

## 2

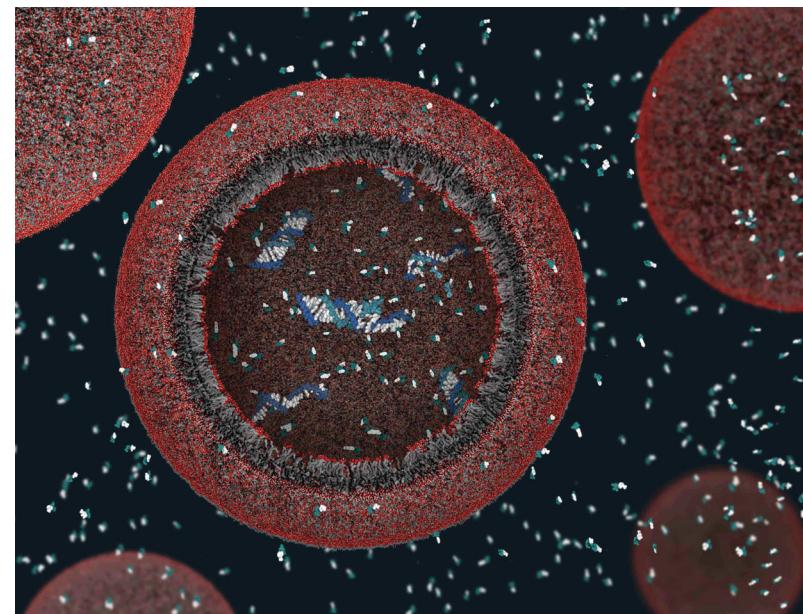
## The Chemical Basis of Life

## THE CHEMICAL ORIGIN OF LIFE

Around 4.6 billion years ago, our solar system was formed from an enormous rotating cloud of gas and dust. For its first billion years, Earth was a tumultuous place, with violent volcanic eruptions and near constant collisions with asteroids. Yet it is during this very time period that scientists believe that life, in the form of primitive cells, first appeared on the planet. Ancient microbes are thought to have formed stratified rock formations, called stromatolites, suggesting that life may have proliferated as early as 3.5 billion years ago.

Researchers have hypothesized that the earliest cells (called protocells) were very simple, made up of just nucleic acids (such as DNA or RNA) surrounded by a membrane, and that these cells may have formed in warm pools of water, or in the ocean near deep-sea vents. A central dilemma, however, has been in understanding how nucleic acid and membrane molecules formed spontaneously on the young planet—how, essentially, biology was born from chemistry.

In 1952, Harold Urey and his graduate student, Stanley Miller, designed an experiment to test whether the



Building life from scratch: computer rendering of a proto-cell, an artificial lipid vesicle containing self-replicating nucleic acids. The ability of such protocells to self-assemble and replicate when supplied with the right mix of chemical building blocks underscores the critical role of chemistry in the origins of life.

SOURCE: Image created by Janet Iwasa. Used by permission.

continued

- 2.1 Covalent Bonds
- 2.2 THE HUMAN PERSPECTIVE:  
Do Free Radicals Cause Aging?
- 2.3 Noncovalent Bonds
- 2.4 Acids, Bases, and Buffers
- 2.5 The Nature of Biological Molecules
- 2.6 Carbohydrates
- 2.7 Lipids
- 2.8 Building Blocks of Protein

- 2.9 Primary and Secondary Structures of Proteins
- 2.10 Tertiary Structure of Proteins
- 2.11 Quaternary Structure of Proteins
- 2.12 Protein Folding
- 2.13 THE HUMAN PERSPECTIVE:  
Protein Misfolding Can Have Deadly Consequences

## CHAPTER OUTLINE

- 2.14 EXPERIMENTAL PATHWAYS:  
Chaperones—Helping Proteins Reach Their Proper Folded State
- 2.15 Proteomics and Interactomics
- 2.16 Protein Engineering
- 2.17 Protein Adaptation and Evolution
- 2.18 Nucleic Acids
- 2.19 The Formation of Complex Macromolecular Structures

**32** conditions on the early Earth favored the spontaneous synthesis of biological molecules. They simulated Earth's early atmosphere by circulating water, methane, ammonia, and hydrogen in a sealed glass apparatus, and introducing energy in the form of heat and electricity (which mimicked the effect of lightning). Over the course of two weeks, the glass became coated with organic compounds that included a variety amino acids and sugars, supporting the idea that the conditions on the young planet may have been ideal for creating the organic compounds that were eventually incorporated into early cells.

In recent years, researchers have found prebiotically feasible pathways to create a number of additional molecules,

including ribonucleotides and fatty acids, the building blocks of RNA and membranes. Some laboratories are now attempting to find conditions that would allow for the formation of protocells.

There are many mysteries that still remain. Molecules created by the Miller-Urey experiment tend to be an even mixture of left- and right-handed isomers (which are mirror images of each other), but all life on Earth use only left-handed amino acids and right-handed sugars. How and why did this selection come about?

## 2.1 Covalent Bonds

We will begin this chapter with a brief examination of the atomic basis of living matter. The cellular level of organization is only a small step from the atomic level, as will become evident when we examine the importance of the movement of a few atoms of a molecule during such activities as muscle contraction or the transport of substances across cell membranes. The properties of cells and their organelles derive directly from the activities of the molecules of which they are composed. Consider a process such as cell division, which can be followed in considerable detail under a simple light microscope. To understand the activities that occur when a cell divides, one needs to know, for example, about the interactions between DNA and protein molecules that cause the chromosomes to condense into rod-shaped packages that can be separated into different cells; the molecular construction of protein-containing microtubules that allows them to disassemble at one moment in the cell and reassemble the next moment in an entirely different cellular location; and the properties of lipid molecules that make the outer cell membrane deformable so that it can be pulled into the middle of a cell, thereby pinching the cell in two. It is impossible even to begin to understand cellular function without a reasonable knowledge of the structure and properties of the major types of biological molecules. This is the goal of the present chapter: to provide an understanding of the chemistry that allows biological events to occur, giving rise to life. We will begin by considering the types of bonds that atoms can form with one another.

The atoms that make up a molecule are joined together by **covalent bonds** in which pairs of electrons are shared between pairs of atoms. The formation of a covalent bond between two atoms is governed by the fundamental principle that an atom is most stable when its outermost electron shell is filled. Consequently, the number of bonds an atom can form depends on the number of electrons needed to fill its outer shell.

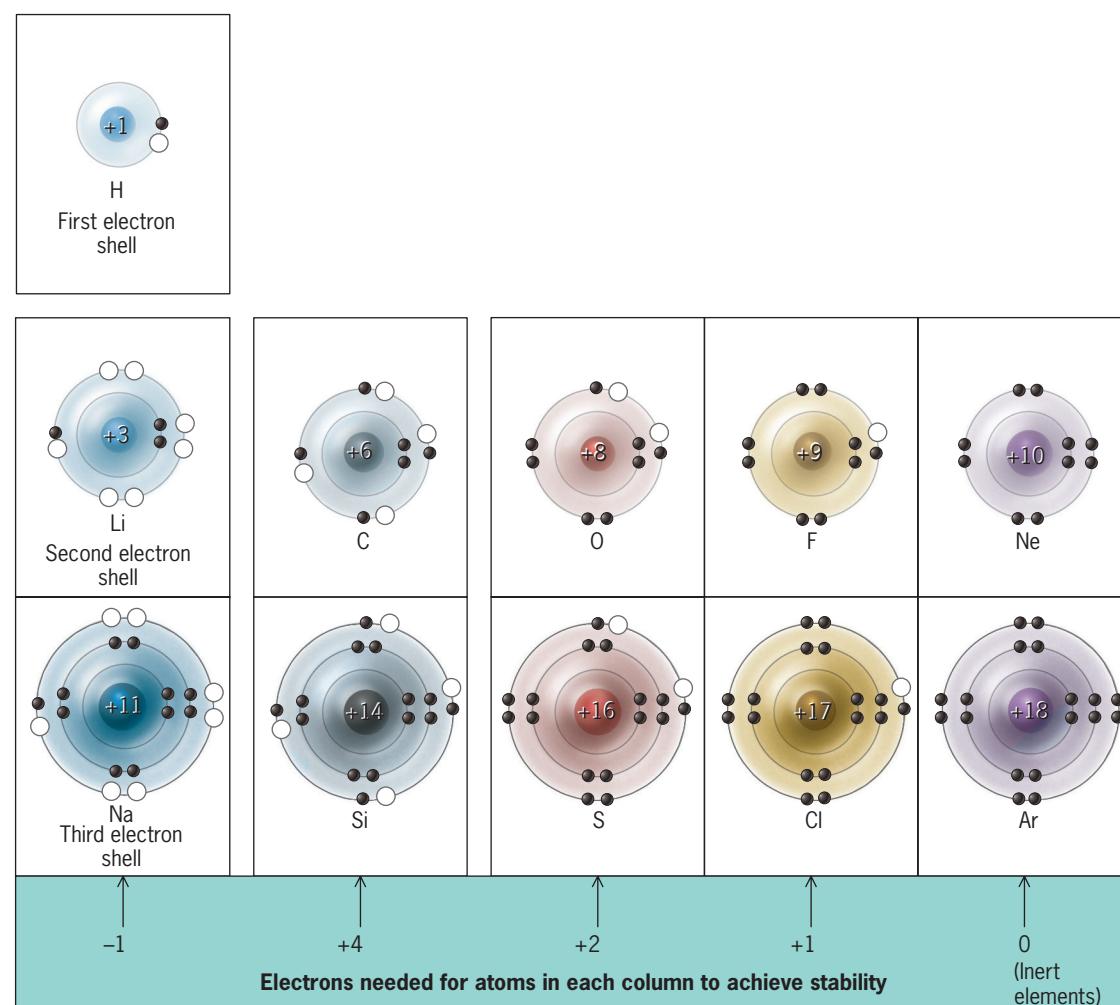
The electronic structure of a number of atoms is shown in **FIGURE 2.1**. The outer (and only) shell of a hydrogen or helium atom is filled when it contains two electrons; the outer shells of the other atoms in Figure 2.1 are filled when they contain eight electrons. Thus, an oxygen atom, with six outer-shell electrons, can fill its outer shell by combining with two hydrogen atoms, forming a molecule of water. The oxygen atom is linked to each hydrogen atom by a *single* covalent bond (denoted as H:O or H—O). The formation of a covalent

bond is accompanied by the release of energy, which must be reabsorbed at some later time if the bond is to be broken. The energy required to cleave C—H, C—C, or C—O covalent bonds is quite large—typically between 80 and 100 kilocalories per mole (kcal/mol)<sup>1</sup> of molecules. By comparison, the thermal energy of a molecule is only 0.6 kcal/mol. The thermal vibrations acting on a molecule are thus far too weak to break a covalent bond, making these bonds stable under most conditions. In this chapter, when we speak of bonds being strong, we mean that the energy required to break the bond is much greater than the thermal energy of the molecule. Conversely, when we talk about bonds being weak, we mean that the energy required to break the bond is of the same magnitude or smaller than the thermal energy.

In many cases, two atoms can become joined by bonds in which more than one pair of electrons are shared. If two electron pairs are shared, as occurs in molecular oxygen (O<sub>2</sub>), the covalent bond is a *double bond*, and if three pairs of electrons are shared (as in molecular nitrogen, N<sub>2</sub>), it is a *triple bond*. Quadruple bonds are not known to occur. The type of bond between atoms has important consequences in determining the shapes of molecules. For example, atoms joined by a single bond are able to rotate relative to one another, whereas the atoms of double (and triple) bonds lack this ability. As illustrated in Figure 6.6, double bonds can function as energy-capturing centers, driving such vital processes as respiration and photosynthesis.

When atoms of the same element bond to one another, as in H<sub>2</sub>, the electron pairs of the outer shell are equally shared between the two bonded atoms. When two unlike atoms are covalently bonded, however, the positively charged nucleus of one atom exerts a greater attractive force on the outer electrons than the other. Consequently, the shared electrons tend to be located more closely to the atom with the greater attractive force, that is, the more **electronegative atom**. Among the atoms most commonly present in biological molecules, nitrogen and oxygen are strongly electronegative.

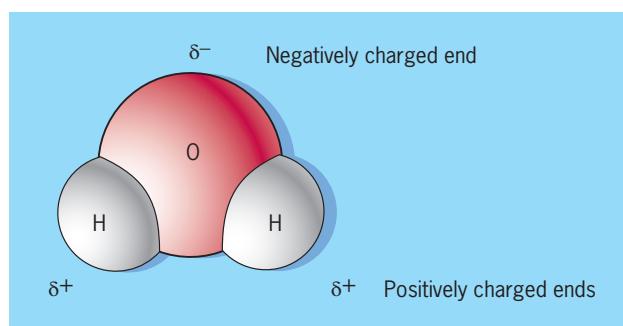
<sup>1</sup>One calorie is the amount of thermal energy required to raise the temperature of one gram of water one degree Celsius. One kilocalorie (kcal) equals 1000 calories (or one large Calorie). In addition to calories, energy can also be expressed in Joules, which is a term that was used historically to measure energy in the form of work. One kilocalorie is equivalent to 4186 Joules. Conversely, 1 Joule = 0.239 calories. A mole is equal to Avogadro's number ( $6 \times 10^{23}$ ) of molecules. A mole of a substance is its molecular weight expressed in grams.



**FIGURE 2.1** A representation of the arrangement of electrons in a number of common atoms. Electrons are present around an atom's nucleus in "clouds" or orbitals that are roughly defined by their boundaries, which may have a spherical or dumbbell shape. Each orbital contains a maximum of two electrons, which is why the electrons (dark dots in the drawing) are grouped in pairs. The innermost shell contains a single orbital (thus two electrons), the second shell contains four orbitals (thus eight electrons), and the third shell also contains four orbitals. The number of outer-shell electrons is a primary determinant of the chemical properties of an element. Atoms with a similar number of outer-shell electrons have similar properties. Lithium (Li) and sodium (Na), for example, have one outer-shell electron, and both are highly reactive metals. Carbon (C) and silicon (Si) atoms can each bond with four different atoms. Because of its size, however, a carbon atom can bond to other carbon atoms, forming long-chained organic molecules, whereas silicon is unable to form comparable molecules. Neon (Ne) and argon (Ar) have filled outer shells, making these atoms highly nonreactive; they are referred to as inert gases.

## Polar and Nonpolar Molecules

Let's examine a molecule of water. Water's single oxygen atom attracts electrons much more forcefully than do either of its hydrogen atoms. As a result, the O—H bonds of a water molecule are said to be *polarized*, such that one of the atoms has a partial negative charge and the other a partial positive charge. This is generally denoted in the following manner:



Molecules, such as water, that have an asymmetric distribution of charge (or *dipole*) are referred to as **polar** molecules. Polar molecules of biological importance contain one or more electronegative atoms, usually O, N, and/or S. Molecules that lack electronegative atoms and strongly polarized bonds, such as molecules that consist entirely of carbon and hydrogen atoms, are said to be **nonpolar**. The presence of strongly polarized bonds is of utmost importance in determining the reactivity of molecules. Large nonpolar molecules, such as waxes and fats, are relatively inert. Some of the more interesting biological molecules, including proteins and phospholipids, contain both polar and nonpolar regions, which behave very differently.

## Ionization

Some atoms are so strongly electronegative that they can capture electrons from other atoms during a chemical reaction. For example, when the elements sodium (a silver-colored metal) and chlorine (a toxic gas) are mixed, the single electron in the outer shell of each

**34** sodium atom migrates to the electron-deficient chlorine atom. As a result, these two atoms are transformed into charged **ions**.



Because the chloride ion has an extra electron (relative to the number of protons in its nucleus), it has a negative charge ( $\text{Cl}^{-}$ ) and is termed an **anion**. The sodium atom, which has lost an electron, has an extra positive charge ( $\text{Na}^{+}$ ) and is termed a **cation**. When present in crystals, these two ions form sodium chloride, or table salt.

The  $\text{Na}^{+}$  and  $\text{Cl}^{-}$  ions depicted above are relatively stable because they possess filled outer shells. A different arrangement of electrons within an atom can produce a highly reactive species, called a *free radical*. The structure of free radicals and their importance in biology are considered in the accompanying Human Perspective.

## REVIEW

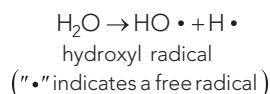
- Oxygen atoms have eight protons in their nucleus. How many electrons do they have? How many orbitals are in the inner electron shell? How many electrons are in the outer shell? How many more electrons can the outer shell hold before it is filled?
- Compare and contrast: a sodium atom and a sodium ion; a single bond and a double bond; an atom of weak and strong electronegativity; the electron distribution around an oxygen atom bound to another oxygen atom and an oxygen atom bound to two hydrogen atoms.

## 2.2 THE HUMAN PERSPECTIVE

### Do Free Radicals Cause Aging?

During the course of this textbook, we will discuss several different biological factors that are thought to contribute to the process of aging. Here we will consider one factor that has long been imagined to drive aging: the gradual accumulation of damage to our body's tissues. This idea is appealing because it introduces a natural element of time—as long as damage occurs at some low constant rate, the longer you live, the more damage you accumulate. But while this idea is appealing, actually testing it requires knowing the source of the damage, and then asking whether changing the rate of damage actually does alter aging. What is getting damaged, and what is doing the damage? The most destructive damage probably occurs to DNA. Alterations in DNA lead to the production of faulty genetic messages that promote gradual cellular deterioration. How does cellular damage occur, and why should it occur more rapidly in a shorter-lived animal, such as a chimpanzee, than a human? The answer may reside at the atomic level.

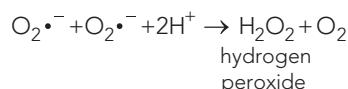
Atoms are stabilized when their shells are filled with electrons. Electron shells consist of orbitals, each of which can hold a maximum of two electrons. Atoms or molecules that have orbitals containing a single unpaired electron tend to be highly unstable—they are called **free radicals**. Free radicals may be formed when a covalent bond is broken such that each portion keeps one-half of the shared electrons, or they may be formed when an atom or molecule accepts a single electron transferred during an oxidation-reduction reaction. For example, water can be converted into free radicals when exposed to radiation from the sun:



Free radicals are extremely reactive and capable of chemically altering many types of molecules, including proteins, nucleic acids, and lipids. This is illustrated by the fact that certain cells of the immune system generate free radicals within their cytoplasm as a means to kill bacteria that these immune cells have ingested.

The formation of hydroxyl radicals is probably a major reason that sunlight is so damaging to skin.

In 1956, Denham Harman of the University of Nebraska proposed that aging results from tissue damage caused by free radicals. Because the subject of free radicals was not one with which biologists and physicians were familiar, the proposal failed to generate significant interest. Then, in 1969, Joe McCord and Irwin Fridovich of Duke University discovered an enzyme, superoxide dismutase (SOD), whose sole function was the destruction of the superoxide radical ( $\text{O}_2^{\cdot-}$ ), a type of free radical formed when molecular oxygen picks up an extra electron. SOD catalyzes the following reaction:



Hydrogen peroxide is also a potentially reactive oxidizing agent, which is why it is often used as a disinfectant and bleaching agent. If it is not rapidly destroyed,  $\text{H}_2\text{O}_2$  can break down to form hydroxyl radicals that attack the cell's macromolecules. Hydrogen peroxide is normally destroyed in the cell by the enzymes catalase or glutathione peroxidase.

Subsequent research has revealed that superoxide radicals are formed within cells during normal oxidative metabolism and that a superoxide dismutase is present in the cells of diverse organisms, from bacteria to humans. In fact, animals possess three different versions (isoforms) of SOD: a cytosolic, mitochondrial, and extracellular isoform. It is estimated that as much as 1–2 percent of the oxygen taken into human mitochondria can be converted to hydrogen peroxide rather than to water, the normal end product of respiration. The importance of SOD is most clearly revealed in studies of mutant bacteria and yeast that lack the enzyme; these cells are unable to grow in the presence of oxygen. Similarly, mice that are lacking the mitochondrial version of the enzyme (SOD2) are not able to survive more than a week or so after birth. Conversely, mice that have been genetically engineered so that their mitochondria contain elevated levels of the  $\text{H}_2\text{O}_2$ -destroying enzyme catalase live 20 percent longer,

on average, than untreated controls. This finding, reported in 2005, marked the first demonstration that enhanced antioxidant defenses can increase the life span of a mammal. Although the destructive potential of free radicals, such as superoxide and hydroxyl radicals, is unquestioned, the importance of these agents as a factor in aging remains controversial. In some cases, perturbations that increase oxygen radicals were found to increase lifespan rather than decrease it. Moreover, the whole concept that aging involves accumulation of random damage has been challenged by the discovery of specific genes in model organisms like yeast and nematode worms that, when mutated, allow the organism to live much longer. The existence of such genes has led to a competing theory known as "programmed aging," which posits that organisms have evolved mechanisms to induce their own decline

after they have passed reproductive age. The degree to which free radical damage determines aging is thus very much an open question.

A related area of research concerns the study of substances called antioxidants that are able to destroy free radicals in the test tube. The sale of these substances provides a major source of revenue for the vitamin/supplements industry. Common antioxidants found in the body include glutathione, vitamins E and C, and beta-carotene (the orange pigment in carrots and other vegetables). Although these substances may prove beneficial in the diet because of their ability to destroy free radicals, studies with rats and mice have failed to provide convincing evidence that they retard the aging process or increase maximum life span.

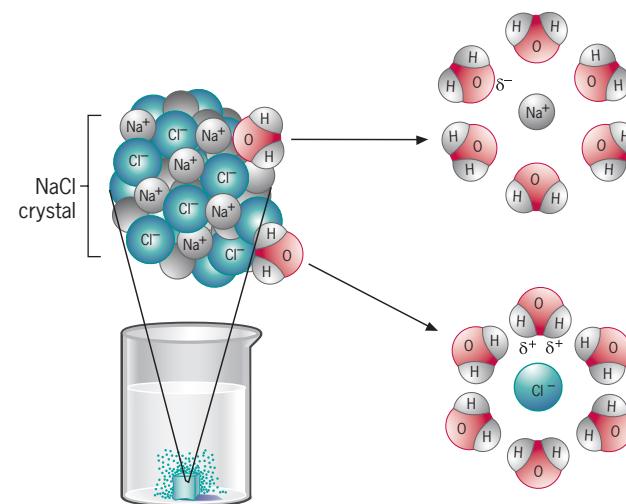
## 2.3 Noncovalent Bonds

Covalent bonds are strong bonds between the atoms that make up a molecule. Interactions between molecules (or between different parts of a large biological molecule) are governed by a variety of weaker linkages called *noncovalent bonds*. **Noncovalent bonds** do not depend on shared electrons but rather on attractive forces between atoms having an opposite charge. Individual noncovalent bonds are weak (about 1 to 5 kcal/mol) and are thus readily broken and reformed. As will be evident throughout this book, this feature allows noncovalent bonds to mediate the dynamic interactions among molecules in the cell.

Even though individual noncovalent bonds are weak, when large numbers of them act in concert, as between the two strands of a DNA molecule or between different parts of a large protein, their attractive forces are additive. Taken as a whole, they provide the structure with considerable stability. We will examine several types of noncovalent bonds that are important in cells.

### Ionic Bonds: Attractions between Charged Atoms

A crystal of table salt is held together by an electrostatic attraction between positively charged  $\text{Na}^+$  and negatively charged  $\text{Cl}^-$  ions. This type of attraction between fully charged components is called an **ionic bond** (or a *salt bridge*). Ionic bonds within a salt crystal may be quite strong. However, if a crystal of salt is dissolved in water, each of the individual ions becomes surrounded by water molecules, which inhibit oppositely charged ions from approaching one another closely enough to form ionic bonds (FIGURE 2.2). Because cells are composed primarily of water, bonds between *free* ions are of little importance. In contrast, weak ionic bonds between oppositely charged groups of large biological molecules are of considerable importance. For example, when negatively charged phosphate atoms in a DNA molecule are closely associated with positively charged groups on the surface of a protein (FIGURE 2.3), ionic bonds between them help hold the complex together. The strength of ionic bonds in a cell is generally weak (about 3 kcal/mol) due to the presence of water, but deep within the core of a protein, where water is often excluded, such bonds can be much stronger.

