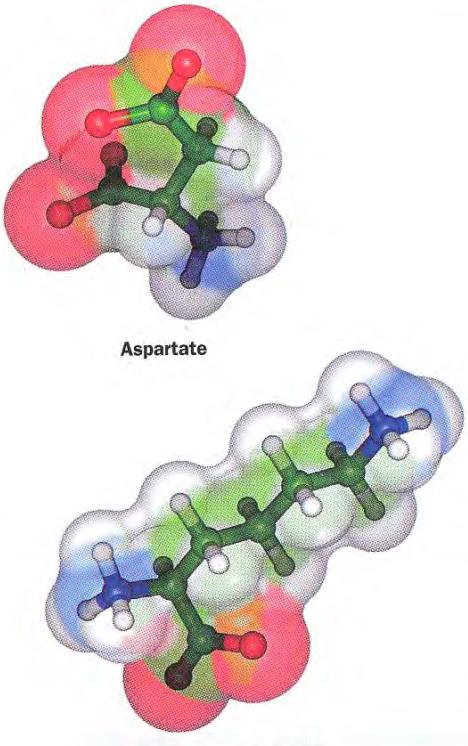


Figure 4-7 | Some amino acids with charged polar side chains. Atoms are represented and colored as in Fig. 4-4.

The allocation of the 20 amino acids among the three different groups is somewhat arbitrary. For example, glycine and alanine, the smallest of the amino acids, and tryptophan, with its heterocyclic ring, might just as well be classified as uncharged polar amino acids. Similarly, tyrosine and cysteine, with their ionizable side chains, might also be thought of as charged polar amino acids, particularly at higher pH values. In fact, the deprotonated side chain of cysteine (which contains the thiolate anion, S^-) occurs in a variety of enzymes, where it actively participates in chemical reactions.

Inclusion of a particular amino acid in one group or another reflects not just the properties of the isolated amino acid, but its behavior when it is part of a polypeptide. The structures of most polypeptides depend on a tendency for polar and ionic side chains to be hydrated and for nonpolar side chains to associate with each other rather than with water. This property of polypeptides is the hydrophobic effect (Section 2-1C) in action. As we shall see, the chemical and physical properties of amino acid side chains also govern the chemical reactivity of the polypeptide. It is worthwhile studying the structures of the 20 standard amino acids in order to

appreciate how they vary in polarity, acidity, aromaticity, bulk, conformational flexibility, ability to cross-link, ability to hydrogen bond, and reactivity toward other groups.

D | The p*K* Values of Ionizable Groups Depend on Nearby Groups

The α-amino acids have two or, for those with ionizable side chains, three acid–base groups. At very low pH values, these groups are fully protonated, and at very high pH values, these groups are unprotonated. At intermediate pH values, the acidic groups tend to be unprotonated, and the basic groups tend to be protonated. Thus, for the amino acid glycine, below pH 2.35 (the p*K* value of its carboxylic acid group), the $^{+}H_3NCH_2COOH$ form predominates. Above pH 2.35, the carboxylic acid is mostly ionized but the amino group is still mostly protonated ($^{+}H_3NCH_2COO^{-}$). Above pH 9.78 (the p*K* value of the amino group), the $H_2NCH_2COO^{-}$ form predominates. Note that *in aqueous solution, the un-ionized form (H_2NCH_2COOH) is present only in vanishingly small quantities.*

The pH at which a molecule carries no net electric charge is known as its **isoelectric point, p*I*.** For the α-amino acids,

$$pI = \tfrac{1}{2}(pK_i + pK_j) \qquad [4\text{-}1]$$

where K_i and K_j are the dissociation constants of the two ionizations involving the neutral species. For monoamino, monocarboxylic acids such as glycine, K_i and K_j represent K_1 and K_2. However, for aspartic and glutamic acids, K_i and K_j are K_1 and K_R, whereas for arginine, histidine, and lysine, these quantities are K_R and K_2.

Of course, amino acid residues in the interior of a polypeptide chain do not have free α-amino and α-carboxyl groups that can ionize (these groups are joined in peptide bonds; Fig. 4-3). Furthermore, the p*K* values of all ionizable groups, including the N- and C-termini, usually differ from the p*K* values listed in Table 4-1 for free amino acids. For example, the p*K* values of α-carboxyl groups in unfolded proteins range from 3.5 to 4.0. In the free amino acids, the p*K* values are much lower, because the positively charged ammonium group electrostatically stabilizes the COO^{-} group, in effect making it easier for the carboxylic acid group to ionize. Similarly, the p*K* values for α-amino groups in proteins range from 7.5 to 8.5. In the free amino acids, the p*K* values are higher, due to the electron-withdrawing character of the nearby carboxylate group, which makes it more difficult for the ammonium group to become deprotonated. In addition, the three-dimensional structure of a folded polypeptide chain may bring polar side chains and the N- and C-termini close together. The resulting electrostatic interactions between these groups may shift their p*K* values up to several pH units from the values for the corresponding free amino acids. For this reason, the p*I* of a polypeptide, which is a function of the p*K* values of its many ionizable groups, is not easily predicted and is usually determined experimentally.