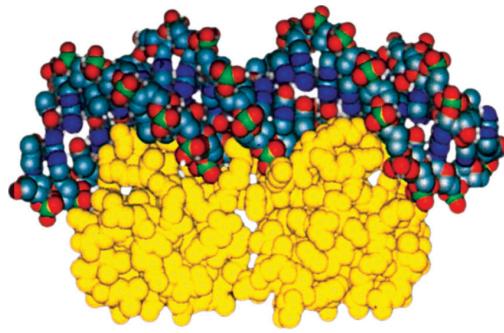
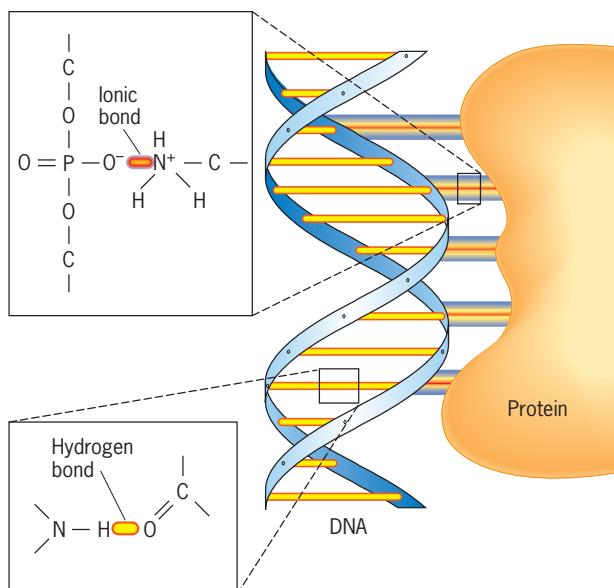


**FIGURE 2.2** The dissolution of a salt crystal. When placed in water, the  $\text{Na}^+$  and  $\text{Cl}^-$  ions of a salt crystal become surrounded by water molecules, breaking the ionic bonds between the two ions. As the salt dissolves, the negatively charged oxygen atoms of the water molecules associate with the positively charged sodium ions, and the positively charged hydrogen atoms of the water molecules associate with the negatively charged chloride ions.

### Hydrogen Bonds

When a hydrogen atom is covalently bonded to an electronegative atom, particularly an oxygen or a nitrogen atom, the single pair of shared electrons is greatly displaced toward the nucleus of the electronegative atom, leaving the hydrogen atom with a partial positive charge. This is illustrated on page 33 for the water molecule. As a result of this shift in charge, the bare, positively charged nucleus of the hydrogen atom can approach near enough to an unshared pair of outer electrons of a second electronegative atom to form an attractive interaction (FIGURE 2.4). This weak attractive interaction is called a **hydrogen bond**. Hydrogen bonds between or within molecules in solution have an energy of roughly 1 kcal/mol, which is similar in magnitude to the thermal energy. As a result, hydrogen bonds in biological molecules are easily broken.

Hydrogen bonds occur between most polar molecules and are particularly important in determining the structure and properties of water (discussed later). Hydrogen bonds also form between polar



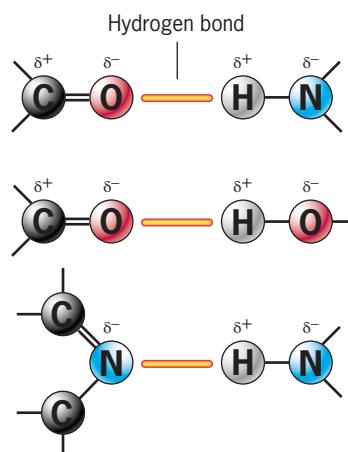
**FIGURE 2.3** Noncovalent ionic bonds play an important role in holding the protein molecule on the right (yellow atoms) to the DNA molecule on the left. Ionic bonds form between positively charged nitrogen atoms in the protein and negatively charged oxygen atoms in the DNA. The DNA molecule itself consists of two separate strands held together by noncovalent hydrogen bonds. Although a single noncovalent bond is relatively weak and easily broken, large numbers of these bonds between two molecules, as between two strands of DNA, make the overall complex quite stable.

SOURCE: Courtesy of Stephen Harrison, Harvard Biochemistry Department.

groups present in large biological molecules, as occurs between the two strands of a DNA molecule (see Figure 2.3). Because their strength is additive, the large number of hydrogen bonds between the strands makes the DNA duplex a stable structure. However, because individual hydrogen bonds are weak, the two strands can be partially separated to allow enzymes access to individual strands of the DNA molecule.

## Hydrophobic Interactions and van der Waals Forces

Because of their ability to interact with water, polar molecules, such as sugars and amino acids (described shortly), are said to be **hydrophilic**, or “water loving.” Nonpolar molecules, such as steroid or fat molecules, are essentially insoluble in water because they lack the charged regions that would attract them to the poles of water molecules. When nonpolar compounds are mixed with water, the nonpolar, **hydrophobic** (“water fearing”) molecules are forced into aggregates, which minimizes their exposure to the polar surroundings

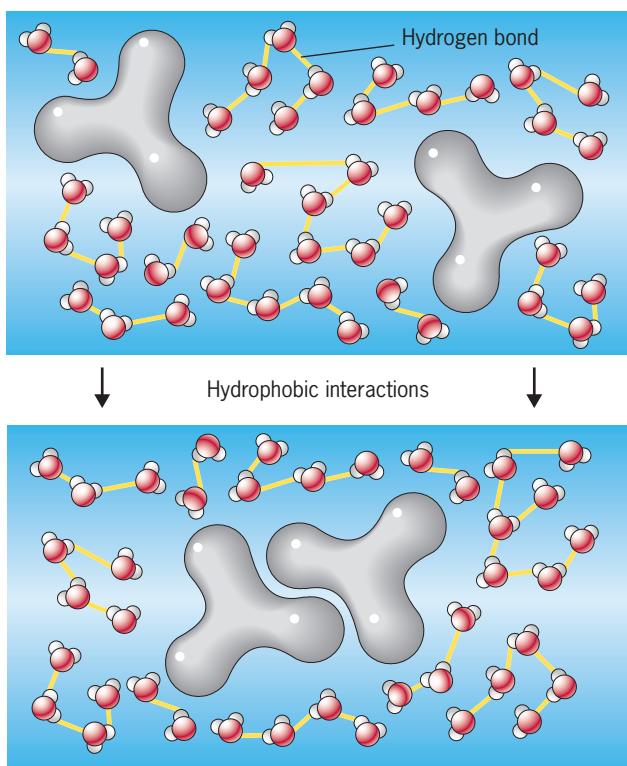


**FIGURE 2.4** Hydrogen bonds form between a bonded electronegative atom, such as nitrogen or oxygen, which bears a partial negative charge, and a bonded hydrogen atom, which bears a partial positive charge. Hydrogen bonds (about 0.18 nm) are typically about twice as long as the much stronger covalent bonds.

(FIGURE 2.5). This association of nonpolar molecules is called a **hydrophobic interaction**. This is why droplets of fat molecules rapidly reappear on the surface of beef or chicken soup even after the liquid is stirred with a spoon.

Hydrophobic interactions of the type just described are not classified as true bonds because they do not result from an attraction between hydrophobic molecules but rather from an energetic drive to exclude water away from the hydrophobic surfaces.<sup>2</sup> In addition to this type of interaction, hydrophobic groups can form weak bonds with one another based on electrostatic attractions. Polar molecules associate because they contain permanent asymmetric charge distributions within their structure. Closer examination of the covalent bonds that make up a nonpolar molecule (such as H<sub>2</sub> or CH<sub>4</sub>) reveals that electron distributions are not always symmetric. The distribution of electrons around an atom at any given instant is a statistical matter and, therefore, varies from one instant to the next. Consequently, at any given time, the electron density may happen to be greater on one side of an atom, even though the atom shares the electrons equally with some other atom. These transient asymmetries in electron distribution result in momentary separations of charge (*dipoles*) within the molecule. If two molecules with temporary dipoles are very close to one another and oriented in the appropriate manner, they experience a weak attractive force, called a **van der Waals force**, that bonds them together. Moreover, the formation of a temporary separation of charge in one molecule can *induce* a similar separation in an adjacent molecule. In this way, additional attractive forces can be generated between nonpolar molecules. A single van der Waals force is very weak (0.1 to 0.3 kcal/mol) and very sensitive to the distance that separates the two atoms (FIGURE 2.6a). As we will see in later chapters, however, biological molecules that interact with one another—for example, an antibody and a protein on the surface of a virus—often possess complementary

<sup>2</sup>This statement reflects an accepted hypothesis that hydrophobic interactions are driven by increased entropy (disorder). When a hydrophobic group projects into an aqueous solvent, the water molecules become ordered in a cage around the hydrophobic group. These solvent molecules become disordered when the hydrophobic group withdraws from the surrounding solvent. A discussion of this and other views can be found in *Nature* 437:640, 2005 and *Curr. Opin. Struct. Biol.* 16:152, 2006.



**FIGURE 2.5** In a hydrophobic interaction, the nonpolar (hydrophobic) molecules are forced into aggregates, which minimizes their exposure to the surrounding water molecules.

shapes. As a result, many atoms of both interactants may have the opportunity to approach each other very closely (Figure 2.6*b*), making van der Waals forces important in biological interactions.

### The Life-Supporting Properties of Water

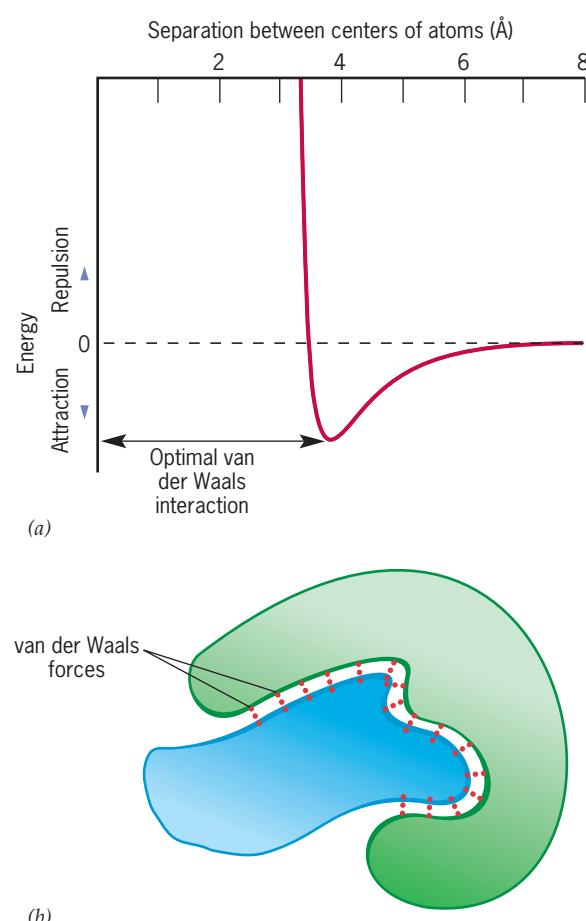
Life on Earth is totally dependent on water, and water may be essential to the existence of life anywhere in the universe. Even though it contains only three atoms, a molecule of water has a unique structure that gives the molecule extraordinary properties.<sup>3</sup> Most importantly,

1. Water is a highly asymmetric molecule with the O atom at one end and the two H atoms at the opposite end.
2. Each of the two covalent bonds in the molecule is highly polarized.
3. All three atoms in a water molecule are adept at forming hydrogen bonds.

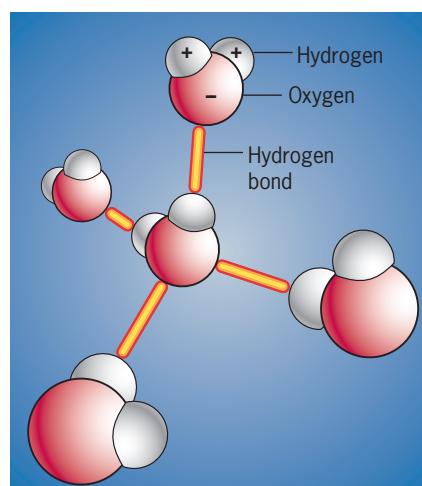
The life-supporting attributes of water stem from these properties.

Each molecule of water can form hydrogen bonds with as many as four other water molecules, producing a highly interconnected network of molecules (FIGURE 2.7). Each hydrogen bond is formed when the partially positive-charged hydrogen of one water molecule becomes aligned next to a partially negative-charged oxygen atom of another water molecule. Because of their extensive hydrogen bonding, water molecules have an unusually strong tendency to adhere to one another. This feature is most evident in the thermal properties of

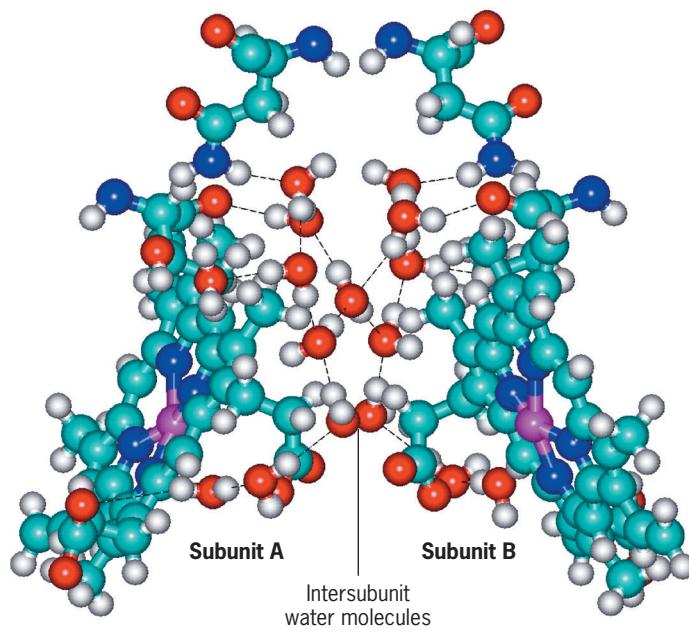
<sup>3</sup>One way to appreciate the structure of water is by comparing it to H<sub>2</sub>S. Like oxygen, sulfur has six outer-shell electrons and forms single bonds with two hydrogen atoms. But because sulfur is a larger atom, it is less electronegative than oxygen, and its ability to form hydrogen bonds is greatly reduced. At room temperature, H<sub>2</sub>S is a gas, not a liquid. In fact, the temperature has to drop to -86°C before H<sub>2</sub>S freezes into a solid.



**FIGURE 2.6** Van der Waals forces. (a) As two atoms approach each other, they experience a weak attractive force that increases up to a specific distance, typically about 4 Å. If the atoms approach more closely, their electron clouds repel one another, causing the atoms to be forced apart. (b) Although individual van der Waals forces are very weak and transient, large numbers of such attractive forces can be formed if two macromolecules have a complementary surface, as is indicated schematically in this figure (see Figure 2.40 for an example).



**FIGURE 2.7** Hydrogen bond formation between neighboring water molecules. Each H atom of the molecule has about four-tenths of a full positive charge, and the single O atom has about eight-tenths of a full negative charge.



**FIGURE 2.8** The importance of water in protein structure. The water molecules (each with a single red oxygen atom and two smaller gray hydrogen atoms) are shown in their ordered locations between the two subunits of a clam hemoglobin molecule.

SOURCE: From Martin Chaplin, *Nature Revs. Mol. Cell Biol.* 7:864, 2006, © 2006, by Macmillan Publishers Limited.

water. For example, when water is heated, most of the thermal energy is consumed in disrupting hydrogen bonds rather than contributing to molecular motion (which is measured as an increased temperature). Similarly, evaporation from the liquid to the gaseous state requires that water molecules break the hydrogen bonds holding them to their neighbors, which is why it takes so much energy to convert water to steam. Mammals take advantage of this property when they sweat because the heat required to evaporate the water is absorbed from the body, which thus becomes cooler.

The small volume of aqueous fluid present within a cell contains a remarkably complex mixture of dissolved substances, or *solutes*. In fact, water is able to dissolve more types of substances than any other solvent. But water is more than just a solvent; it determines the structure of biological molecules and the types of interactions in which they can engage. Water is the fluid matrix around which the insoluble fabric of the cell is constructed. It is also the medium through which materials move from one compartment of the cell to another; it is a reactant or product in many cellular reactions; and it protects the cell in many ways—from excessive heat, cold, or damaging radiation.

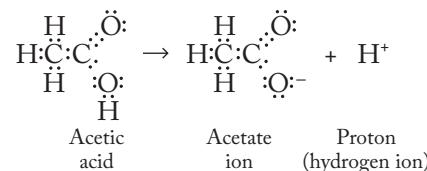
Water is such an important factor in a cell because it is able to form weak interactions with so many different types of chemical groups. Recall from page 35 how water molecules, with their strongly polarized O—H bonds, form a shell around ions, separating the ions from one another. Similarly, water molecules form hydrogen bonds with organic molecules that contain polar groups, such as amino acids and sugars, which ensure their solubility within the cell. Water also plays a key role in maintaining the structure and function of macromolecules and the complexes that they form (such as membranes). **FIGURE 2.8** shows the ordered arrangement of water molecules between two subunits of a protein molecule. The water molecules are hydrogen bonded to each other and to specific amino acids of the protein.

## REVIEW

1. Describe some of the properties that distinguish covalent and noncovalent bonds.
2. Why do polar molecules, such as table sugar, dissolve so readily in water? Why do fat droplets form on the surface of an aqueous solution? Why does sweating help cool the body?

## 2.4 Acids, Bases, and Buffers

Protons are not only found within atomic nuclei, they are also released into the medium whenever a hydrogen atom loses a shared electron. Consider acetic acid—the distinctive ingredient of vinegar—which can undergo the following reaction, described as a *dissociation*.



A molecule that is capable of releasing (donating) a hydrogen ion is termed an **acid**. The proton released by the acetic acid molecule in the previous reaction does not remain in a free state; instead, it combines with another molecule. Possible reactions involving a proton include

- Combination with a water molecule to form a hydronium ion ( $\text{H}_3\text{O}^+$ ).



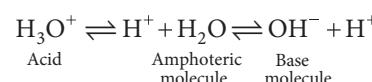
- Combination with a hydroxyl ion ( $\text{OH}^-$ ) to form a molecule of water.



- Combination with an amino group ( $-\text{NH}_2$ ) in a protein to form a charged amine.



Any molecule that is capable of accepting a proton is defined as a **base**. Acids and bases exist in pairs, or *couples*. When the acid loses a proton (as when acetic acid gives up a hydrogen ion), it becomes a base (in this case, acetate ion), which is termed the *conjugate base* of the acid. Similarly, when a base (such as an  $-\text{NH}_2$  group) accepts a proton, it forms an acid (in this case  $-\text{NH}_3^+$ ), which is termed the *conjugate acid* of that base. Thus, the acid always contains one more positive charge than its conjugate base. Water is an example of an *amphoteric* molecule, that is, one that can serve both as an acid and a base:



We will discuss another important group of amphoteric molecules, the amino acids, on page 49.

Acids vary markedly in the ease with which they give up a proton. The more readily the proton is lost, that is, the less strong the attraction of a conjugate base for its proton, the stronger the acid. Hydrogen

**TABLE 2.1** Strengths of Acids and Bases

	Acids		Bases	
	Very weak	Weak	Strong	Very weak
Very weak	$\text{H}_2\text{O}$	$\text{NH}_4^+$	$\text{OH}^-$	Strong
Weak	$\text{H}_2\text{S}$	$\text{CH}_3\text{COOH}$	$\text{NH}_3$	Weak
Strong	$\text{H}_3\text{O}^+$	$\text{H}_2\text{CO}_3$	$\text{H}_2\text{O}$	Very weak
	$\text{HCl}$		$\text{Cl}^-$	
	$\text{H}_2\text{SO}_4$		$\text{SO}_4^{2-}$	

chloride is a very strong acid, one that will readily transfer its proton to water molecules. The conjugate base of a strong acid, such as HCl, is a weak base (Table 2.1). Acetic acid, in contrast, is a relatively weak acid because for the most part it remains undissociated when dissolved in water. In a sense, one can consider the degree of dissociation of an acid in terms of the competition for protons among the components of a solution. Water is a better competitor, that is, a stronger base, than chloride ion, so HCl completely dissociates. In contrast, acetate ion is a stronger base than water, so it remains largely as undissociated acetic acid.

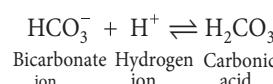
The acidity of a solution is measured by the concentration of hydrogen ions<sup>4</sup> and is expressed in terms of pH.

$$pH = -\log[H^+]$$

where  $[H^+]$  is the molar concentration of protons. For example, a solution having a pH of 5 contains a hydrogen ion concentration of  $10^{-5} M$ . Because the pH scale is logarithmic, an increase of one pH unit corresponds to a tenfold decrease in  $H^+$  concentration (or a tenfold increase in  $OH^-$  concentration). Stomach juice (pH 1.8), for example, has nearly one million times the  $H^+$  concentration of blood (pH 7.4).

A water molecule can dissociate into a hydroxyl ion and a proton,  $\text{H}_2\text{O} \rightarrow \text{H}^+ + \text{OH}^-$ . In pure water, the concentration of both  $\text{H}^+$  and  $\text{OH}^-$  is approximately  $10^{-7} \text{ M}$ . Pure water thus has a pH of 7.0 and since we view water as being the standard solvent into which other molecules are dissolved, it is common to refer to pH 7 as “neutral”.

Most biological processes are acutely sensitive to pH because changes in hydrogen ion concentration affect the ionic state of biological molecules. For example, as the hydrogen ion concentration increases, the  $-\text{NH}_2$  group of the amino acid arginine becomes protonated to form  $-\text{NH}_3^+$ , which can disrupt the activity of the entire protein. Even slight changes in pH can impede biological reactions. Organisms, and the cells they comprise, are protected from pH fluctuations by **buffers**—compounds that react with free hydrogen or hydroxyl ions, thereby resisting changes in pH. Buffer solutions usually contain a weak acid together with its conjugate base. Blood, for example, is buffered by carbonic acid and bicarbonate ions, which normally hold blood pH at about 7.4.



If the hydrogen ion concentration rises (as occurs during exercise), the bicarbonate ions combine with the excess protons, removing them from solution. Conversely, excess  $\text{OH}^-$  ions (which are generated during hyperventilation) are neutralized by protons derived from carbonic acid. The pH of the fluid within the cell is regulated in a similar manner by a phosphate buffer system consisting of  $\text{H}_2\text{PO}_4^-$  and  $\text{HPO}_4^{2-}$ .

<sup>4</sup>In aqueous solutions, protons do not exist in the free state, but rather as  $\text{H}_3\text{O}^+$  or  $\text{H}_5\text{O}_2^+$ . For the sake of simplicity, we will refer to them simply as protons or hydrogen ions.

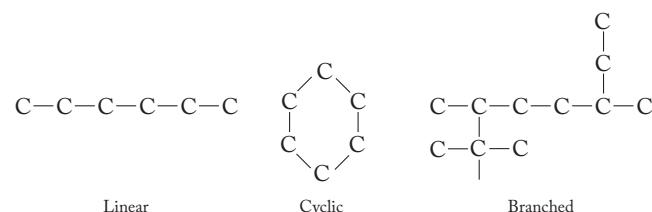
REVIEW

1. If you were to add hydrochloric acid to water, what effect would this have on the hydrogen ion concentration? on the pH? on the ionic charge of any proteins in solution?
  2. What is the relationship between a base and its conjugate acid?

## 2.5 The Nature of Biological Molecules

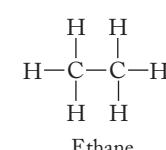
The bulk of an organism is water. If the water is evaporated away, most of the remaining dry weight consists of molecules containing atoms of carbon. When first discovered, it was thought that carbon-containing molecules were present only in living organisms and thus were referred to as *organic molecules* to distinguish them from *inorganic molecules* found in the inanimate world. As chemists learned to synthesize more and more of these carbon-containing molecules in the lab, the mystique associated with organic compounds disappeared. The compounds produced by living organisms are called **biochemicals**.

The chemistry of life centers around the chemistry of the carbon atom. The essential quality of carbon that has allowed it to play this role is the incredible number of molecules it can form. Having four outer-shell electrons, a carbon atom can bond with up to four other atoms. Most importantly, each carbon atom is able to bond with other carbon atoms so as to construct molecules with backbones containing long chains of carbon atoms. Carbon-containing backbones may be linear, branched, or cyclic.

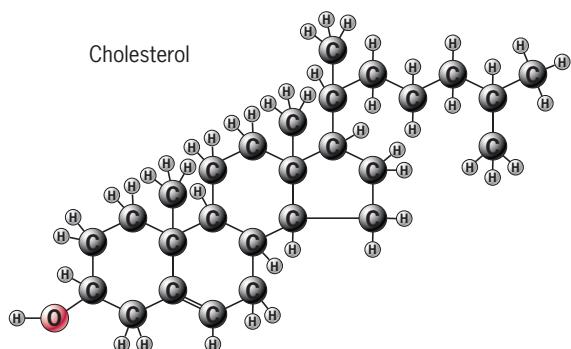


Cholesterol, whose structure is depicted in **FIGURE 2.9**, illustrates various arrangements of carbon atoms.

Both the size and electronic structure of carbon make it uniquely suited for generating large numbers of molecules, several hundred thousand of which are known. In contrast, silicon, which is just below carbon in the periodic table and also has four outer-shell electrons (see Figure 2.1), is too large for its positively charged nucleus to attract the outer-shell electrons of neighboring atoms with sufficient force to hold such large molecules together. We can best understand the nature of biological molecules by starting with the simplest group of organic molecules, the *hydrocarbons*, which contain only carbon and hydrogen atoms. The molecule ethane ( $C_2H_6$ ) is a simple hydrocarbon



consisting of two atoms of carbon in which each carbon is bonded to the other carbon as well as to three atoms of hydrogen. As more

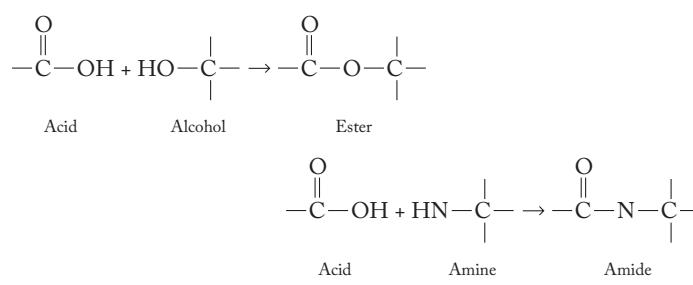


**FIGURE 2.9 Cholesterol**, whose structure illustrates how carbon atoms (represented by the black balls) are able to form covalent bonds with as many as four other carbon atoms. As a result, carbon atoms can be linked together to form the backbones of a virtually unlimited variety of organic molecules. The carbon backbone of a cholesterol molecule includes four rings, which is characteristic of steroids (e.g., estrogen, testosterone, cortisol). The cholesterol molecule shown here is drawn as a ball-and-stick model, which is another way that molecular structure is depicted.

carbons are added, the skeletons of organic molecules increase in length and their structure becomes more complex.

## Functional Groups

Hydrocarbons do not occur in significant amounts within most living cells (though they constitute the bulk of the fossil fuels formed from the remains of ancient plants and animals). Many of the organic molecules that are important in biology contain chains of carbon atoms, like hydrocarbons, but certain hydrogen atoms are replaced by various **functional groups**. Functional groups are particular groupings of atoms that often behave as a unit and give organic molecules their physical properties, chemical reactivity, and solubility in aqueous solution. Some of the more common functional groups are listed in Table 2.2. Two of the most common linkages between functional groups are **ester bonds**, which form between carboxylic acids and alcohols, and **amide bonds**, which form between carboxylic acids and amines.



**TABLE 2.2** Functional Groups

Methyl	Hydroxyl	Carboxyl	Amino	Phosphate	Carbonyl	Sulphydryl

Most of the groups in Table 2.2 contain one or more electronegative atoms (N, P, O, and/or S) and make organic molecules more polar, more water soluble, and more reactive. Several of these functional groups can ionize and become positively or negatively charged. The effect on molecules by the substitution of various functional groups is readily demonstrated. The hydrocarbon ethane ( $\text{CH}_3\text{CH}_3$ ) depicted on page 39 is a toxic, flammable gas. Replace one of the hydrogens with a hydroxyl group ( $-\text{OH}$ ) and the molecule ( $\text{CH}_3\text{CH}_2\text{OH}$ ) becomes palatable—it is ethyl alcohol (or ethanol). Substitute a carboxyl group ( $-\text{COOH}$ ) and the molecule becomes acetic acid ( $\text{CH}_3\text{COOH}$ ), the strong-tasting ingredient in vinegar. Substitute a sulphydryl group ( $-\text{SH}$ ), and you have formed  $\text{CH}_3\text{CH}_2\text{SH}$ , a strong, foul-smelling agent, ethyl mercaptan, used by biochemists in studying enzyme reactions.

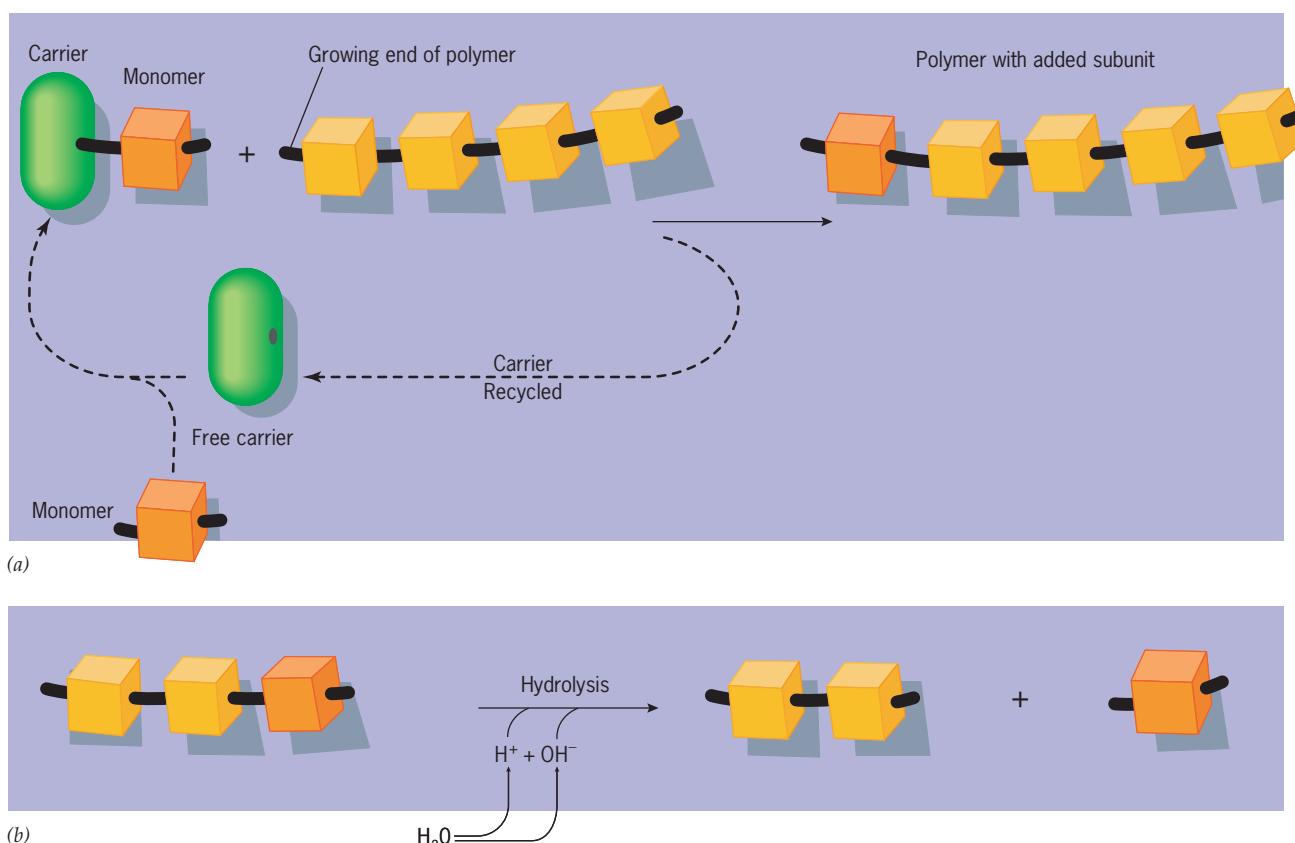
## A Classification of Biological Molecules by Function

The organic molecules commonly found within living cells can be divided into several categories based on their role in metabolism.

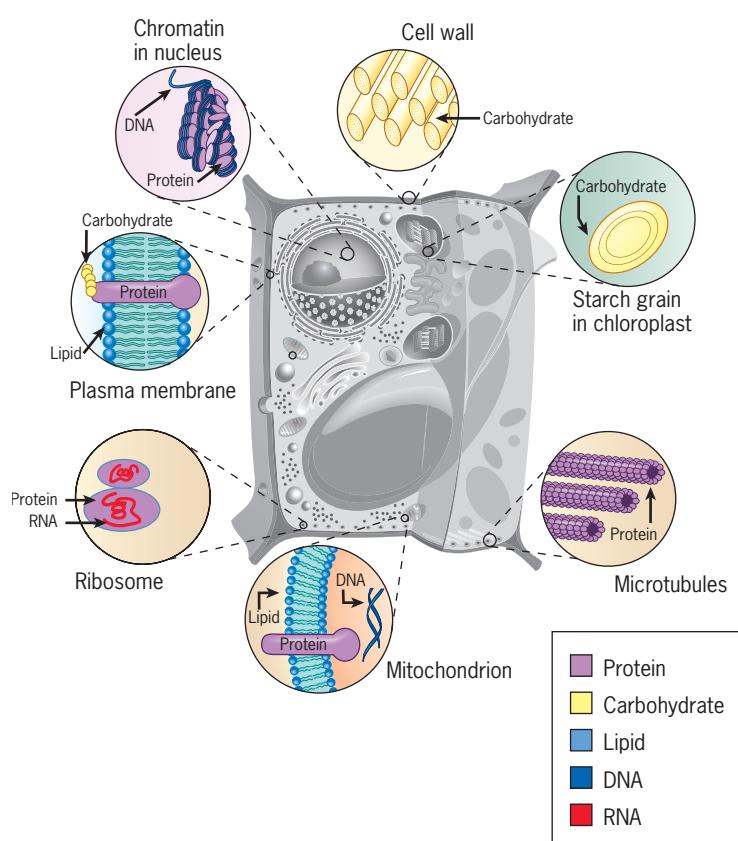
**1. Macromolecules.** The molecules that form the structure and carry out the activities of cells are huge, highly organized molecules called **macromolecules**, which contain anywhere from dozens to millions of carbon atoms. Because of their size and the intricate shapes that macromolecules can assume, some of these molecular giants can perform complex tasks with great precision and efficiency. The presence of macromolecules, more than any other characteristic, endows organisms with the properties of life and sets them apart chemically from the inanimate world.

Macromolecules can be divided into four major categories: proteins, nucleic acids, polysaccharides, and certain lipids. The first three types are *polymers* composed of a large number of low-molecular-weight building blocks, or *monomers*. These macromolecules are constructed from monomers by a process of *polymerization* that resembles coupling railroad cars onto a train (FIGURE 2.10). The basic structure and function of each type of macromolecule are similar in all organisms. It is not until you look at the specific sequences of monomers that make up individual macromolecules that the diversity among organisms becomes apparent. The localization of these molecules in a number of cellular structures is shown in an overview in FIGURE 2.11.

**2. The building blocks of macromolecules.** Most of the macromolecules within a cell have a short lifetime compared with the cell itself; with the exception of the cell's DNA, they are continually broken down and replaced by new macromolecules. Consequently, most cells contain a supply (or *pool*) of low-molecular-weight precursors that are ready to be incorporated into macromolecules. These include sugars, which are the precursors of polysaccharides; amino acids, which are the precursors of proteins; nucleotides, which are the precursors of nucleic acids; and fatty acids, which are incorporated into lipids.



**FIGURE 2.10** Monomers and polymers; polymerization and hydrolysis. (a) Polysaccharides, proteins, and nucleic acids consist of monomers (subunits) linked together by covalent bonds. Free monomers do not simply react with each other to become macromolecules. Rather, each monomer is first activated by attachment to a carrier molecule that helps the monomer to chemically react with the end of the growing macromolecule. (b) A macromolecule is disassembled by hydrolysis of the bonds that join the monomers together. Hydrolysis is the splitting of a bond by water. All of these reactions are catalyzed by specific enzymes.



**FIGURE 2.11** An overview of the types of biological molecules that make up various cellular structures.

**3. Metabolic intermediates (metabolites).** The molecules in a cell have complex chemical structures and must be synthesized in a step-by-step sequence beginning with specific starting materials. In the cell, each series of chemical reactions is termed a **metabolic pathway**. The cell starts with compound A and converts it to compound B, then to compound C, and so on, until some functional end product (such as an amino acid building block of a protein) is produced. The compounds formed along the pathways leading to the end products might have no function per se and are called **metabolic intermediates**.

**4. Molecules of miscellaneous function.** This is obviously a broad category of molecules but not as large as you might expect; the vast bulk of the dry weight of a cell is made up of macromolecules and their direct precursors. The molecules of miscellaneous function include such substances as vitamins, which function primarily as adjuncts to proteins; certain steroid or amino acid hormones; molecules involved in energy storage, such as ATP; regulatory molecules such as cyclic AMP; and metabolic waste products such as urea.

### REVIEW

- What properties of a carbon atom are critical to life?
- Draw the structures of four different functional groups. How would each of these groups alter the solubility of a molecule in water?

## 42 2.6 Carbohydrates

**Carbohydrates** (or **glycans**, as they are often called) include simple sugars (or *monosaccharides*) and all larger molecules constructed of sugar building blocks. Carbohydrates function primarily as stores of chemical energy and as durable building materials for biological construction. Most sugars have the general formula  $(\text{CH}_2\text{O})_n$ . The sugars of importance in cellular metabolism have values of  $n$  that range from 3 to 7. Sugars containing three carbons are known as *triose*s, those with four carbons as *tetrose*s, those with five carbons as *pentose*s, those with six carbons as *hexose*s, and those with seven carbons as *heptose*s.

### The Structure of Simple Sugars

Each sugar molecule consists of a backbone of carbon atoms linked together in a linear array by single bonds. Each of the carbon atoms of the backbone is linked to a single hydroxyl group, except for one that bears a *carbonyl* ( $\text{C}=\text{O}$ ) group. If the carbonyl group is located at an internal position (to form a ketone group), the sugar is a *ketose*, such as fructose, which is shown in **FIGURE 2.12a**. If the carbonyl is located at one end of the sugar, it forms an aldehyde group and the molecule is known as an *aldose*, as exemplified by glucose, which is shown in **Figure 2.12b-f**. Because of their large numbers of hydroxyl groups, sugars tend to be highly water soluble.

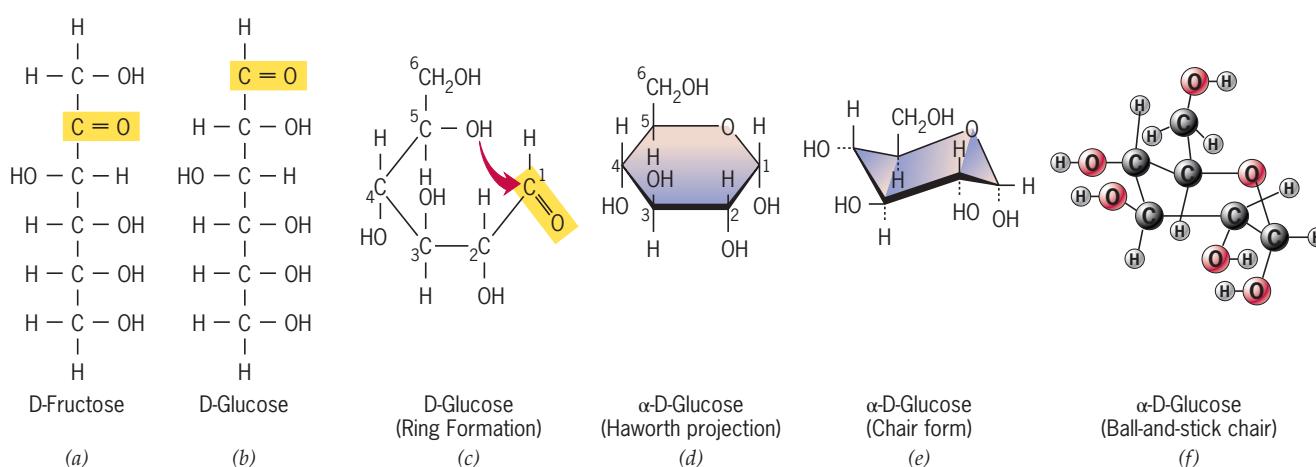
Although the straight-chained formulas shown in **Figure 2.12a,b** are useful for comparing the structures of various sugars, sugars having five or more carbons spontaneously self-react (**Figure 2.12c**) to produce a closed, or ring-containing, molecule. Only a tiny fraction of sugar molecules in solution are found in the open-chain linear form; the rest are in the ring form. The linear form is biochemically important because the aldehyde group at the end of the chain is reactive and can react with proteins, notably hemoglobin. Patients with diabetes have higher levels of sugar in their blood, and this sugar in its open chain form reacts with hemoglobin to produce a modified hemoglobin called Hemoglobin A1c, which is often used in blood tests to track the progress of diabetes. Similar reactions of linear-form sugar with proteins involved in cholesterol metabolism are one of the reasons that diabetes

causes heart disease. The open chain form of sugars is thus quite important in medicine, but the vast majority of sugars are found in the ring form, and it is in this form that they are used as building blocks to build other types of carbohydrates. The ring forms of sugars are usually depicted as flat (*planar*) structures (**Figure 2.12d**) lying perpendicular to the plane of the paper with the thickened line situated closest to the reader. The H and OH groups lie parallel to the plane of the paper, projecting either above or below the ring of the sugar. In actual fact, the sugar ring is not a planar structure, but usually exists in a three-dimensional conformation resembling a chair (**Figure 2.12e,f**).

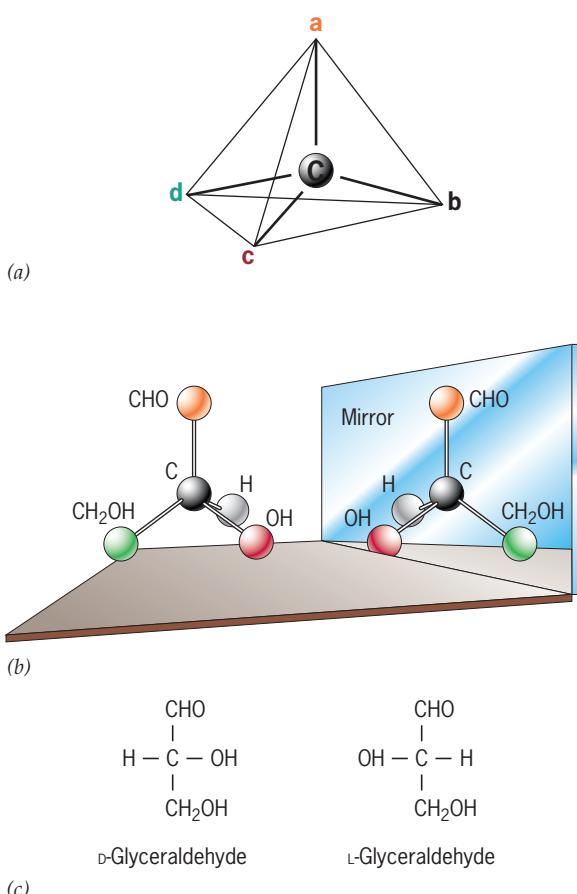
### Stereoisomerism

As noted earlier, a carbon atom can bond with four other atoms. The arrangement of the groups around a carbon atom can be depicted as in **FIGURE 2.13a** with the carbon placed in the center of a tetrahedron and the bonded groups projecting into its four corners. **Figure 2.13b** depicts a molecule of glyceraldehyde, which is the only aldotriose. The second carbon atom of glyceraldehyde is linked to four different groups ( $-\text{H}$ ,  $-\text{OH}$ ,  $-\text{CHO}$ , and  $-\text{CH}_2\text{OH}$ ). If the four groups bonded to a carbon atom are all different, as in glyceraldehyde, then two possible configurations exist that cannot be superimposed on one another. These two molecules (termed *stereoisomers* or *enantiomers*) have essentially the same chemical reactivities, but their structures are mirror images (not unlike a pair of right and left human hands). By convention, the molecule is called *D*-glyceraldehyde if the hydroxyl group of carbon 2 projects to the right, and *L*-glyceraldehyde if it projects to the left (**Figure 2.13c**). Because it acts as a site of stereoisomerism, carbon 2 is referred to as an *asymmetric* carbon atom.

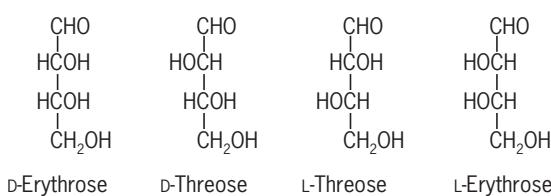
As the backbone of sugar molecules increases in length, so too does the number of asymmetric carbon atoms and, consequently, the number of stereoisomers. Aldotetroses have two asymmetric carbons and thus can exist in four different configurations (**FIGURE 2.14**). Similarly, there are eight different aldopentoses and 16 different aldochexoses. The designation of each of these sugars as *D* or *L* is based by convention on the arrangement of groups attached to the asymmetric carbon atom farthest from the aldehyde (the carbon associated with the aldehyde is



**FIGURE 2.12** The structures of sugars. (a) Straight-chain formula of fructose, a ketohexose [keto, indicating the carbonyl (yellow)], is located internally, and hexose because it consists of six carbons]. (b) Straight-chain formula of glucose, an aldohexose (aldo because the carbonyl is located at the end of the molecule). (c) Self-reaction in which glucose is converted from an open chain to a closed ring (a pyranose ring). (d) Glucose is commonly depicted in the form of a flat (planar) ring lying perpendicular to the page with the thickened line situated closest to the reader and the H and OH groups projecting either above or below the ring. The basis for the designation  $\alpha$ -D-glucose is discussed in the following section. (e) The chair conformation of glucose, which depicts its three-dimensional structure more accurately than the flattened ring of part d. (f) A ball-and-stick model of the chair conformation of glucose.



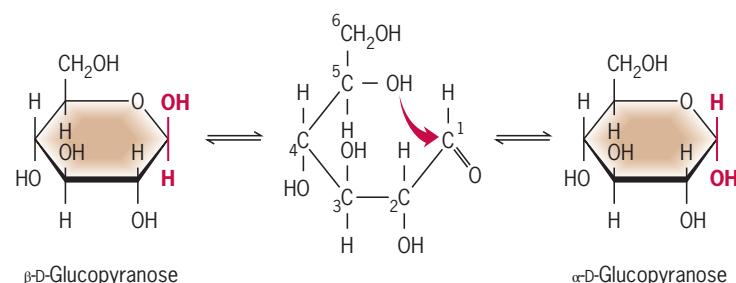
**FIGURE 2.13** Stereoisomerism of glyceraldehyde. (a) The four groups bonded to a carbon atom (labeled a, b, c, and d) occupy the four corners of a tetrahedron with the carbon atom at its center. (b) Glyceraldehyde is the only three-carbon aldose; its second carbon atom is bonded to four different groups ( $\text{—H}$ ,  $\text{—OH}$ ,  $\text{—CHO}$ , and  $\text{—CH}_2\text{OH}$ ). As a result, glyceraldehyde can exist in two possible configurations that are not superimposable, but instead are mirror images of each other as indicated. These two stereoisomers (or enantiomers) can be distinguished by the configuration of the four groups around the asymmetric (or chiral) carbon atom. Solutions of these two isomers rotate plane-polarized light in opposite directions and, thus, are said to be optically active. (c) Straight-chain formulas of glyceraldehyde. By convention, the D-isomer is shown with the OH group on the right.