E | Amino Acid Names Are Abbreviated

The three-letter abbreviations for the 20 standard amino acids given in Table 4-1 are widely used in the biochemical literature. Most of these abbreviations are taken from the first three letters of the name of the corresponding amino acid and are pronounced as written. The symbol **Glx**

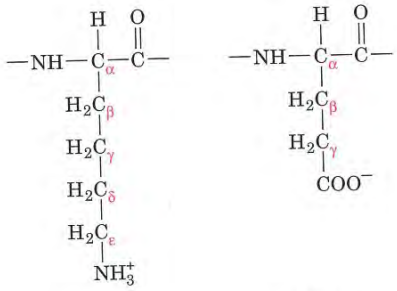
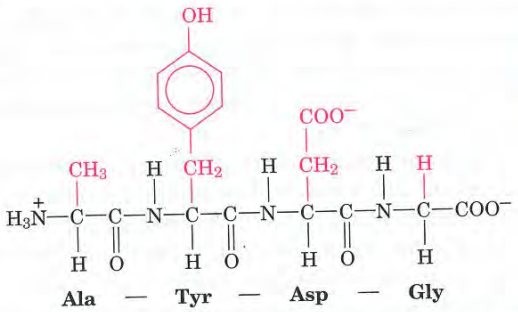


Figure 4-8 | Greek nomenclature for amino acids. The carbon atoms are assigned sequential letters in the Greek alphabet, beginning with the carbon next to the carbonyl group.

indicates Glu or Gln, and similarly, **Asx** means Asp or Asn. This ambiguous notation stems from laboratory experience: Asn and Gln are easily hydrolyzed to Asp and Glu, respectively, under the acidic or basic conditions often used to recover them from proteins. Without special precautions, it is impossible to tell whether a detected Glu was originally Glu or Gln, and likewise for Asp and Asn.

The one-letter symbols for the amino acids are also given in Table 4-1. This more compact code is often used when comparing the amino acid sequences of several similar proteins. Note that the one-letter symbol is usually the first letter of the amino acid's name. However, for sets of residues that have the same first letter, this is true only of the most abundant residue of the set.

Amino acid residues in polypeptides are named by dropping the suffix, usually **-ine,** in the name of the amino acid and replacing it by **-yl.** Polypeptide chains are described by starting at the N-terminus and proceeding to the C-terminus. The amino acid at the C-terminus is given the name of its parent amino acid. Thus, the compound



is called alanyltyrosylaspartylglycine. Obviously, such names for polypeptide chains of more than a few residues are extremely cumbersome. The tetrapeptide above can also be written as Ala-Tyr-Asp-Gly using the three-letter abbreviations, or AYDG using the one-letter symbols.

The various atoms of the amino acid side chains are often named in sequence with the Greek alphabet, starting at the carbon atom adjacent to the peptide carbonyl group. Therefore, as Fig. 4-8 indicates, the Lys residue is said to have an ε-amino group and Glu has a γ-carboxyl group. Unfortunately, this labeling system is ambiguous for several amino acids. Consequently, standard numbering schemes for organic molecules are also employed (and are indicated in Table 4-1 for the heterocyclic side chains).

2 Stereochemistry

With the exception of glycine, all the amino acids recovered from polypeptides are **optically active;** that is, they rotate the plane of polarized light. The direction and angle of rotation can be measured using an instrument known as a **polarimeter** (Fig. 4-9).

Optically active molecules are asymmetric; that is, they are not superimposable on their mirror image in the same way that a left hand is not

CHECK YOUR UNDERSTANDING

Describe the overall structure of an amino acid.
Be able to identify the peptide bonds, amino acid residues, and the N- and C-termini of a polypeptide.
Draw the structures of the 20 standard amino acids and give their one- and three-letter abbreviations.
Classify the 20 standard amino acids by polarity, structure, type of functional group, and acid–base properties.
Why do the p*K* values of ionizable groups differ between free amino acids and amino acid residues in polypeptides?

LEARNING OBJECTIVES

- Understand that amino acids and other biological compounds are chiral molecules whose configurations can be depicted by Fischer projections.
- Understand that amino acids in proteins all have the L stereochemical configuration.