**FIGURE 2.14 Aldotetroses.** Because they have two asymmetric carbon atoms, aldotetroses can exist in four configurations.

designated C1). If the hydroxyl group of this carbon projects to the right, the aldose is a D-sugar; if it projects to the left, it is an L-sugar. The enzymes present in living cells can distinguish between the D and L forms of a sugar. Typically, only one of the stereoisomers (such as D-glucose and L-fucose) is used by cells.

The self-reaction in which a straight-chain glucose molecule is converted into a six-membered (pyranose) ring was shown in

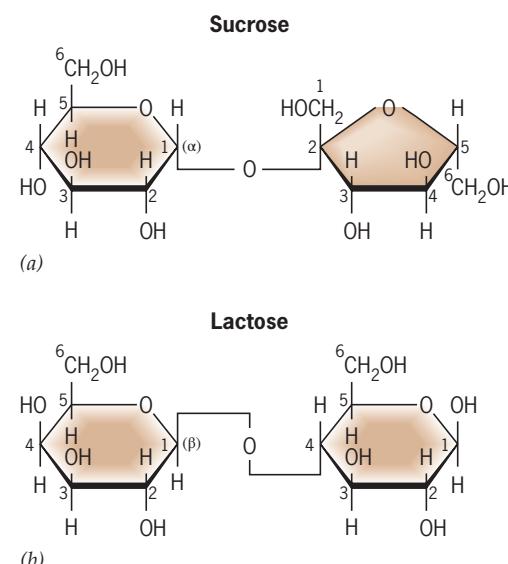


**FIGURE 2.15 Formation of an  $\alpha$ - and  $\beta$ -pyranose.** When a molecule of glucose undergoes self-reaction to form a pyranose ring (i.e., a six-membered ring), two stereoisomers are generated. The two isomers are in equilibrium with each other through the open-chain form of the molecule. By convention, the molecule is an  $\alpha$ -pyranose when the OH group of the first carbon projects below the plane of the ring, and a  $\beta$ -pyranose when the hydroxyl group projects upward.

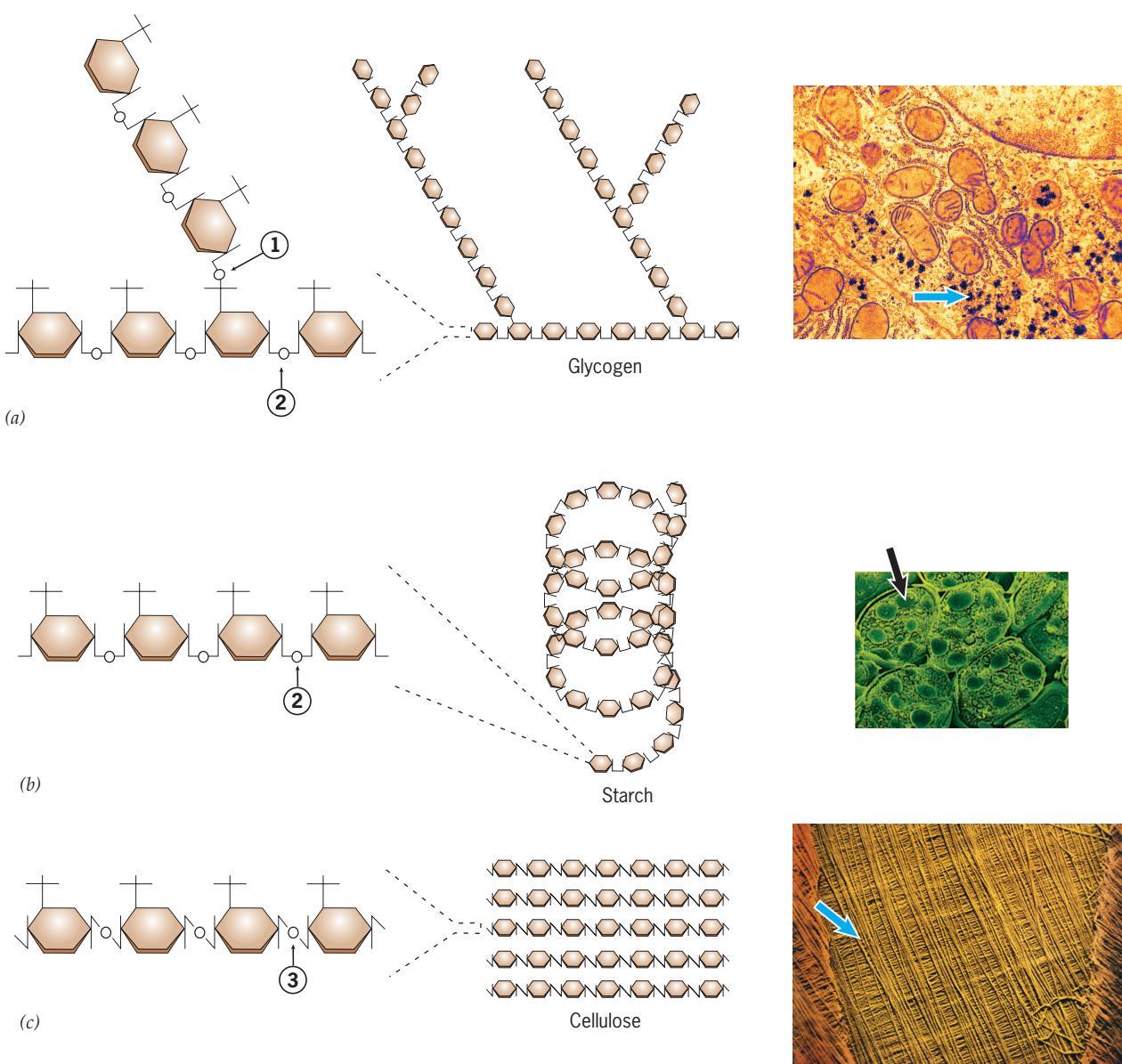
Figure 2.12c. Unlike its precursor in the open chain, the C1 of the ring bears four different groups and thus becomes a new center of asymmetry within the sugar molecule. Because of this extra asymmetric carbon atom, each type of pyranose exists as  $\alpha$  and  $\beta$  stereoisomers (**FIGURE 2.15**). By convention, the molecule is an  $\alpha$ -pyranose when the OH group of the first carbon projects below the plane of the ring, and a  $\beta$ -pyranose when the hydroxyl projects upward. The difference between the two forms has important biological consequences, resulting, for example, in the compact shape of glycogen and starch molecules and the extended conformation of cellulose (discussed later).

## Linking Sugars Together

Sugars can be joined to one another by covalent **glycosidic bonds** to form larger molecules. Glycosidic bonds form by reaction between carbon atom C1 of one sugar and the hydroxyl group of another sugar, generating a  $\text{---C---O---C---}$  linkage between the two sugars. As discussed below (and indicated in **FIGURES 2.16** and **2.17**), sugars can be joined by quite a variety of different glycosidic bonds. Molecules



**FIGURE 2.16 Disaccharides.** Sucrose and lactose are two of the most common disaccharides. Sucrose is composed of glucose and fructose joined by an  $\alpha(1 \rightarrow 2)$  linkage, whereas lactose is composed of glucose and galactose joined by a  $\beta(1 \rightarrow 4)$  linkage.



**FIGURE 2.17** Three polysaccharides with identical sugar monomers but dramatically different properties. Glycogen (a), starch (b), and cellulose (c) are each composed entirely of glucose subunits, yet their chemical and physical properties are very different due to the distinct ways that the monomers are linked together (three different types of linkages are indicated by the circled numbers). Glycogen molecules are the most highly branched, starch molecules assume a helical arrangement, and cellulose molecules are unbranched and highly extended. Whereas glycogen and starch are energy stores, cellulose molecules are bundled together into tough fibers that are suited for their structural role. Colorized electron micrographs show glycogen granules in a liver cell, starch grains (amyloplasts) in a plant seed, and cellulose fibers in a plant cell wall; each is indicated by an arrow.

SOURCE: (a) Don W. Fawcett/Photo Researchers, Inc.; (b) Jeremy Burgess/Photo Researchers, Inc.; (c) Biophoto Associates/Photo Researchers, Inc.

composed of only two sugar units are *disaccharides* (Figure 2.16). Disaccharides serve primarily as readily available energy stores. Sucrose, or table sugar, is a major component of plant sap, which carries chemical energy from one part of the plant to another. Lactose, present in the milk of most mammals, supplies newborn mammals with fuel for early growth and development. Lactose in the diet is hydrolyzed by the enzyme lactase, which is present in the plasma membranes of the cells that line the intestine. Many people lose this enzyme after childhood and find that eating dairy products causes digestive discomfort.

Sugars may also be linked together to form small chains called **oligosaccharides** (*oligo* = few). Most often such chains are found covalently attached to lipids and proteins, converting them into glycolipids and glycoproteins, respectively. Oligosaccharides are

particularly important on the glycolipids and glycoproteins of the plasma membrane, where they project from the cell surface (see Figure 4.4a). Because oligosaccharides may be composed of many different combinations of sugar units, these carbohydrates can play an informational role; that is, they can serve to distinguish one type of cell from another and help mediate specific interactions of a cell with its surroundings.

## Polysaccharides

By the middle of the nineteenth century, it was known that the blood of people suffering from diabetes had a sweet taste due to an elevated level of glucose, the key sugar in energy metabolism. Claude Bernard, a prominent French physiologist of the period, was looking for the cause of diabetes by investigating the source of blood sugar. It was

assumed at the time that any sugar present in a human or an animal had to have been previously consumed in the diet. Working with dogs, Bernard found that, even if the animals were placed on a diet totally lacking carbohydrates, their blood still contained a normal amount of glucose. Clearly, glucose could be formed in the body from other types of compounds.

After further investigation, Bernard found that glucose enters the blood from the liver. Liver tissue, he found, contains an insoluble polymer of glucose he named **glycogen**. Bernard concluded that various food materials (such as proteins) were carried to the liver where they were chemically converted to glucose and stored as glycogen. Then, as the body needed sugar for fuel, the glycogen in the liver was transformed to glucose, which was released into the bloodstream to satisfy glucose-depleted tissues. In Bernard's hypothesis, the balance between glycogen formation and glycogen breakdown in the liver was the prime determinant in maintaining the relatively constant (*homeostatic*) level of glucose in the blood.

Bernard's hypothesis proved to be correct. The molecule he named glycogen is a type of **polysaccharide**—a polymer of sugar units joined by glycosidic bonds.

**GLYCOGEN AND STARCH: NUTRITIONAL POLYSACCHARIDES** Glycogen is a branched polymer containing only one type of monomer: glucose (Figure 2.17a). Most of the sugar units of a glycogen molecule are joined to one another by  $\alpha(1 \rightarrow 4)$  glycosidic bonds (type 2 bond in Figure 2.17a). Branch points contain a sugar joined to three neighboring units rather than to two, as in the unbranched segments of the polymer. The extra neighbor, which forms the branch, is linked by an  $\alpha(1 \rightarrow 6)$  glycosidic bond (type 1 bond in Figure 2.17a).

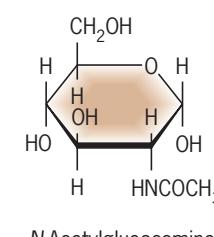
Glycogen serves as a storehouse of surplus chemical energy in most animals. Human skeletal muscles, for example, typically contain enough glycogen to fuel about 30 minutes of moderate activity. Depending on various factors, glycogen typically ranges in molecular weight from about one to four million daltons. When stored in cells, glycogen is highly concentrated in what appears as dark-staining, irregular granules in electron micrographs (Figure 2.17a, right).

Most plants bank their surplus chemical energy in the form of **starch**, which like glycogen is also a polymer of glucose. Potatoes and cereals, for example, consist primarily of starch. Starch is actually a mixture of two different polymers, amylose and amylopectin. Amylose is an unbranched, helical molecule whose sugars are joined by  $\alpha(1 \rightarrow 4)$  linkages (Figure 2.17b), whereas amylopectin is branched. Amylopectin differs from glycogen in being much less branched and having an irregular branching pattern. Starch is stored as densely packed granules, or *starch grains*, which are enclosed in membrane-bound organelles (*plastids*) within the plant cell (Figure 2.17b, right). Although animals don't synthesize starch, they possess an enzyme (*amylase*) that readily hydrolyzes it.

**CELLULOSE, CHITIN, AND GLYCOSAMINOGLYCANS: STRUCTURAL POLYSACCHARIDES** Whereas some polysaccharides constitute easily digested energy stores, others form tough, durable structural materials. Cotton and linen, for example, consist largely of **cellulose**, which is the major component of plant cell walls. Cotton textiles owe their durability to the long, unbranched cellulose molecules, which are ordered into side-by-side aggregates to form molecular cables (Figure 2.17c, right panel) that are ideally constructed to resist pulling (tensile) forces. Like glycogen and starch,

cellulose consists solely of glucose monomers; its properties differ dramatically from these other polysaccharides because the glucose units are joined by  $\beta(1 \rightarrow 4)$  linkages (bond 3 in Figure 2.17c) rather than  $\alpha(1 \rightarrow 4)$  linkages. Ironically, multicellular animals (with rare exception) lack the enzyme needed to degrade cellulose, which happens to be the most abundant organic material on Earth and rich in chemical energy. Animals that "make a living" by digesting cellulose, such as termites and sheep, do so by harboring bacteria and protozoa that synthesize the necessary enzyme, cellulase. Cellulose is an important component of *dietary fiber*, a broad term that includes all the polysaccharides we eat that cannot be digested by human enzymes.

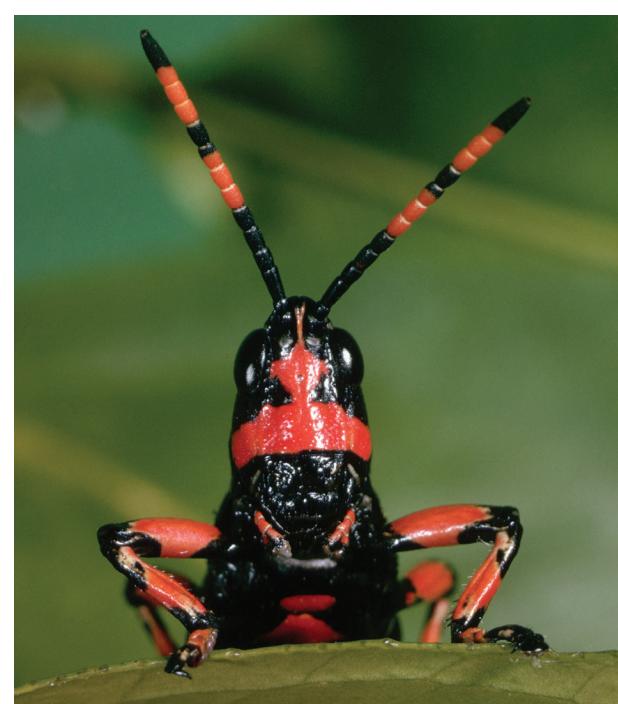
Not all biological polysaccharides consist of glucose monomers. **Chitin** is an unbranched polymer of the sugar *N*-acetylglucosamine, which is similar in structure to glucose but has an acetyl amino group instead of a hydroxyl group bonded to the second carbon atom of the ring.



*N*-Acetylglucosamine

Chitin occurs widely as a structural material among invertebrates, particularly in the outer covering of insects, spiders, and crustaceans. Chitin is a tough, resilient, yet flexible material not unlike certain plastics. Insects owe much of their success to this highly adaptive polysaccharide (FIGURE 2.18).

Another group of polysaccharides that has a more complex structure is the **glycosaminoglycans** (or **GAGs**). Unlike other polysaccharides, they have the structure —A—B—A—B—, where A and B represent two different sugars. The best-studied GAG is



**FIGURE 2.18** Chitin is the primary component of the outer skeleton of this grasshopper.

SOURCE: Anthony Bannister/Gallo Images/© Corbis.

- 46** heparin, which is secreted by cells in the lungs and other tissues in response to tissue injury. Heparin inhibits blood coagulation, thereby preventing the formation of clots that can block the flow of blood to the heart or lungs. Heparin accomplishes this feat by activating an inhibitor (antithrombin) of a key enzyme (thrombin) that is required for blood coagulation. Heparin, which is normally extracted from pig tissue, has been used for decades to prevent blood clots in patients following major surgery. Unlike heparin, most GAGs are found in the spaces that surround cells, and their structure and function are discussed in Section 7.3. The most complex polysaccharides are found in plant cell walls (Section 7.14).

### REVIEW

1. Name three polysaccharides composed of polymers of glucose. How do these macromolecules differ from one another?

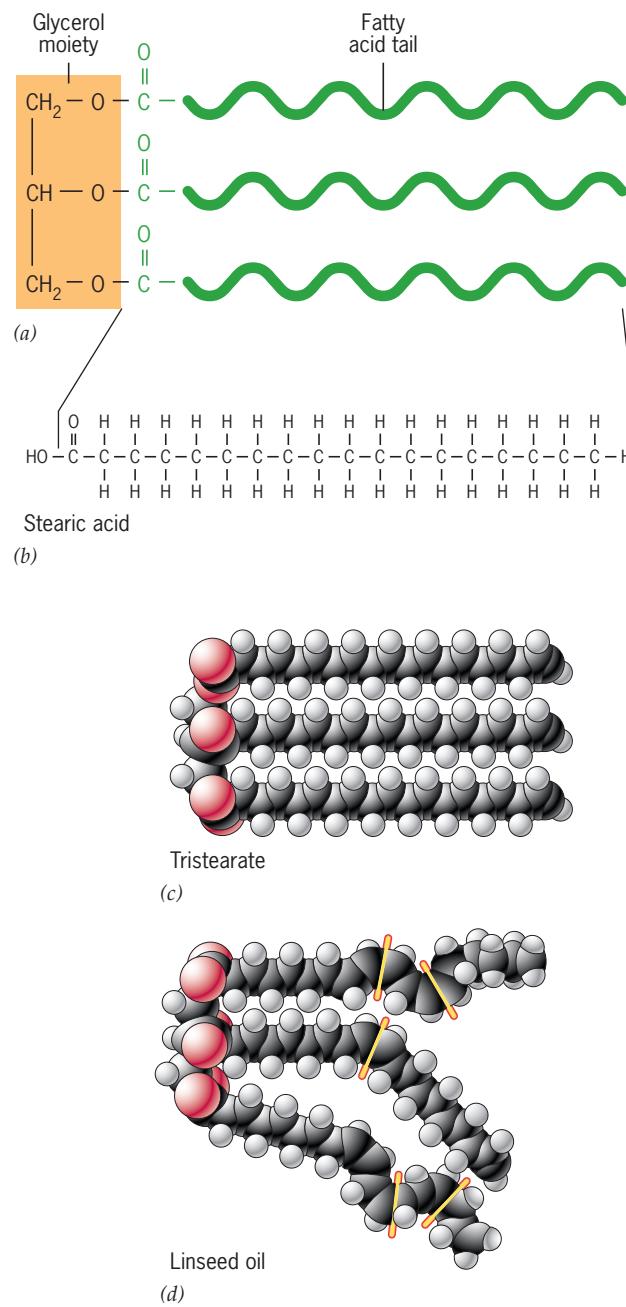
## 2.7 Lipids

Lipids are a diverse group of nonpolar biological molecules whose common properties are their ability to dissolve in organic solvents, such as chloroform or benzene, and their inability to dissolve in water—a property that explains many of their varied biological functions. Lipids of importance in cellular function include fats, steroids, and phospholipids.

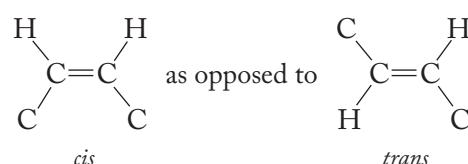
### Fats

**Fats** consist of a glycerol molecule linked by ester bonds to three fatty acids; the composite molecule is termed a **triacylglycerol** (FIGURE 2.19A), also known as a triglyceride. We will begin by considering the structure of **fatty acids**. Fatty acids are long, unbranched hydrocarbon chains with a single carboxyl group at one end (Figure 2.19b). Because the two ends of a fatty acid molecule have a very different structure, they also have different properties. The hydrocarbon chain is hydrophobic, whereas the carboxyl group ( $-\text{COOH}$ ), which bears a negative charge at physiological pH, is hydrophilic. Molecules having both hydrophobic and hydrophilic regions are said to be **amphipathic**; such molecules have unusual and biologically important properties. The properties of fatty acids can be appreciated by considering the use of a familiar product: soap, which consists of fatty acids. In past centuries, soaps were made by heating animal fat in strong alkali ( $\text{NaOH}$  or  $\text{KOH}$ ) to break the bonds between the fatty acids and the glycerol. Today, most soaps are made synthetically. Soaps owe their grease-dissolving capability to the fact that the hydrophobic end of each fatty acid can embed itself in the grease, whereas the hydrophilic end can interact with the surrounding water. As a result, greasy materials are converted into complexes (*micelles*) that can be dispersed by water (FIGURE 2.20).

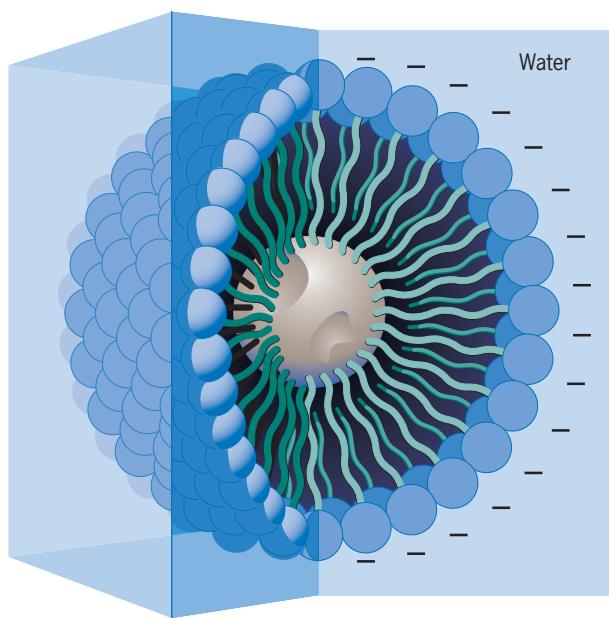
Fatty acids differ from one another in the length of their hydrocarbon chain and the presence or absence of double bonds. Fatty acids present in cells typically vary in length from 14 to 20 carbons. Fatty acids that lack double bonds, such as stearic acid (Figure 2.19b), are described as **saturated**; those possessing double bonds are **unsaturated**. Naturally occurring fatty acids have double bonds in the *cis* configuration. Double bonds (of the *cis* configuration)



**FIGURE 2.19** Fats and fatty acids. (a) The basic structure of a triacylglycerol (also called a triglyceride or a neutral fat). The glycerol moiety, indicated in orange, is linked by three ester bonds to the carboxyl groups of three fatty acids whose tails are indicated in green. (b) Stearic acid, an 18-carbon saturated fatty acid that is common in animal fats. (c) Space-filling model of tristearate, a triacylglycerol containing three identical stearic acid chains. (d) Space-filling model of linseed oil, a triacylglycerol derived from flax seeds that contains three unsaturated fatty acids (linoleic, oleic, and linolenic acids). The sites of unsaturation, which produce kinks in the molecule, are indicated by the yellow-orange bars.



produce kinks in a fatty acid chain. Consequently, the more double bonds that fatty acid chains possess, the less effectively these long



**FIGURE 2.20 Soaps consist of fatty acids.** In this schematic drawing of a soap micelle, the nonpolar tails of the fatty acids are directed inward, where they interact with the greasy matter to be dissolved. The negatively charged heads are located at the surface of the micelle, where they interact with the surrounding water. Membrane proteins, which also tend to be insoluble in water, can also be solubilized in this way by extraction of membranes with detergents.

chains can be packed together. This lowers the temperature at which a fatty acid-containing lipid melts. Tristearate, whose fatty acids lack double bonds (Figure 2.19c), is a common component of animal fats and remains in a solid state well above room temperature. In contrast, the profusion of double bonds in vegetable fats accounts for their liquid state—both in the plant cell and on the grocery shelf—and for their being labeled as “polyunsaturated.” Fats that are liquid at room temperature are described as **oils**. Figure 2.19d shows the structure of linseed oil, a highly volatile lipid extracted from flax seeds, that remains a liquid at a much lower temperature than does tristearate. Solid shortenings, such as margarine, are formed from unsaturated vegetable oils by chemically reducing the double bonds with hydrogen atoms (a process termed *hydrogenation*). The hydrogenation process also converts some of the *cis* double bonds into *trans* double bonds, which are straight rather than kinked. This process generates partially hydrogenated or trans-fats. Eating trans fat increases the risk of heart disease, and artificial trans fat is now banned in several countries, with others considering similar measures.

A molecule of fat can contain three identical fatty acids (as in Figure 2.19c), or it can be a *mixed fat*, containing more than one fatty acid species (as in Figure 2.19d). Most natural fats, such as olive oil or butterfat, are mixtures of molecules having different fatty acid species.

Fats are very rich in chemical energy; a gram of fat contains over twice the energy content of a gram of carbohydrate (for reasons discussed in Section 3.9). Carbohydrates function primarily as a short-term, rapidly available energy source, whereas fat reserves store energy on a long-term basis. It is estimated that a person of average size contains about 0.5 kilograms (kg) of carbohydrate, primarily in the form of glycogen. This amount of carbohydrate provides approximately 2000 kcal of total energy. During the course of a strenuous day's exercise, a person can virtually deplete his or her body's entire store of carbohydrate. In contrast, the average person contains

approximately 16 kg of fat (equivalent to 144,000 kcal of energy), and as we all know, it can take a very long time to deplete our store of this material.

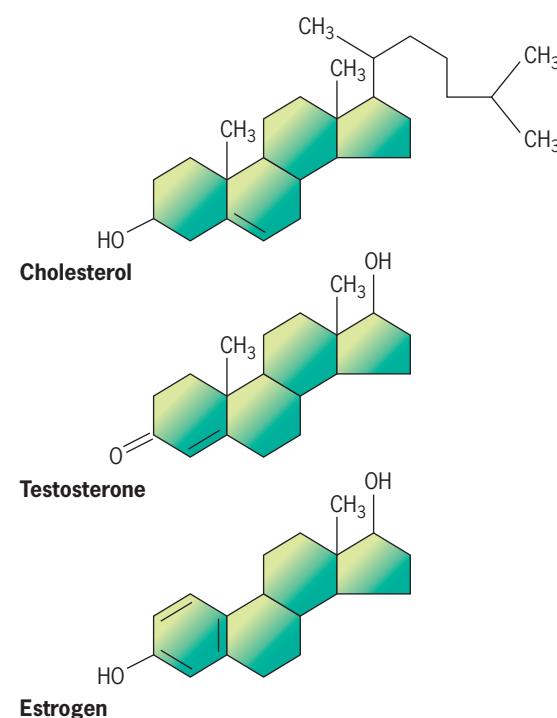
Because they lack polar groups, fats are extremely insoluble in water and are stored in cells in the form of dry lipid droplets. Because lipid droplets do not contain water as do glycogen granules, they represent an extremely concentrated storage fuel. In many animals, fats are stored in special cells (*adipocytes*) whose cytoplasm is filled with one or a few large lipid droplets. Adipocytes exhibit a remarkable ability to change their volume to accommodate varying quantities of fat.

## Steroids

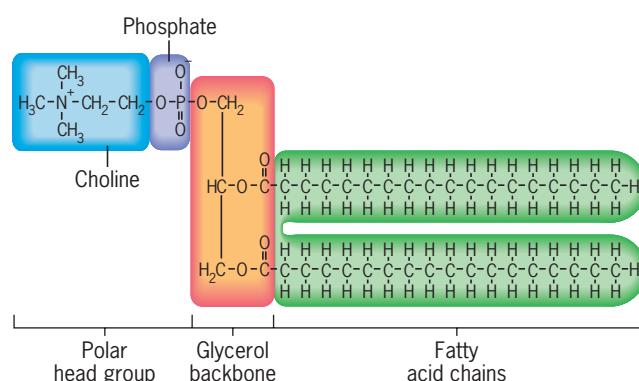
Steroids are built around a characteristic four-ringed hydrocarbon skeleton. One of the most important steroids is *cholesterol*, a component of animal cell membranes and a precursor for the synthesis of a number of steroid hormones, such as testosterone, progesterone, and estrogen (FIGURE 2.21). Cholesterol is largely absent from plant cells, which is why vegetable oils are considered “cholesterol-free,” but plant cells may contain large quantities of related compounds.

## Phospholipids

The chemical structure of a common phospholipid is shown in FIGURE 2.22. The molecule resembles a fat (triacylglycerol), but has only two fatty acid chains rather than three; it is a *diacylglycerol*. The third hydroxyl of the glycerol backbone is covalently bonded to a phosphate group, which in turn is covalently bonded to a small polar group, such as choline, as shown in Figure 2.22. Thus, unlike fat molecules, phospholipids contain two ends that have very different



**FIGURE 2.21 The structure of steroids.** All steroids share the basic four-ring skeleton. The seemingly minor differences in chemical structure between cholesterol, testosterone, and estrogen generate profound biological differences.



**FIGURE 2.22** The phospholipid phosphatidylcholine. The molecule consists of a glycerol backbone whose hydroxyl groups are covalently bonded to two fatty acids and a phosphate group. The negatively charged phosphate is also bonded to a small, positively charged choline group. The end of the molecule that contains the phosphorylcholine is hydrophilic, whereas the opposite end, consisting of the fatty acid tail, is hydrophobic. The structure and function of phosphatidylcholine and other phospholipids are discussed at length in Section 4.2.

properties: the end containing the phosphate group has a distinctly hydrophilic character; the other end composed of the two fatty acid tails has a distinctly hydrophobic character. Because phospholipids function primarily in cell membranes, and because the properties of cell membranes depend on their phospholipid components, they will be discussed further in Sections 4.2 and 15.6 in connection with cell membranes.

### REVIEW

1. Describe the properties of three different types of lipid molecules. What are their respective biological roles?

## 2.8 Building Blocks of Proteins

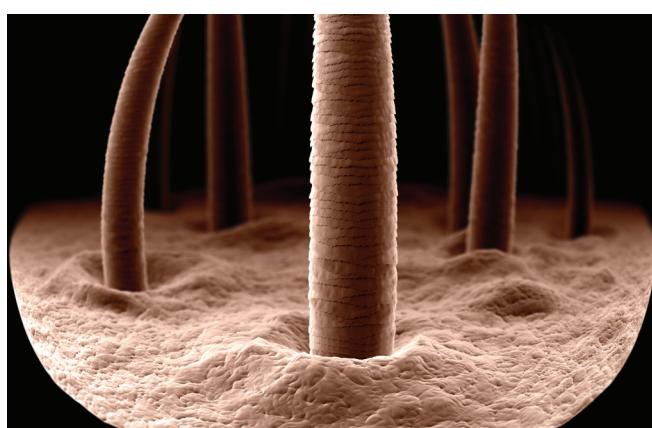
Proteins are the macromolecules that carry out virtually all of a cell's activities; they are the molecular tools and machines that make things happen. As enzymes, proteins vastly accelerate the rate of



(a)



(b)



(c)



(d)

**FIGURE 2.23** Four examples of the thousands of biological structures composed predominantly of protein. These include (a) feathers, which are adaptations in birds for thermal insulation, flight, and sex recognition; and (b) spider webs, made of a protein-based silk that is among the strongest materials known. (c) human hair, composed of the protein keratin, and (d) fingernails, which are also composed of keratin.

SOURCE: (a) Darrell Gulin/Getty Images, Inc.; (b) © gabriela schaufelberger / iStockphoto; (c) © Science Picture Co; (d) Tamara83 / Shutterstock.

metabolic reactions; as structural cables, proteins provide mechanical support both within cells and outside their perimeters (**FIGURE 2.23a**); as hormones, growth factors, and gene activators, proteins perform a wide variety of regulatory functions; as membrane receptors and transporters, proteins determine what a cell reacts to and what types of substances enter or leave the cell; as contractile filaments and molecular motors, proteins constitute the machinery for biological movements. Among their many other functions, proteins act as antibodies, serve as toxins, form blood clots, absorb or refract light (Figure 2.23b), and transport substances from one part of the body to another.

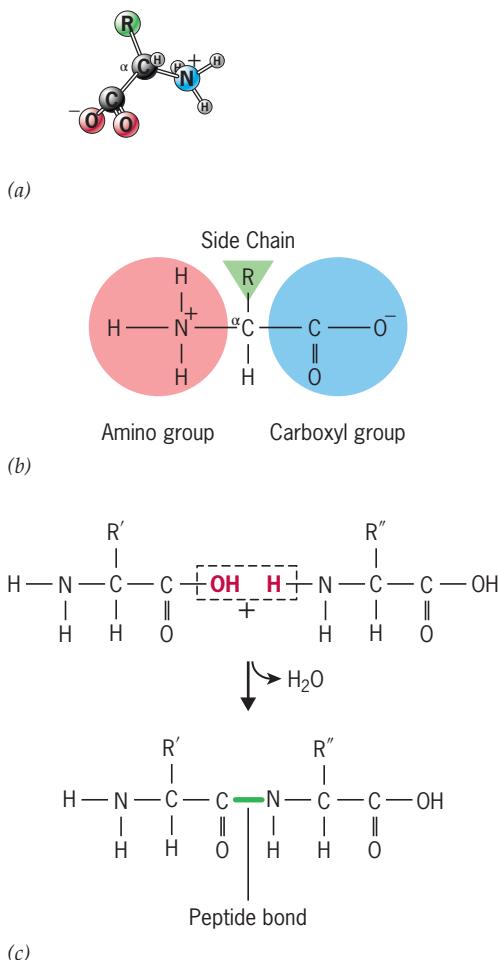
How can one type of molecule have so many varied functions? The explanation resides in the virtually unlimited molecular structures that proteins, *as a group*, can assume. Each protein, however, has a unique and defined structure that enables it to carry out a particular function. Most importantly, proteins have shapes and surfaces that allow them to interact *selectively* with other molecules. Proteins, in other words, exhibit a high degree of **specificity**. It is possible, for example, for a particular DNA-cutting enzyme to recognize a segment of DNA containing one specific sequence of eight nucleotides, while ignoring all the other 65,535 possible sequences composed of this number of nucleotides.

## The Structures of Amino Acids

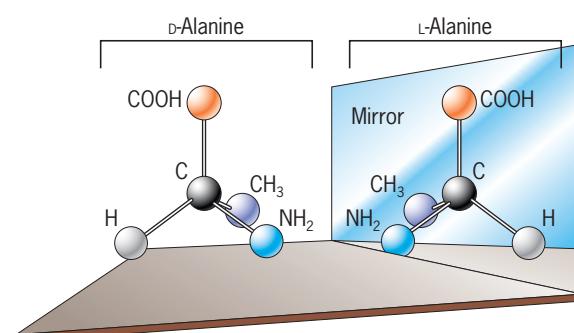
Proteins are polymers made of amino acid monomers. Each protein has a unique sequence of amino acids that gives the molecule its unique properties. Many of the capabilities of a protein can be understood by examining the chemical properties of its constituent amino acids. Twenty different amino acids are commonly used in the construction of proteins, whether from a virus or a human. There are two aspects of amino acid structure to consider: that which is common to all of them and that which is unique to each. We will begin with the shared properties.

All amino acids have a carboxyl group and an amino group, which are separated from each other by a single carbon atom, the  $\alpha$ -carbon (**FIGURE 2.24a,b**). In a neutral aqueous solution, the  $\alpha$ -carboxyl group loses its proton and exists in a negatively charged state ( $-\text{COO}^-$ ), and the  $\alpha$ -amino group accepts a proton and exists in a positively charged state ( $\text{NH}_3^+$ ) (Figure 2.24b). We saw on page 43 that carbon atoms bonded to four different groups can exist in two configurations (*stereoisomers*) that cannot be superimposed on one another. Amino acids also have asymmetric carbon atoms. With the exception of glycine, the  $\alpha$ -carbon of amino acids bonds to four different groups so that each amino acid can exist in either a **D** or an **L** form (**FIGURE 2.25**). Amino acids used in the synthesis of a protein on a ribosome are always **L**-amino acids. The “selection” of **L**-amino acids must have occurred very early in cellular evolution and has been conserved for billions of years. Microorganisms, however, use **D**-amino acids in the synthesis of certain small peptides, including those of the cell wall and several antibiotics (e.g., gramicidin A).

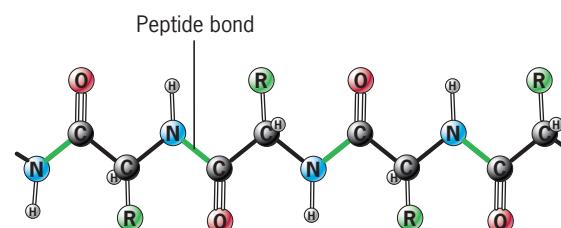
During the process of protein synthesis, each amino acid becomes joined to two other amino acids, forming a long, continuous, unbranched polymer called a **polypeptide chain**. The amino acids that make up a polypeptide chain are joined by **peptide bonds** that result from the linkage of the carboxyl group of one amino acid to the amino group of its neighbor, with the elimination of a molecule of water (Figure 2.24c). A polypeptide chain composed of a string of amino acids joined by peptide bonds has the following backbone:



**FIGURE 2.24** Amino acid structure. Ball-and-stick model (a) and chemical formula (b) of a generalized amino acid in which R can be any of a number of chemical groups (see Figure 2.26). (c) The formation of a peptide bond occurs by the condensation of two amino acids, drawn here in the uncharged state. In the cell, this reaction occurs on a ribosome as an amino acid is transferred from a carrier (a tRNA molecule) onto the end of the growing polypeptide chain (see Figure 11.49).



**FIGURE 2.25** Amino acid stereoisomerism. Because the  $\alpha$ -carbon of all amino acids except glycine is bonded to four different groups, two stereoisomers can exist. The **D** and **L** forms of alanine are shown.



- 50** The “average” polypeptide chain contains about 450 amino acids. The longest known polypeptide, found in the muscle protein titin, contains more than 30,000 amino acids. Once incorporated into a polypeptide chain, amino acids are termed *residues*. The residue on one end of the chain, the *N-terminus*, contains an amino acid with a free (unbonded)  $\alpha$ -amino group, whereas the residue at the opposite end, the *C-terminus*, has a free  $\alpha$ -carboxyl group. In addition to amino acids, many proteins contain other types of components that are added after the polypeptide is synthesized. These include carbohydrates (to form glycoproteins), metal-containing groups (to form metalloproteins), and organic groups (e.g., flavoproteins).

## The Properties of the Side Chains

The backbone, or main chain, of the polypeptide is composed of that part of each amino acid that is common to all of them. The **side chain** or **R group** (Figure 2.24), bonded to the  $\alpha$ -carbon, is highly variable among the 20 building blocks, and it is this variability that ultimately gives proteins their diverse structures and activities. If the various amino acid side chains are considered together, they exhibit a large variety of structural features, ranging from fully charged to hydrophobic, and they can participate in a wide variety of covalent and noncovalent bonds. As discussed in the following chapter, the side chains of the “active sites” of enzymes can facilitate (catalyze) many different organic reactions. The assorted characteristics of the side chains of the amino acids are important in both *intramolecular* interactions, which determine the structure and activity of the molecule, and *intermolecular* interactions, which determine the relationship of a polypeptide with other molecules, including other polypeptides (page 60).

Amino acids are classified on the character of their side chains. They fall roughly into four categories: polar and charged, polar and uncharged, nonpolar, and those with unique properties (FIGURE 2.26).

**1. Polar, charged** (Fig. 2.26a). Amino acids of this group include aspartic acid, glutamic acid, lysine, and arginine. These four amino acids contain side chains that become fully charged; that is, the side chains contain relatively strong organic acids and bases. The ionization reactions of glutamic acid and lysine are shown in FIGURE 2.27. At physiologic pH, the side chains of these amino acids are almost always present in the fully charged state. Consequently, they are able to form ionic bonds with other charged species in the cell. For example, the positively charged arginine residues of histone proteins are linked by ionic bonds to the negatively charged phosphate groups of DNA (see Figure 2.3). Histidine is also considered a polar, charged amino acid, although in most cases it is only partially charged at physiologic pH. In fact, because of its ability to gain or lose a proton in physiologic pH ranges, histidine is a particularly important residue in the active site of many proteins (as in Figure 3.13).

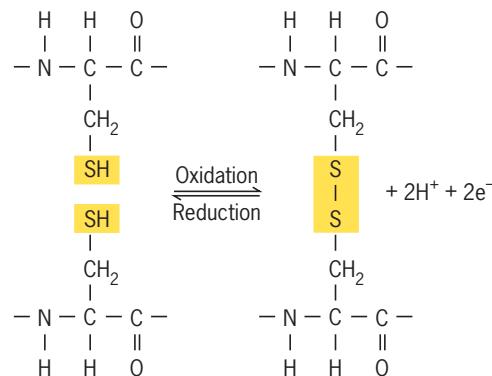
**2. Polar, uncharged** (Fig. 2.26b). The side chains of these amino acids have a partial negative or positive charge and thus can form hydrogen bonds with other molecules including water. These amino acids are often quite reactive. Included in this category are asparagine and glutamine (the amides of aspartic acid and glutamic acid), threonine, serine, and tyrosine.

**3. Nonpolar** (Fig. 2.26c). The side chains of these amino acids are hydrophobic and are unable to form electrostatic bonds or interact with water. The amino acids of this category are alanine, valine, leucine, isoleucine, tryptophan, phenylalanine, and methionine. The side chains of the nonpolar amino acids generally lack oxygen

and nitrogen. They vary primarily in size and shape, which allows one or another of them to pack tightly into a particular space within the core of a protein, associating with one another as the result of van der Waals forces and hydrophobic interactions.

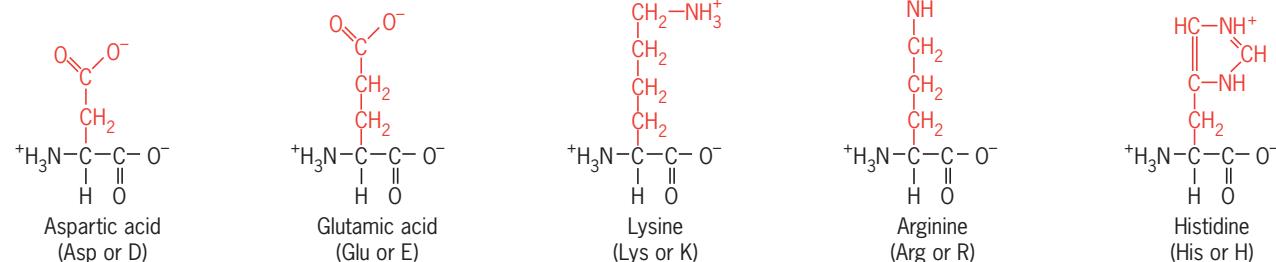
**4. The other three amino acids**—glycine, proline, and cysteine (Fig. 2.26d)—have unique properties that separate them from the others. The side chain of glycine consists of only a hydrogen atom, and glycine is a very important amino acid for just this reason. Owing to its lack of a side chain, glycine residues provide a site where the backbones of two polypeptides (or two segments of the same polypeptide) can approach one another very closely. In addition, glycine is more flexible than other amino acids and allows parts of the backbone to move or form a hinge. Proline is unique in having its  $\alpha$ -amino group as part of a ring (making it an imino acid). Proline is a hydrophobic amino acid that does not readily fit into an ordered secondary structure, such as an  $\alpha$  helix (page 54), often producing kinks or hinges. Cysteine contains a reactive sulphydryl ( $-SH$ ) group and is often covalently linked to another cysteine residue, as a **disulfide** ( $-SS-$ ) **bridge**.

### Cysteine



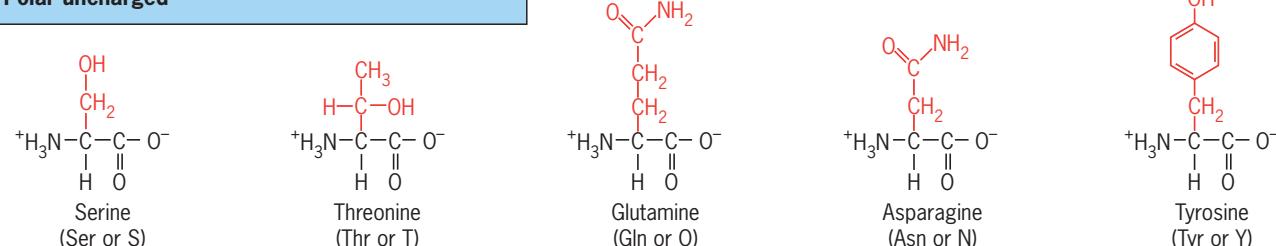
Disulfide bridges often form between two cysteines that are distant from one another in the polypeptide backbone or even in two separate polypeptides. Disulfide bridges help stabilize the intricate shapes of proteins, particularly those present outside of cells where they are subjected to added physical and chemical stress. For example, hair is made of a large number of filaments composed of the cysteine-rich protein keratin, cross-linked together by disulfide bridges. When someone gets a “perm” to make their hair curlier, the hairdresser puts the hair into curlers and then adds a reducing agent that breaks the disulfide bridges, letting the keratin filaments slide past each other. When the reducing agent is washed out, disulfide bridges re-form, locking the keratin in the new positions and causing the hair to be permanently locked into the shape imparted by the curlers. The same principle is used in hair-straightening treatments except that in this case the hairs are pulled out straight before the reducing agent is applied.

Not all of the amino acids described in this section are found in all proteins, nor are the various amino acids distributed in an equivalent manner. A number of other amino acids are also found in proteins, but they arise by alterations to the side chains of the 20 basic amino acids *after* their incorporation into a polypeptide chain. For this reason they are called **posttranslational modifications (PTMs)**. Dozens of different types of PTMs have been documented. The most widespread and important PTM is the reversible addition of a phosphate group to a serine, threonine, or tyrosine residue. Lysine

**Polar charged**

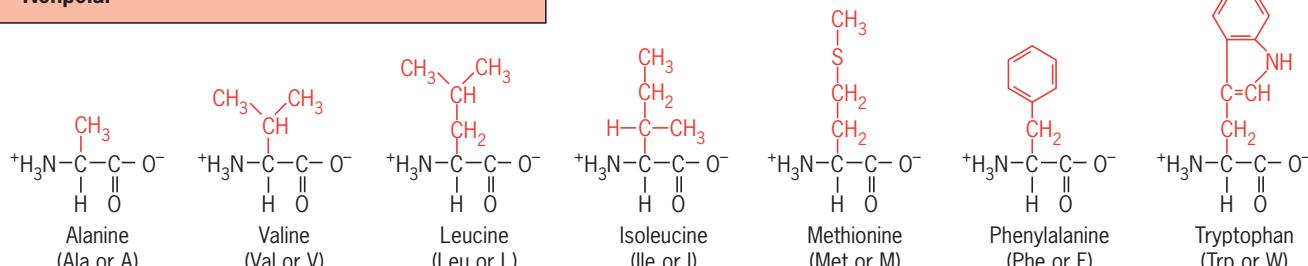
Properties of side chains (R groups):

Hydrophilic side chains act as acids or bases which tend to be fully charged (+ or -) under physiologic conditions.  
Side chains form ionic bonds and are often involved in chemical reactions.

**Polar uncharged**

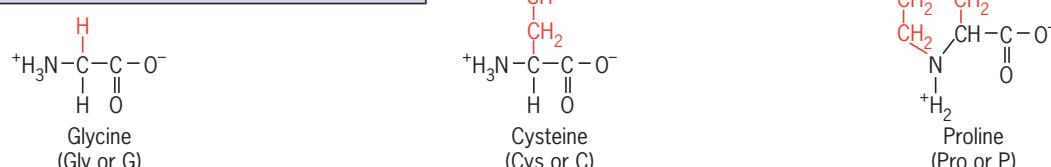
Properties of side chains:

Hydrophilic side chains tend to have partial + or - charge allowing them to participate in chemical reactions, form H-bonds, and associate with water.

**Nonpolar**

Properties of side chains:

Hydrophobic side chain consists almost entirely of C and H atoms. These amino acids tend to form the inner core of soluble proteins, buried away from the aqueous medium. They play an important role in membranes by associating with the lipid bilayer.

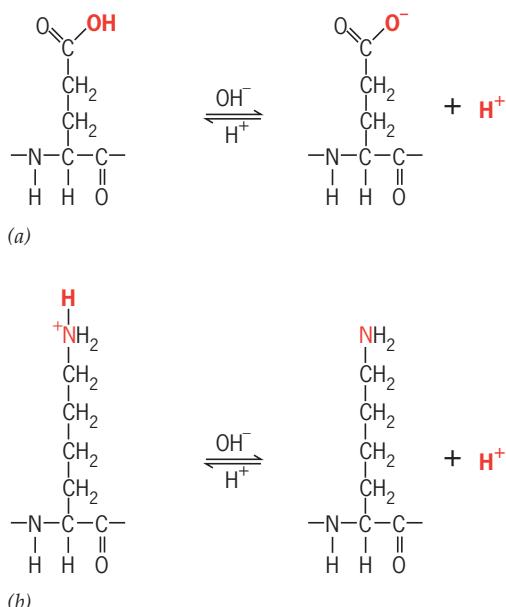
**Side chains with unique properties**

Side chain consists only of hydrogen atom and can fit into either a hydrophilic or hydrophobic environment. Glycine often resides at sites where two polypeptides come into close contact.

Though side chain has polar, uncharged character, it has the unique property of forming a covalent bond with another cysteine to form a disulfide link.

Though side chain has hydrophobic character, it has the unique property of creating kinks in polypeptide chains and disrupting ordered secondary structure.

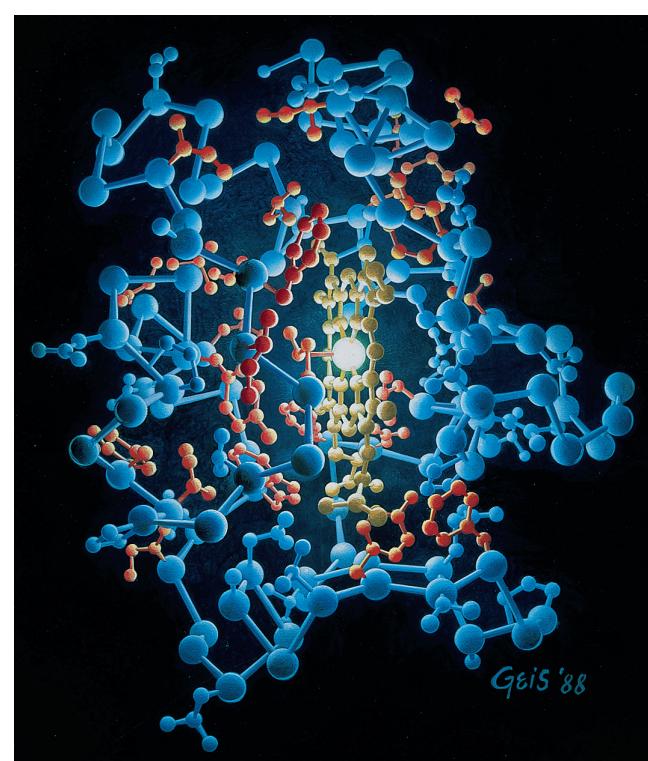
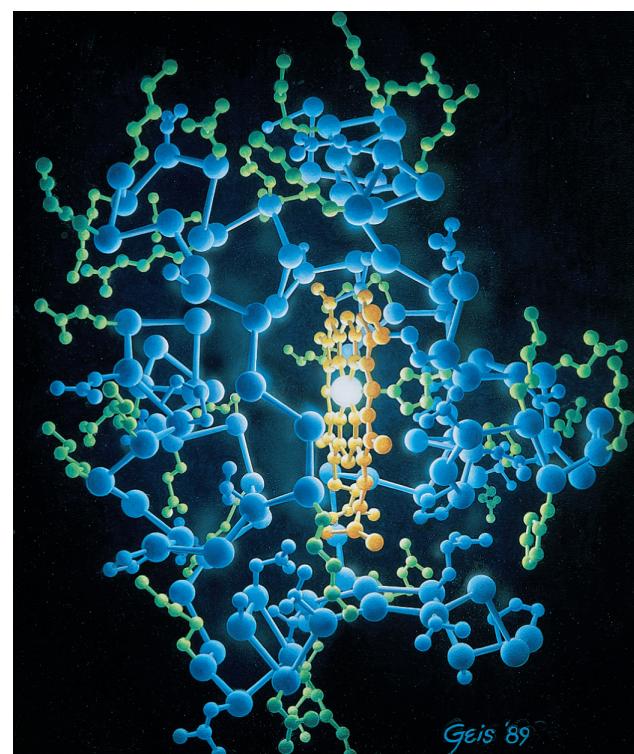
**FIGURE 2.26** The chemical structure of amino acids. These 20 amino acids represent those most commonly found in proteins and, more specifically, those encoded by DNA. Other amino acids occur as the result of a modification to one of those shown here. The amino acids are arranged into four groups based on the character of their side chains, as described in the text. All molecules are depicted as free amino acids in their ionized state as they would exist in solution at neutral pH.



**FIGURE 2.27** The ionization of charged, polar amino acids. (a) The side chain of glutamic acid loses a proton when its carboxylic acid group ionizes. The degree of ionization of the carboxyl group depends on the pH of the medium: the greater the hydrogen ion concentration (the lower the pH), the smaller the percentage of carboxyl groups that are present in the ionized state. Conversely, a rise in pH leads to an increased ionization of the proton from the carboxyl group, increasing the percentage of negatively charged glutamic acid side chains. The pH at which 50 percent of the side chains are ionized and 50 percent are unionized is called the pK, which is 4.4 for the side chain of free glutamic acid. At physiologic pH, virtually all of the glutamic acid residues of a polypeptide are negatively charged. (b) The side chain of lysine becomes ionized when its amino group gains a proton. The greater the hydroxyl ion concentration (the higher the pH), the smaller the percentage of amino groups that are positively charged. The pH at which 50 percent of the side chains of lysine are charged and 50 percent are uncharged is 10.0, which is the pK for the side chain of free lysine. At physiologic pH, virtually all of the lysine residues of a polypeptide are positively charged. Once incorporated into a polypeptide, the pK of a charged group can be greatly influenced by the surrounding environment.

acetylation is another widespread and important PTM affecting thousands of proteins in a mammalian cell. PTMs can generate dramatic changes in the properties and function of a protein, most notably by modifying its three-dimensional structure, level of activity, localization within the cell, life span, and/or its interactions with other molecules. The presence or absence of a single phosphate group on a key regulatory protein has the potential to determine whether a cell will behave as a cancer cell or a normal cell. Because of PTMs, a single polypeptide can exist as a number of distinct biological molecules.

The ionic, polar, or nonpolar character of amino acid side chains is very important in protein structure and function. Most soluble (i.e., nonmembrane) proteins are constructed so that the polar residues are situated at the surface of the molecule where they can associate with the surrounding water and contribute to the protein's solubility in aqueous solution (FIGURE 2.28a). In contrast, the nonpolar residues are situated predominantly in the core of the molecule (Figure 2.28b). The hydrophobic residues of the protein interior are often tightly packed together, creating a type of three-dimensional jigsaw puzzle in which water molecules are generally excluded. Hydrophobic interactions among the nonpolar side chains of these



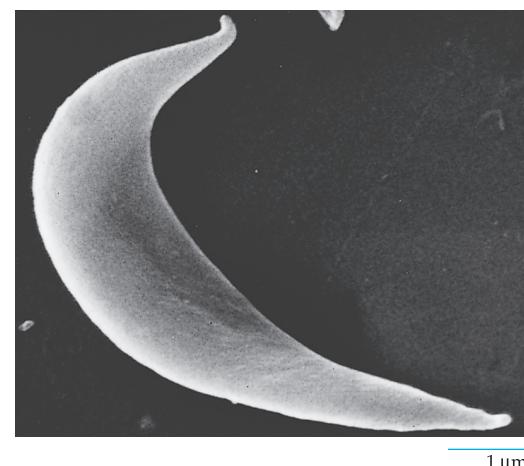
**FIGURE 2.28** Disposition of hydrophilic and hydrophobic amino acid residues in the soluble protein cytochrome c. (a) The hydrophilic side chains, which are shown in green, are located primarily at the surface of the protein where they contact the surrounding aqueous medium. (b) The hydrophobic residues, which are shown in red, are located primarily within the center of the protein, particularly in the vicinity of the central heme group.

SOURCE: Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.

residues are a driving force during protein folding (page 61) and contribute substantially to the overall stability of the protein. In many enzymes, reactive polar groups project into the nonpolar interior, giving the protein its catalytic activity. For example, a nonpolar environment can enhance ionic interactions between charged groups that would be lessened by competition with water in an aqueous environment. Some reactions that might proceed at an imperceptibly slow rate in water can occur in millionths of a second within the protein.

### REVIEW

- What are the major properties that distinguish different amino acids from one another? What roles do these differences play in the structure and function of proteins?
- What are the properties of glycine, proline, and cysteine that distinguish these amino acids?



**FIGURE 2.29** Scanning electron micrograph of a red blood cell from a person with sickle cell anemia. Compare with the micrograph of a normal red blood cell of Figure 4.32a.

SOURCE: Courtesy of J. T. Thornwaite, B. F. Cameron, and R. C. Leif.

## 2.9 Primary and Secondary Structures of Proteins

Nowhere in biology is the intimate relationship between form and function better illustrated than with proteins. The structure of most proteins is completely defined and predictable. Each amino acid in one of these giant macromolecules is located at a specific site within the structure, giving the protein the precise shape and reactivity required for the job at hand. Protein structure can be described at several levels of organization, each emphasizing a different aspect and each dependent on different types of interactions. Customarily, four such levels are described: *primary*, *secondary*, *tertiary*, and *quaternary*. The first, primary structure, concerns the amino acid sequence of a protein, whereas the latter three levels concern the organization of the molecule in space. To understand the mechanism of action and biological function of a protein it is essential to know how that protein is constructed.

### Primary Structure

The primary structure of a polypeptide is the specific linear sequence of amino acids that constitute the chain. With 20 different building blocks, the number of different polypeptides that can be formed is  $20^n$ , where  $n$  is the number of amino acids in the chain. Because most polypeptides contain well over 100 amino acids, the variety of possible sequences is essentially unlimited. The information for the precise order of amino acids in every protein that an organism can produce is encoded within the genome of that organism.

As we will see later, the amino acid sequence provides the information required to determine a protein's three-dimensional shape and thus its function. The sequence of amino acids, therefore, is all-important, and changes that arise in the sequence as a result of genetic mutations in the DNA may not be readily tolerated. The earliest and best-studied example of this relationship is the change in the amino acid sequence of hemoglobin that causes the disease *sickle cell anemia*. This severe, inherited anemia results solely from a single

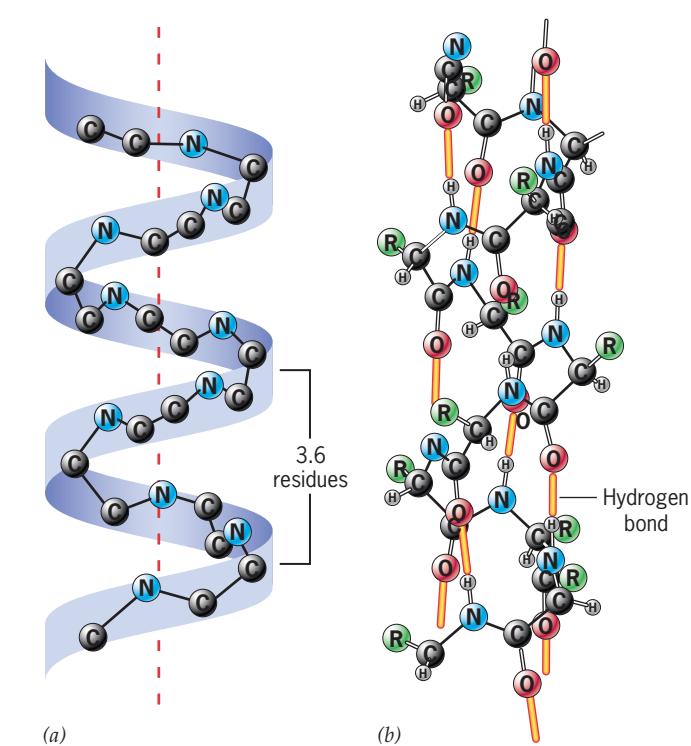
change in amino acid sequence within the hemoglobin molecule: A nonpolar valine residue is present where a charged glutamic acid is normally located. This change in hemoglobin structure can have a dramatic effect on the shape of red blood cells, converting them from disk-shaped cells to sickle-shaped cells (FIGURE 2.29), which tend to clog small blood vessels, causing pain and life-threatening crises. Not all amino acid changes have such a dramatic effect, as evidenced by the differences in amino acid sequence in the same protein among related organisms. The degree to which changes in the primary sequence are tolerated depends on the degree to which the shape of the protein or the critical functional residues are disturbed.

The first amino acid sequence of a protein was determined by Frederick Sanger and co-workers at Cambridge University in the early 1950s. Beef insulin was chosen for the study because of its availability and its small size—two polypeptide chains of 21 and 30 amino acids each. The sequencing of insulin was a momentous feat in the newly emerging field of molecular biology. It revealed that proteins, the most complex molecules in cells, have a definable substructure that is neither regular nor repeating, unlike those of polysaccharides. Each particular polypeptide, whether insulin or some other species, has a precise sequence of amino acids that does not vary from one molecule to another. With the advent of techniques for rapid DNA sequencing (see Section 18.22), the primary structure of a polypeptide can be deduced from the nucleotide sequence of the encoding gene. In the past few years, the complete sequences of the genomes of hundreds of organisms, including humans, have been determined. This information will eventually allow researchers to learn about every protein that an organism can manufacture. However, translating information about primary sequence into knowledge of higher levels of protein structure remains a formidable challenge.

### Secondary Structure

All matter exists in space and therefore has a three-dimensional shape. Proteins are formed by linkages among vast numbers of atoms; consequently their shape is complex. The term **conformation** refers to the three-dimensional arrangement of the atoms of a molecule, that is, to their spatial organization. Secondary structure describes the

**54** conformation of portions of the polypeptide chain. Early studies on secondary structure were carried out by Linus Pauling and Robert Corey of the California Institute of Technology. By studying the structure of simple peptides consisting of a few amino acids linked together, Pauling and Corey concluded that polypeptide chains exist in preferred conformations that provide the maximum possible number of hydrogen bonds between neighboring amino acids.

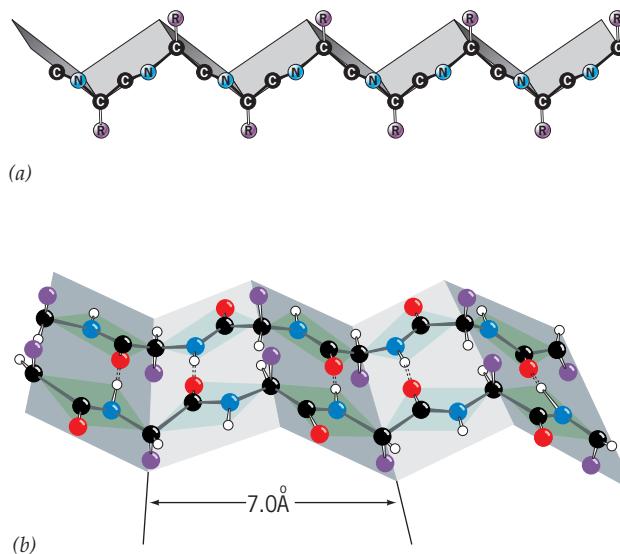


**FIGURE 2.30** The  $\alpha$  helix. (a) The helical path around a central axis taken by the polypeptide backbone in a region of  $\alpha$  helix. Each complete ( $360^\circ$ ) turn of the helix corresponds to 3.6 amino acid residues. The distance along the axis between adjacent residues is  $15\text{ \AA}$ . (b) The arrangement of the atoms of the backbone of the  $\alpha$  helix and the hydrogen bonds that form between amino acids. Because of the helical rotation, the peptide bonds of every fourth amino acid come into close proximity. The approach of the carbonyl group ( $\text{C=O}$ ) of one peptide bond to the imine group ( $\text{H-N}$ ) of another peptide bond results in the formation of hydrogen bonds between them. The hydrogen bonds (orange bars) are essentially parallel to the axis of the cylinder and thus hold the turns of the chain together.

SOURCE: (a, b) Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.

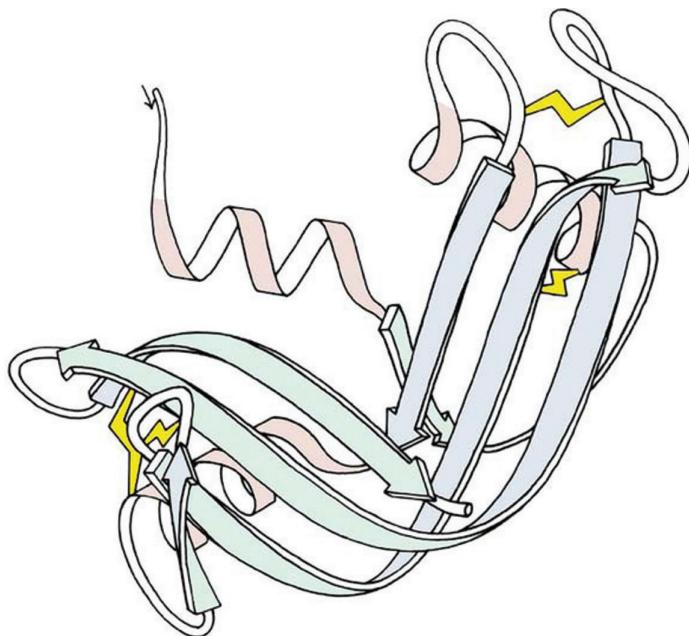
The second conformation proposed by Pauling and Corey was the **beta ( $\beta$ ) sheet**, which consists of several segments of a polypeptide lying side by side (FIGURE 2.31a). Unlike the coiled, cylindrical form of the  $\alpha$  helix, the backbone of each segment of polypeptide (or  **$\beta$  strand**) in a  $\beta$  sheet assumes a folded or pleated conformation (Figure 2.31a). Like the  $\alpha$  helix, the  $\beta$  sheet is also characterized by a large number of hydrogen bonds, but these are oriented perpendicular to the long axis of the polypeptide chain and project across from one part of the chain to another (Figure 2.31b). The strands of a  $\beta$  sheet can be arranged either parallel to one another, with neighboring strands running in the same direction, or antiparallel, with neighboring strands running in opposite directions. Figure 2.31b shows an antiparallel  $\beta$  sheet. Like the  $\alpha$  helix, the  $\beta$  sheet has also been found in many different proteins. Because  $\beta$  strands are highly extended, the  $\beta$  sheet resists pulling (tensile) forces. Silk is composed of a protein containing an extensive amount of  $\beta$  sheet; silk fibers are thought to owe their strength to this architectural feature. Remarkably, a single fiber of spider silk, which may be a tenth the thickness of a human hair, is roughly five times stronger than a steel fiber of comparable weight.

Those portions of a polypeptide chain not organized into an  $\alpha$  helix or a  $\beta$  sheet may consist of hinges, turns, loops, or finger-like extensions. Often, these are the most flexible portions of a polypeptide chain and the sites of greatest biological activity. For example, antibody molecules are known for their specific interactions with other molecules (antigens); these interactions are mediated by a series of loops at one end of the antibody molecule (see Figures 17.15 and 17.16). The various types of secondary structures are most simply depicted as shown in FIGURE 2.32:  $\alpha$  helices are represented by helical ribbons,  $\beta$  strands as flattened arrows, and connecting segments as thinner strands.



**FIGURE 2.31** The  $\beta$  sheet. (a) Each polypeptide of a  $\beta$  sheet assumes an extended but pleated conformation referred to as a  $\beta$  strand. The pleats result from the location of the  $\alpha$ -carbons above and below the plane of the sheet. Successive side chains (R groups in the figure) project upward and downward from the backbone. The distance along the axis between adjacent residues is  $3.5\text{ \AA}$ . (b) A  $\beta$  sheet consists of a number of  $\beta$  strands that lie parallel to one another and are joined together by a regular array of hydrogen bonds between the carbonyl and imine groups of the neighboring backbones. Neighboring segments of the polypeptide backbone may lie either parallel (in the same N-terminal  $\rightarrow$  C-terminal direction) or antiparallel (in the opposite N-terminal  $\rightarrow$  C-terminal direction).

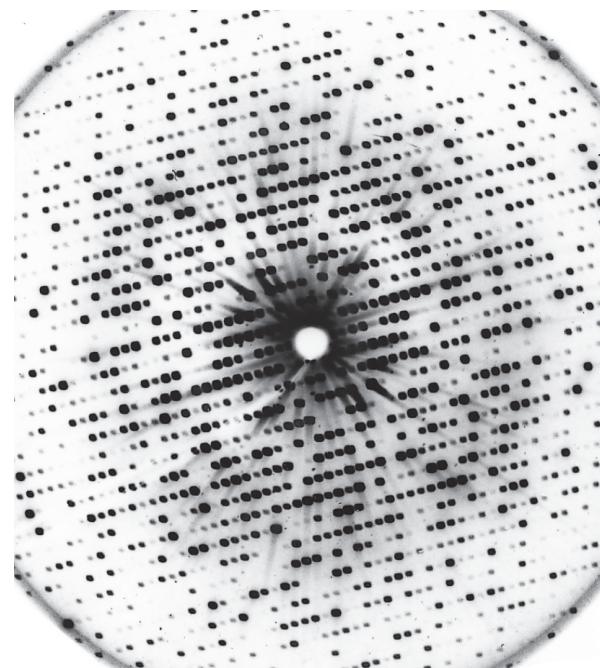
SOURCE: (b): Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.



**FIGURE 2.32** A ribbon model of ribonuclease. The regions of  $\alpha$  helix are depicted as spirals and  $\beta$  strands as flattened ribbons with the arrows indicating the N-terminal  $\rightarrow$  C-terminal direction of the polypeptide. Those segments of the chain that do not adopt a regular secondary structure (i.e., an  $\alpha$  helix or  $\beta$  strand) consist largely of loops and turns. Disulfide bonds are shown in yellow. SOURCE: Hand drawn by Jane S. Richardson.

#### REVIEW

- How are the properties of an  $\alpha$  helix different from a  $\beta$  strand? How are they similar?



**FIGURE 2.33** An X-ray diffraction pattern of myoglobin. The pattern of spots is produced as a beam of X-rays is diffracted by the atoms in the protein crystal, causing the X-rays to strike the film at specific sites. Information derived from the position and intensity (darkness) of the spots can be used to calculate the positions of the atoms in the protein that diffracted the beam, leading to complex structures such as that shown in Figure 2.35. SOURCE: Courtesy of John C. Kendrew.

## 2.10 Tertiary Structure of Proteins

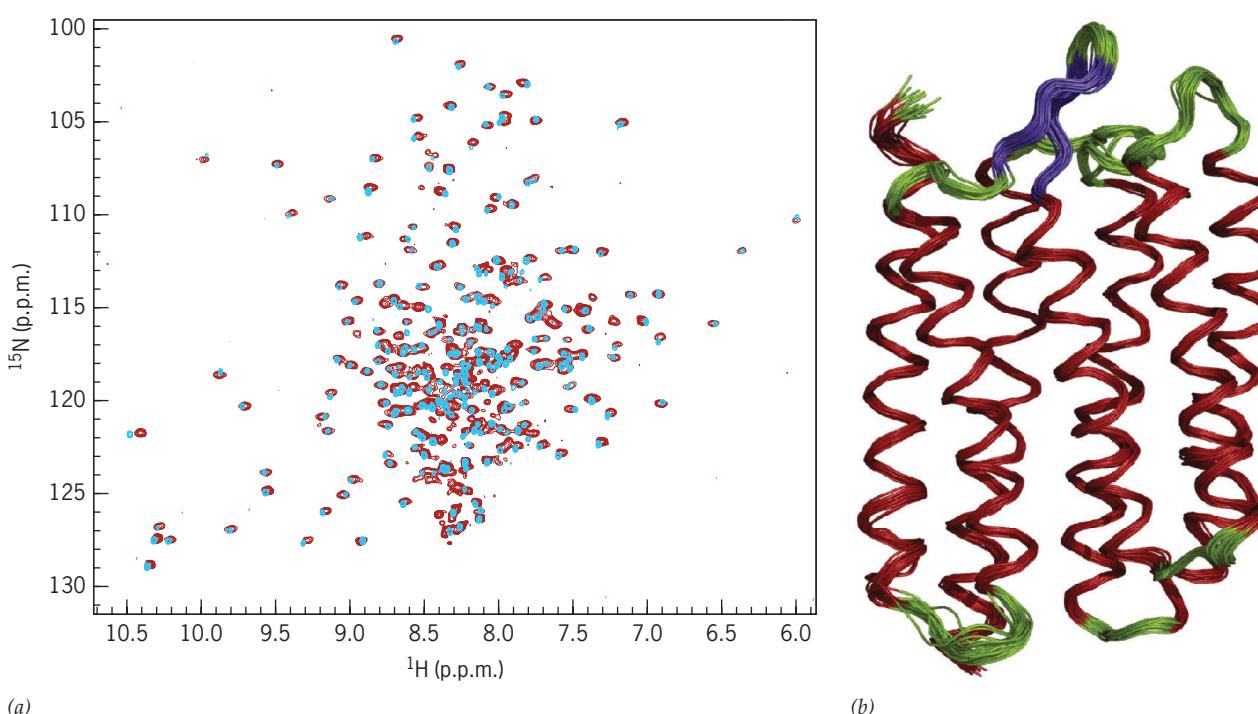
The next level above secondary structure is tertiary structure, which describes the conformation of the entire polypeptide. Whereas secondary structure is stabilized primarily by hydrogen bonds between atoms that form the peptide bonds of the backbone, tertiary structure is stabilized by an array of noncovalent bonds between the diverse side chains of the protein. Secondary structure is largely limited to a small number of conformations, but tertiary structure is virtually unlimited.

The detailed tertiary structure of a protein is usually determined using the technique of **X-ray crystallography**. In this technique (which is described in more detail in Sections 3.6 and 18.16), a crystal of the protein is bombarded by a thin beam of X-rays, and the radiation that is scattered (diffracted) by the electrons of the protein's atoms is allowed to strike a radiation-sensitive detector, forming an image of spots, such as those of **FIGURE 2.33**. These patterns of spots do not directly show the structure of the protein, but computer programs based on the mathematics of diffraction can then be used to derive the structure responsible for producing the pattern. Tertiary structure can also be determined by nuclear magnetic resonance (NMR) spectroscopy, a method in which proteins in solution are placed in a powerful magnetic field and probed with radio waves to produce a spectrum (**FIGURE 2.34a**) which provides information about the distances between atoms. As with X-ray crystallography, the spectrum itself does not directly show the protein structure, but

computer programs can derive the structures most likely to give the observed spectrum (Figure 2.34b). The two methods, X-ray crystallography and NMR, have different strengths and weaknesses: X-ray crystallography can provide higher resolution structures for larger proteins but is limited by the need to get any given protein to form pure crystals. NMR does not require crystallization, can provide information about dynamic changes in protein structure, and can rapidly reveal drug binding sites on a protein, but the method becomes increasingly difficult to apply as the size of the protein increases.

For many years it was presumed that all proteins had a fixed three-dimensional structure, which gave each protein its unique properties and specific functions. It came as a surprise to discover over the past decade or so that many proteins of higher organisms contain sizable segments that lack a defined conformation. Examples of proteins containing these types of unstructured (or *disordered*) segments can be seen in the models of the PrP protein in Figure 1 on page 63 and the histone tails in Figure 12.13c. The disordered regions in these proteins are depicted as dashed lines in the images, conveying the fact that these segments of the polypeptide (like pieces of spaghetti) can occupy many different positions and, thus, cannot be studied by X-ray crystallography. Disordered segments tend to have a predictable amino acid composition, being enriched in charged and polar residues and deficient in hydrophobic residues. You might be wondering whether proteins lacking a fully defined structure could be engaged in a useful function. In fact, the disordered regions of such proteins play key roles in vital cellular processes, often binding to DNA or to other proteins. Remarkably, these segments often undergo a physical transformation once they bind to an appropriate partner and are then seen to possess a defined, folded structure.

Most proteins can be categorized on the basis of their overall conformation as being either **fibrous proteins**, which have an



**FIGURE 2.34** NMR spectroscopy reveals tertiary structure without crystallization. (a) An NMR spectrum of the membrane-spanning helix protein sensory rhodopsin II. (b) Solving the structure of sensory rhodopsin II by NMR, showing thirty computed solutions consistent with the measured NMR spectra. The fact that all solutions agree on the overall shape indicates that we should have high confidence in the solution.

SOURCE: From Antoine Guatier et al., *Nature Struct. Mol. Biol.*, 17, 768. 2010.

elongated shape, or **globular proteins**, which have a compact shape. Most proteins that act as structural materials outside living cells are fibrous proteins, such as collagens and elastins of connective tissues, keratins of hair and skin, and silk. These proteins resist pulling or shearing forces to which they are exposed. In contrast, most proteins within the cell are globular proteins.

### Myoglobin: The First Globular Protein Whose Tertiary Structure Was Determined

The polypeptide chains of globular proteins are folded and twisted into complex shapes. Distant points on the linear sequence of amino acids are brought next to each other and linked by various types of bonds. The first glimpse at the tertiary structure of a globular protein came in 1957 through the X-ray crystallographic studies of John Kendrew and his colleagues at Cambridge University using X-ray diffraction patterns such as that shown in Figure 2.33. The protein they reported on was myoglobin.

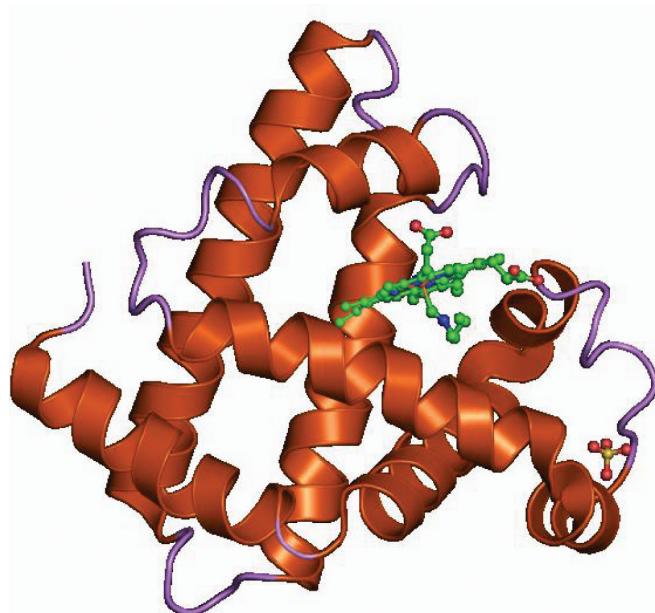
Myoglobin functions in muscle tissue as a storage site for oxygen; the oxygen molecule is bound to an iron atom in the center of a heme group. (The heme is an example of a *prosthetic group*, i.e., a portion of the protein that is not composed of amino acids, which is joined to the polypeptide chain after its assembly on the ribosome.) It is the heme group of myoglobin that gives most muscle tissue its reddish color. The first report on the structure of myoglobin provided a low-resolution profile sufficient to reveal that the molecule was compact (globular) and that the polypeptide chain was folded back on itself in a complex arrangement. There was no evidence of regularity or symmetry within the molecule, such as that revealed in the earlier description of the DNA double helix.

The earliest crude profile of myoglobin revealed eight rod-like stretches of  $\alpha$  helix ranging from 7 to 24 amino acids in length.

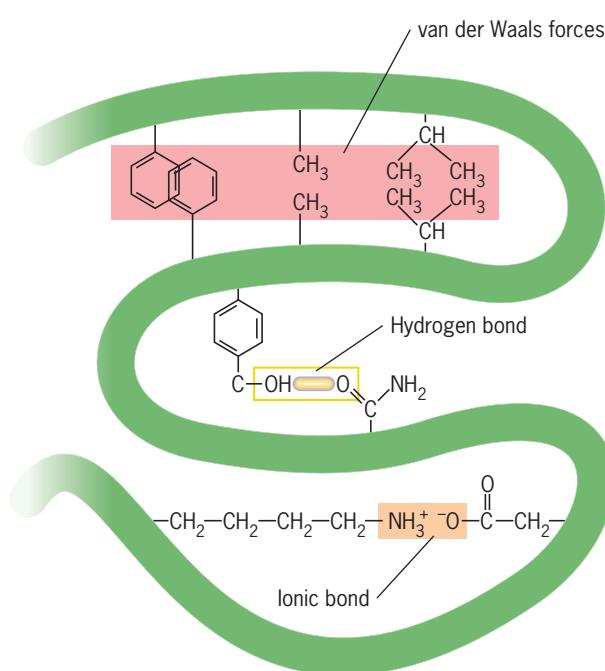
Altogether, approximately 75 percent of the 153 amino acids in the polypeptide chain are in the  $\alpha$ -helical conformation. This is an unusually high percentage compared with that for other proteins that have since been examined. No  $\beta$  sheet was found. Subsequent analyses of myoglobin using additional X-ray diffraction data provided a much more detailed picture of the molecule (FIGURES 2.35 and 3.16). For example, it was shown that the heme group is situated within a pocket of hydrophobic side chains that promotes the binding of oxygen without the oxidation (loss of electrons) of the iron atom. Myoglobin contains no disulfide bonds; the tertiary structure of the protein is held together exclusively by noncovalent interactions. All of the noncovalent bonds thought to occur between side chains within proteins—hydrogen bonds, ionic bonds, and van der Waals forces—have been found (FIGURE 2.36). Unlike myoglobin, most globular proteins contain both  $\alpha$  helices and  $\beta$  sheets. Most importantly, these early landmark studies revealed that each protein has a unique tertiary structure that can be correlated with its amino acid sequence and its biological function.

### Tertiary Structure May Reveal Unexpected Similarities between Proteins

Similarity in primary sequence is often used to decide whether two proteins may have similar structure and function. Sometimes, however, proteins that appeared to be unrelated at the primary sequence level have been found to have similar tertiary structures. Because it is the tertiary structure that determines the interactions and enzymatic activity of a protein, such structural similarity indicates that these proteins may have similar functions. FIGURE 2.37 shows two proteins, actin from eukaryotic cells and MreB from bacteria, whose primary sequence shows no similarity but whose tertiary structures



**FIGURE 2.35** The three-dimensional structure of myoglobin. The tertiary structure of myoglobin. Most of the amino acids are part of  $\alpha$  helices, shown in red. The nonhelical regions occur primarily as turns, where the polypeptide chain changes direction, shown in green. The position of the heme is indicated in yellow and blue with the coordinated iron atom in orange.  
SOURCE: This work is licensed under the Creative Commons Attribution-ShareAlike 3.0 License.

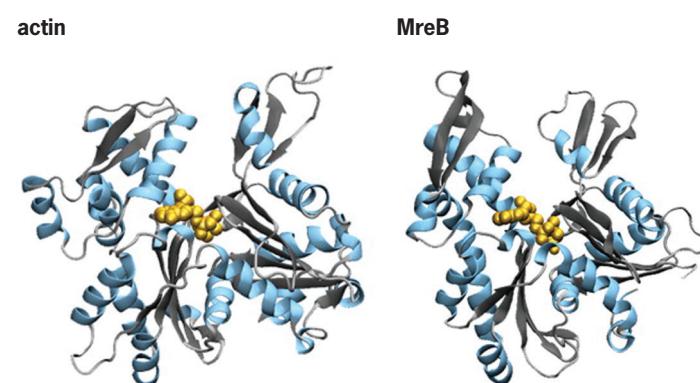


**FIGURE 2.36** Types of noncovalent bonds maintaining the conformation of proteins.

are clearly related. As discussed in Section 9.7, both proteins form cytoskeletal filaments.

## Protein Domains

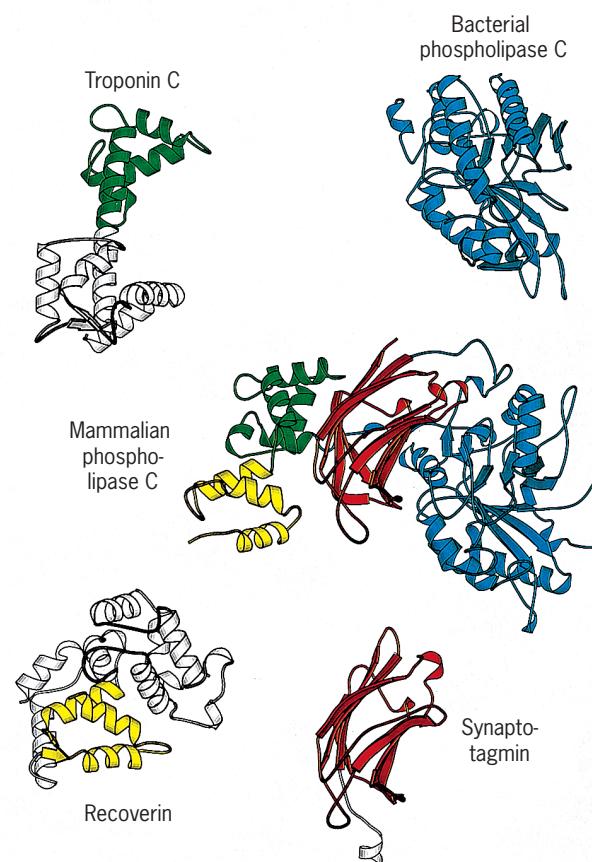
Unlike myoglobin, most eukaryotic proteins are composed of two or more spatially distinct modules, or **domains**, that fold independent of one another. For example, the mammalian enzyme phospholipase C, shown in the central part of **FIGURE 2.38**, consists



**FIGURE 2.37** Different sequence, similar structure. Two proteins, actin and MreB, whose primary sequences are completely different but that share a common tertiary structure.

SOURCE: From Bill Wickstead and Keith Gull, *J. Cell Biol.* 194:513–525, Fig. 1.

of four distinct domains, colored differently in the drawing. The different domains of a polypeptide often represent parts that function in a semi-independent manner. For example, they might bind different factors, such as a coenzyme and a substrate or a DNA strand and another protein, or they might move relatively independent of one another. Protein domains are often identified with a specific function. For example, proteins containing a PH domain bind to membranes containing a specific phospholipid, whereas



**FIGURE 2.38** Proteins are built of structural units, or domains. The mammalian enzyme phospholipase C is constructed of four domains, indicated in different colors. The catalytic domain of the enzyme is shown in blue. Each of the domains of this enzyme can be found independently in other proteins as indicated by the matching color.

SOURCE: From Liisa Holm and Chris Sander, *Structure* 5:167, © 2007, with permission from Elsevier.

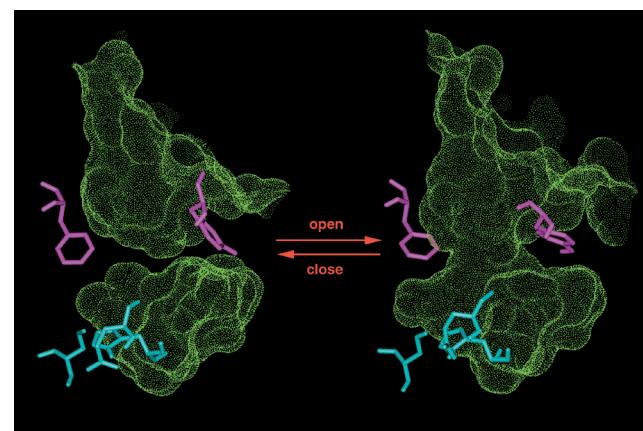
**58** proteins containing a chromodomain bind to a methylated lysine residue in another protein. The functions of a newly identified protein can often be predicted by the domains of which it is made.

Many polypeptides containing more than one domain are thought to have arisen during evolution by the fusion of genes that encoded different ancestral proteins, with each domain representing a part that was once a separate molecule. Each domain of the mammalian phospholipase C molecule, for example, has been identified as a homologous unit in another protein (Figure 2.38). Some domains have been found only in one or a few proteins. Other domains have been shuffled widely about during evolution, appearing in a variety of proteins whose other regions show little or no evidence of an evolutionary relationship. Shuffling of domains creates proteins with unique combinations of activities. On average, mammalian proteins tend to be larger and contain more domains than proteins of less complex organisms, such as fruit flies and yeast.

## Dynamic Changes within Proteins

Although X-ray crystallographic structures possess exquisite detail, they are static images frozen in time. Proteins, in contrast, are not static and inflexible, but capable of considerable internal movements. Proteins are, in other words, molecules with “moving parts.” Because they are tiny, nanometer-sized objects, proteins can be greatly influenced by the energy of their environment. Random, small-scale fluctuations in the arrangement of the bonds within a protein create an incessant thermal motion within the molecule. Spectroscopic techniques, such as nuclear magnetic resonance (NMR), can monitor dynamic movements within proteins, and they reveal shifts in hydrogen bonds, waving movements of external side chains, and the full rotation of the aromatic rings of tyrosine and phenylalanine residues about one of the single bonds. The important role that such movements can play in a protein’s function is illustrated in studies of acetylcholinesterase, the enzyme responsible for degrading acetylcholine molecules that are left behind following the transmission of an impulse from one nerve cell to another (Section 4.18). When the tertiary structure of acetylcholinesterase was first revealed by X-ray crystallography, there was no obvious pathway for acetylcholine molecules to enter the enzyme’s catalytic site, which was situated at the bottom of a deep gorge in the molecule. In fact, the narrow entrance to the site was completely blocked by the presence of a number of bulky amino acid side chains. Using high-speed computers, researchers have been able to simulate the random movements of thousands of atoms within the enzyme, a feat that cannot be accomplished using experimental techniques. These molecular dynamic (MD) simulations indicated that movements of side chains within the protein would lead to the rapid opening and closing of a “gate” that would allow acetylcholine molecules to diffuse into the enzyme’s catalytic site (FIGURE 2.39). The X-ray crystallographic structure of a protein (i.e., its *crystal structure*) can be considered an average structure, or “ground state.” A protein can undergo dynamic excursions from the ground state and assume alternate conformations that are accessible based on the amount of energy that the protein contains.

Predictable (nonrandom) movements within a protein that are triggered by the binding of a specific molecule are described as **conformational changes**. Conformational changes typically involve the coordinated movements of various parts of the molecule. A comparison of the polypeptides depicted in Figure 3a and 3b on page 69 shows the dramatic conformational change that occurs in a bacterial protein (GroEL) when it interacts with another protein (GroES).



**FIGURE 2.39** Dynamic movements within the enzyme acetylcholinesterase. A portion of the enzyme is depicted here in two different conformations: (1) a closed conformation (left) in which the entrance to the catalytic site is blocked by the presence of aromatic rings that are part of the side chains of tyrosine and phenylalanine residues (shown in purple) and (2) an open conformation (right) in which the aromatic rings of these side chains have swung out of the way, opening the “gate” to allow acetylcholine molecules to enter the catalytic site. These images are constructed using computer programs that take into account a host of information about the atoms that make up the molecule, including bond lengths, bond angles, electrostatic attraction and repulsion, van der Waals forces, etc. Using this information, researchers are able to simulate the movements of the various atoms over a very short time period, which provides images of the conformations that the protein can assume. An animation of this image can be found on the Web at <http://mccammon.ucsd.edu>

SOURCE: Courtesy of J. Andrew McCammon.

Virtually every activity in which a protein takes part is accompanied by conformational changes within the molecule (see <http://molmovdb.org> to watch examples). The conformational change in the protein myosin that occurs during muscle contraction is shown in Figure 9.52. In this case, binding of myosin to an actin molecule leads to a small ( $20^\circ$ ) rotation of the head of the myosin, which results in a 50 to 100 Å movement of the adjacent actin filament. The importance of this dynamic event can be appreciated if you consider that the movements of your body result from the additive effect of millions of conformational changes taking place within the contractile proteins of your muscles.

### REVIEW |

1. What are some of the differences between X-ray crystallography and NMR for determining protein structure? What can NMR tell us that X-ray crystallography cannot?

## 2.11 Quaternary Structure of Proteins

Whereas many proteins such as myoglobin are composed of only one polypeptide chain, most are found in complexes of more than one chain, or **subunit**. The subunits may be linked by covalent disulfide bonds, but most often they are held together by noncovalent bonds as occur

typically between hydrophobic “patches” on the complementary surfaces of neighboring polypeptides. Protein complexes are said to have **quaternary structure**. Depending on the protein, the polypeptide chains may be identical or nonidentical. A protein complex composed of two identical subunits is described as a *homodimer*, whereas a protein complex composed of two nonidentical subunits is a *heterodimer*. A ribbon drawing of a homodimeric protein is depicted in **FIGURE 2.40a**. The two subunits of the protein are drawn in different colors, and the hydrophobic residues that form the complementary sites of contact are indicated.

### The Structure of Hemoglobin

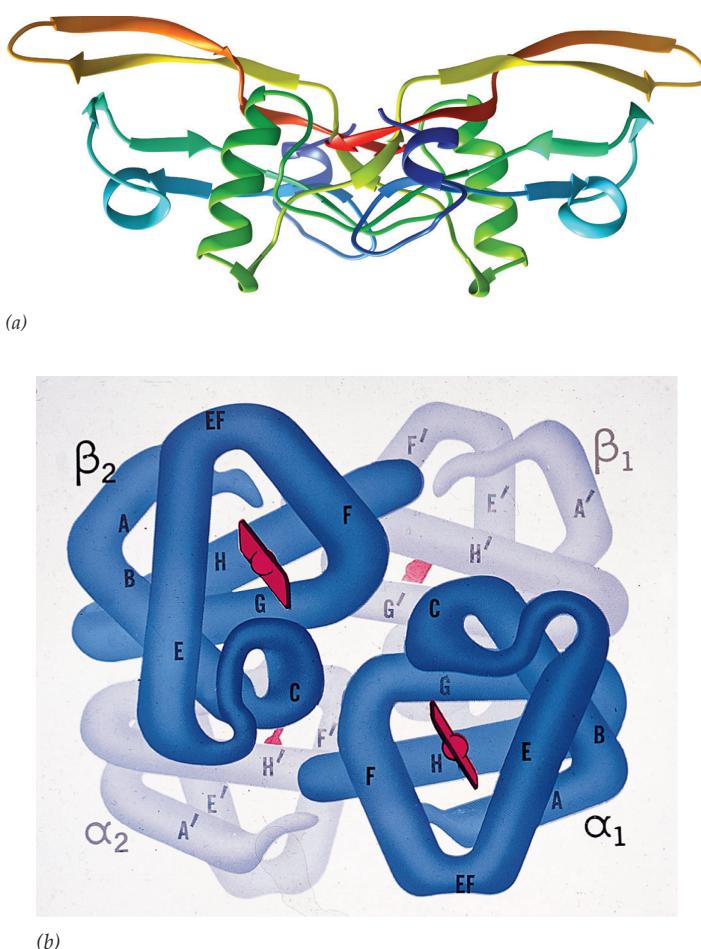
The best-studied multisubunit protein is hemoglobin, the O<sub>2</sub>-carrying protein of red blood cells. A molecule of human hemoglobin consists of two α-globin and two β-globin polypeptides (Figure 2.40b), each of which binds a single molecule of oxygen. Elucidation of the three-dimensional structure of hemoglobin by Max Perutz of Cambridge University in 1959 was one of the early landmarks in the

study of molecular biology. Perutz demonstrated that each of the four globin polypeptides of a hemoglobin molecule has a tertiary structure similar to that of myoglobin, a fact that strongly suggested the two proteins had evolved from a common ancestral polypeptide with a common O<sub>2</sub>-binding mechanism. Perutz also compared the structure of the oxygenated and deoxygenated versions of hemoglobin. In doing so, he discovered that the binding of oxygen was accompanied by the movement of the bound iron atom closer to the plane of the heme group. This seemingly inconsequential shift in position of a single atom pulled on an α helix to which the iron is connected, which in turn led to a series of increasingly larger movements within and between the subunits. This finding revealed for the first time that the complex functions of proteins may be carried out by means of small changes in their conformation.

### Protein–Protein Interactions

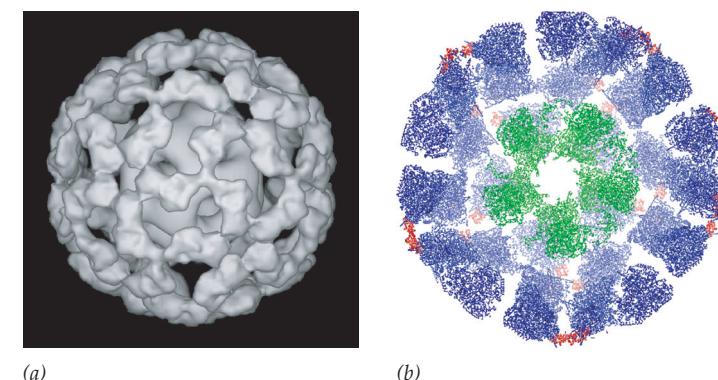
Even though hemoglobin consists of four subunits, it is still considered a single protein with a single function. Many examples are known in which different proteins, each with a specific function, become physically associated to form a much larger **multiprotein complex**. One of the first multiprotein complexes to be discovered and studied was the pyruvate dehydrogenase complex of the bacterium *Escherichia coli*, which consists of 60 polypeptide chains constituting three different enzymes (**FIGURE 2.41**). The enzymes that make up this complex catalyze a series of reactions connecting two metabolic pathways, glycolysis and the TCA cycle (see Figure 5.7). Because the enzymes are so closely associated, the product of one enzyme can be channeled directly to the next enzyme in the sequence without becoming diluted in the aqueous medium of the cell.

Multiprotein complexes that form within the cell are not necessarily stable assemblies, such as the pyruvate dehydrogenase complex. In fact, most proteins interact with other proteins in highly dynamic patterns, associating and dissociating depending on conditions within the cell at any given time. Interacting proteins tend to have complementary surfaces. Often a projecting portion of one molecule fits into a pocket



**FIGURE 2.40** Proteins with quaternary structure. (a) Drawing of transforming growth factor-β2 (TGF-β2), a protein that is a dimer composed of two identical subunits. (b) Drawing of a hemoglobin molecule, which consists of two α-globin chains and two β-globin chains (a heterotetramer) joined by noncovalent bonds. When the four globin polypeptides are assembled into a complete hemoglobin molecule, the kinetics of O<sub>2</sub> binding and release are quite different from those exhibited by isolated polypeptides. This is because the binding of O<sub>2</sub> to one polypeptide causes a conformational change in the other polypeptides that alters their affinity for O<sub>2</sub> molecules.

SOURCE: (a) From www.rcsb.org; (b) Illustration, Irving Geis. Image from Irving Geis reproduced with permission from AAAS. Collection/ Howard Hughes Medical Institute. Rights owned by HHMI. Reproduced by permission only.



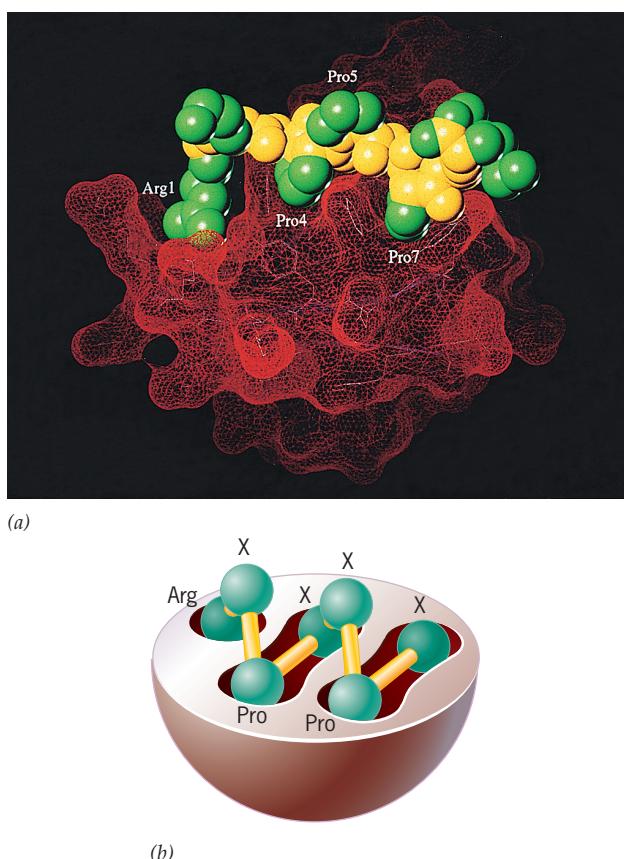
**FIGURE 2.41** Pyruvate dehydrogenase: a multiprotein complex.

(a) Three-dimensional reconstruction of the bacterial pyruvate dehydrogenase E1E2 subcomplex determined by electron microscopy. Its molecular mass is 11 million daltons. (b) Molecular structure of the pyruvate dehydrogenase complex determined by fitting crystal structures of subunit proteins into the large-scale structure seen in electron microscopy. The core of the complex consists of dihydrolipoyl acetyltransferase molecules (green). Pyruvate dehydrogenase tetramers (blue) form an outer shell around the core. The full pyruvate dehydrogenase complex contains a third protein that is not resolved here, and is even larger.

SOURCE: From Jacqueline Milne et al., 2002. *Embo J.* 21:5587–98.

within its partner. Once the two molecules have come into close contact, their interaction is stabilized by noncovalent bonds.

The reddish-colored object in **FIGURE 2.42a** is called an SH3 domain, and it is found as part of more than 200 different proteins involved in molecular signaling. The surface of an SH3 domain contains shallow hydrophobic “pockets” that become filled by complementary “knobs” projecting from another protein (Figure 2.42b). A large number of different structural domains have been identified that, like SH3, act as adaptors to mediate interactions between proteins. In many cases, protein–protein interactions are regulated by modifications, such as the addition of a phosphate group to a key amino acid, which act as a switch to turn on or off the protein’s ability to bind a protein partner. As more and more complex molecular activities have been discovered, the importance of interactions among proteins has become increasingly apparent. For example, such diverse processes as DNA synthesis, ATP formation, and RNA processing are all accomplished by “molecular machines” that consist of a large number of interacting proteins, some of which form stable relationships and others transient liaisons. Several hundred different protein complexes have been purified in large-scale studies on yeast.



**FIGURE 2.42** Protein–protein interactions. (a) A model illustrating the complementary molecular surfaces of portions of two interacting proteins. The reddish-colored molecule is an SH3 domain of the enzyme PI3K, whose function is discussed in Chapter 15. This domain binds specifically to a variety of proline-containing peptides, such as the one shown in the space-filling model at the top of the figure. The proline residues in the peptide, which fit into hydrophobic pockets on the surface of the enzyme, are indicated. The polypeptide backbone of the peptide is colored yellow, and the side chains are colored green. (b) Schematic model of the interaction between an SH3 domain and a peptide showing the manner in which certain residues of the peptide fit into hydrophobic pockets in the SH3 domain.

SOURCE: From Hongtao Yu and Stuart Schreiber, *Cell* 76:940, © 1994, with permission from Elsevier. Courtesy of Hongtao Yu and Stuart Schreiber.

## REVIEW |

1. Describe the difference between primary, secondary, tertiary, and quaternary structure.

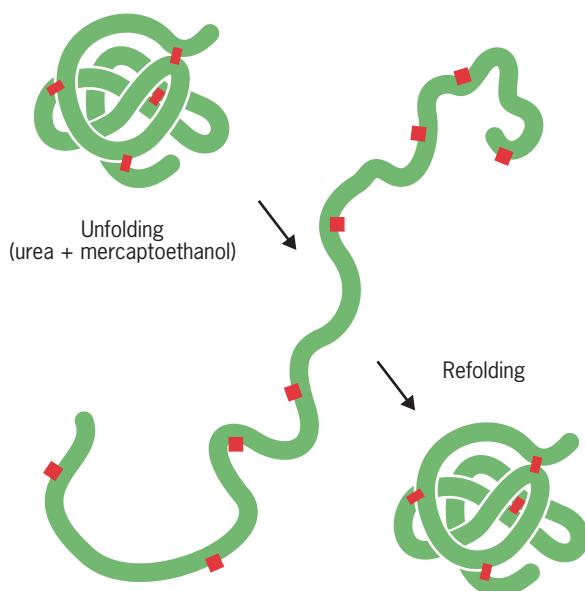
## 2.12 Protein Folding

The elucidation of the tertiary structure of myoglobin in the late 1950s led to an appreciation for the complexity of protein architecture. An important question immediately arose: How does such a complex, folded, asymmetric organization arise in the cell? The first insight into this problem began with a serendipitous observation in 1956 by Christian Anfinsen at the National Institutes of Health. Anfinsen was studying the properties of ribonuclease A, a small enzyme that consists of a single polypeptide chain of 124 amino acids with four disulfide bonds linking various parts of the chain. The disulfide bonds of a protein are typically broken (reduced) by adding a reducing agent, such as mercaptoethanol, which converts each disulfide bridge to a pair of sulphydryl (—SH) groups (see drawing, page 50). To make all of the disulfide bonds accessible to the reducing agent, Anfinsen found that the molecule had to first be partially unfolded. The unfolding or disorganization of a protein is termed **denaturation**, and it can be brought about by a variety of agents, including detergents, organic solvents, radiation, heat, and compounds such as urea and guanidine chloride, all of which interfere with the various interactions that stabilize a protein’s tertiary structure.

When Anfinsen treated ribonuclease molecules with mercaptoethanol and concentrated urea, he found that the preparation lost all of its enzymatic activity, which would be expected if the protein molecules had become totally unfolded. When he removed the urea and mercaptoethanol from the preparation, he found, to his surprise, that the molecules regained their normal enzymatic activity. The active ribonuclease molecules that had re-formed from the unfolded protein were indistinguishable both structurally and functionally from the correctly folded (i.e., **native**) molecules present at the beginning of the experiment (**FIGURE 2.43**). After extensive study, Anfinsen concluded that the linear sequence of amino acids contained all of the information required for the formation of the polypeptide’s three-dimensional conformation. Ribonuclease, in other words, is capable of **self-assembly**. As discussed in Chapter 3, events tend to progress toward states of lower energy. According to this concept, the tertiary structure that a polypeptide chain assumes after folding is the accessible structure with the lowest energy, which makes it the most thermodynamically stable structure that can be formed by that chain. It would appear that evolution selects for those amino acid sequences that generate a polypeptide chain capable of spontaneously arriving at a meaningful native state in a biologically reasonable time period.

## Dynamics of Protein Folding

There have been numerous controversies in the study of protein folding. Many of these controversies stem from the fact that the field is characterized by highly sophisticated experimental, spectroscopic, and computational procedures that are required to study complex molecular events that typically occur on a microsecond timescale. These efforts have often yielded conflicting results and have generated data that are open to more than one interpretation. For the sake of simplicity, we will restrict the discussion to “simple” proteins, such as ribonuclease, that consist of a single domain. One fundamental issue that has been



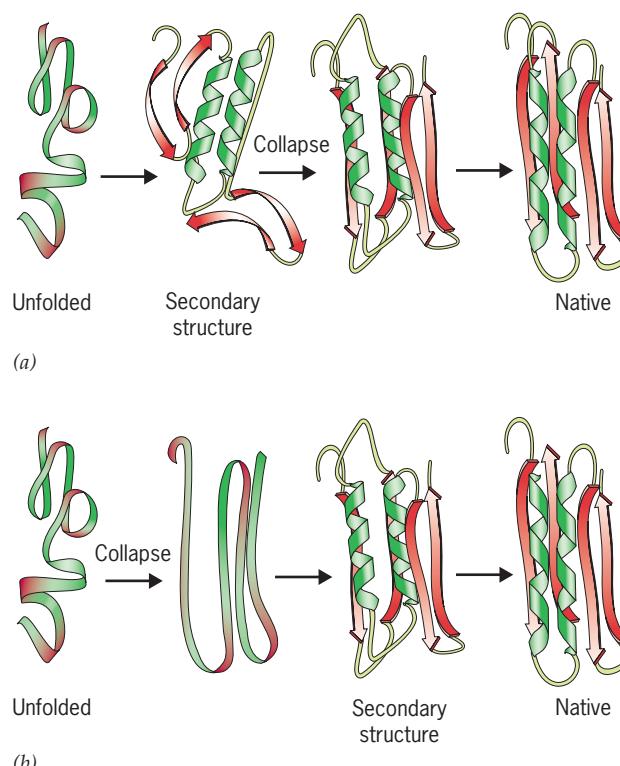
**FIGURE 2.43** Denaturation and refolding of ribonuclease. A native ribonuclease molecule (with intramolecular disulfide bonds indicated) is reduced and unfolded with  $\beta$ -mercaptoethanol and 8 M urea. After removal of these reagents, the protein undergoes spontaneous refolding.

SOURCE: From C. J. Epstein, R. F. Goldberger, and C. B. Anfinsen, *Cold Spring Harbor Symp. Quant. Biol.* 28:439, 1963. Reprinted with permission from Cold Spring Harbor Laboratory Press.

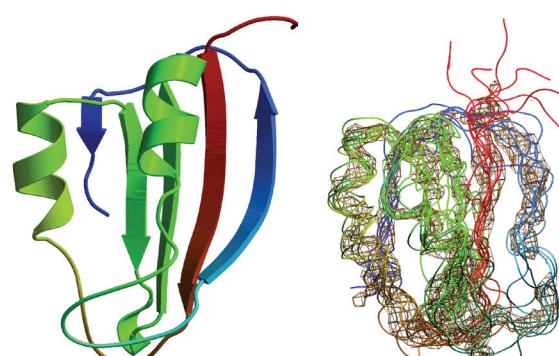
extensively debated is whether all of the members of a population of unfolded proteins of a single species fold along a similar pathway or fold by means of a diverse set of routes that somehow converge upon the same native state. Recent studies that have simulated folding with the aid of high speed computers have suggested that both views are right, and that proteins initially explore a wide range of different conformations when they first begin to fold but eventually funnel down into an increasingly restricted set of possible configurations (see Figure 2.45).

Another issue that has been roundly debated concerns the types of events that occur at various stages during the folding process. In the course depicted in **FIGURE 2.44a**, protein folding is initiated by interactions among neighboring residues that lead to the formation of much of the secondary structure of the molecule. Once the  $\alpha$  helices and  $\beta$  sheets are formed, subsequent folding is driven by hydrophobic interactions that bury nonpolar residues together in the central core of the protein. According to an alternate scheme shown in Figure 2.44b, the first major event in protein folding is the hydrophobic collapse of the polypeptide to form a compact structure in which the backbone adopts a native-like topology. Only after this collapse does significant secondary structure develop. Recent studies indicate that the two pathways depicted in Figure 2.44 lie at opposite extremes and that most proteins probably fold by a middle-of-the-road scheme in which secondary structure formation and compaction occur simultaneously. These early folding events lead to the formation of a partially folded, transient structure that resembles the native protein but lacks many of the specific interactions between amino acid side chains that are present in the fully folded molecule (**FIGURE 2.45**).

If the information that governs folding is embedded in a protein's amino acid sequence, then it should be possible to predict the tertiary structure of the protein just from its sequence. Such *de novo* structural prediction has been a holy grail in protein science and still remains an unsolved problem. However, if a protein of interest is closely related at the primary sequence level with another protein



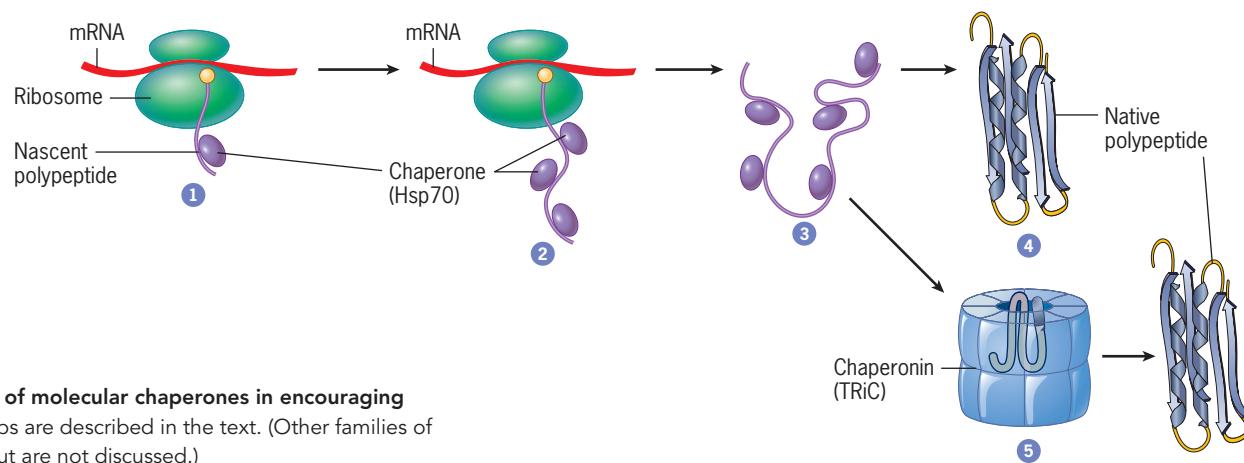
**FIGURE 2.44** Two alternate pathways by which a newly synthesized or denatured protein could achieve its native conformation. Curled segments represent  $\alpha$  helices, and arrows represent  $\beta$  strands.



**FIGURE 2.45** Along the folding pathway. The image on the left shows the native tertiary structure of the enzyme acyl-phosphatase. The image on the right is the transition structure, which represents the state of the molecule at the top of an energy barrier that must be crossed if the protein is going to reach the native state. The transition structure consists of numerous individual lines because it is a set (ensemble) of closely related structures. The overall architecture of the transition structure is similar to that of the native protein, but many of the finer structural features of the fully folded protein have yet to emerge. Conversion of the transition state to the native protein involves completing secondary structure formation, tighter packing of the side chains, and finalizing the burial of hydrophobic side chains from the aqueous solvent.

SOURCE: From K. Lindorff-Larsen, et al, *Trends Biochem. Sci.* 30:14, 2005, Fig. 1B. © 2005, with permission from Elsevier. Image from Christopher Dobson.

whose tertiary structure is known, then it does become possible to make a reasonable guess about the tertiary structure of the unknown protein by aligning the amino acids of the unknown protein onto the corresponding amino acids in the protein whose structure is known, a process known as **threading**. The fact that primary sequence



**FIGURE 2.46** The role of molecular chaperones in encouraging protein folding. The steps are described in the text. (Other families of chaperones are known but are not discussed.)

determines the folding of a protein means that alterations in this sequence have the potential to change the way a protein folds, leading to an abnormal tertiary structure. In fact, many mutations responsible for inherited disorders have been found to alter a protein's three-dimensional structure. In some cases, the consequences of protein misfolding can be fatal. Two examples of fatal neurodegenerative diseases that result from abnormal protein folding are discussed in Section 2.13.

### The Role of Molecular Chaperones

Not all proteins are able to assume their final tertiary structure by a simple process of self-assembly. This is not because the primary structure of these proteins lacks the required information for proper folding, but rather because proteins undergoing folding have to be prevented from interacting nonselectively with other molecules in the crowded compartments of the cell. Several families of proteins have evolved whose function is to help unfolded or misfolded proteins achieve their proper three-dimensional conformation. These “helper proteins” are called **molecular chaperones**, and they selectively bind to short stretches of hydrophobic amino acids that tend to be exposed in non-native proteins but buried in proteins having a native conformation.

**FIGURE 2.46** depicts the activities of two families of molecular chaperones that operate in the cytosol of eukaryotic cells. Molecular chaperones are involved in a multitude of activities within cells, ranging from the import of proteins into organelles (see Figure 8.47a) to the prevention and reversal of protein aggregation. We will restrict the discussion to their actions on newly synthesized proteins.

Polypeptide chains are synthesized on ribosomes by the addition of amino acids, one at a time, beginning at the chain's N-terminus (step 1, Figure 2.46). Chaperones of the Hsp70 family bind to elongating polypeptide chains as they emerge from an exit channel within the large subunit of the ribosome (step 2). Hsp70 chaperones are thought to prevent these partially formed polypeptides (i.e., *nascent* polypeptides) from binding to other proteins in the cytosol, which would cause them either to aggregate or misfold. Once their synthesis has been completed (step 3), many of these proteins are simply released by the chaperones into the cytosol where they spontaneously fold into their native state (step 4). Other proteins are repeatedly bound and released by chaperones until they finally reach their fully folded state. Many of the larger polypeptides are transferred from Hsp70 proteins to a different type of chaperone called a *chaperonin* (step 5). Chaperonins are cylindrical protein complexes that contain chambers in which newly synthesized polypeptides can fold without interference from other macromolecules in the cell. TRiC is a chaperonin thought to assist in the folding of up to 15 percent of the polypeptides synthesized in mammalian cells. The discovery and mechanism of action of Hsp70 and chaperonins are discussed in depth in the Experimental Pathways in Section 2.14 on page 67.

### REVIEW |

- Given that proteins act as molecular machines, explain why conformational changes are so important in protein function.

## 2.13 THE HUMAN PERSPECTIVE

### Protein Misfolding Can Have Deadly Consequences

In April 1996 a paper was published in the medical journal *Lancet* that generated widespread alarm in the populations of Europe. The paper described a study of 10 persons afflicted with Creutzfeld-Jakob disease (CJD), a rare, fatal disorder that attacks the brain, causing a loss of motor coordination and dementia. Like numerous other diseases, CJD can occur as an inherited disease that runs in certain families or as a sporadic form that appears in individuals who have no family history of the disease. Unlike virtually every

other inheritable disease, however, CJD can also be acquired. Until recently, persons who had acquired CJD had been recipients of organs or organ products that were donated by a person with undiagnosed CJD. The cases described in the 1996 *Lancet* paper had also been acquired, but the apparent source of the disease was contaminated beef that the infected individuals had eaten years earlier. The contaminated beef was derived from cattle raised in England that had contracted a neurodegenerative disease that

caused the animals to lose motor coordination and develop demented behavior. The disease became commonly known as "mad cow disease." Patients who have acquired CJD from eating contaminated beef can be distinguished by several criteria from those who suffer from the classical forms of the disease. To date, roughly 200 people have died of CJD acquired from contaminated beef, and the numbers of such deaths have been declining.<sup>1</sup>

A disease that runs in families can invariably be traced to a faulty gene, whereas diseases that are acquired from a contaminated source can invariably be traced to an infectious agent. How can the same disease be both inherited and infectious? The answer to this question has emerged gradually over the past several decades, beginning with observations by D. Carleton Gajdusek in the 1960s on a strange malady that once afflicted the native population of Papua, New Guinea. Gajdusek showed that these islanders were contracting a fatal neurodegenerative disease—which they called "kuru"—during a funeral ritual in which they ate the brain tissue of a recently deceased relative. Autopsies of the brains of patients who had died of kuru showed a distinct pathology, referred to as *spongiform encephalopathy*, in which certain brain regions were riddled with microscopic holes (vacuolations), causing the tissue to resemble a sponge.

It was soon shown that the brains of islanders suffering from kuru were strikingly similar in microscopic appearance to the brains of persons afflicted with CJD. This observation raised an important question: Did the brain of a person suffering from CJD, which was known to be an inherited disease, contain an infectious agent? In 1968, Gajdusek showed that when extracts prepared

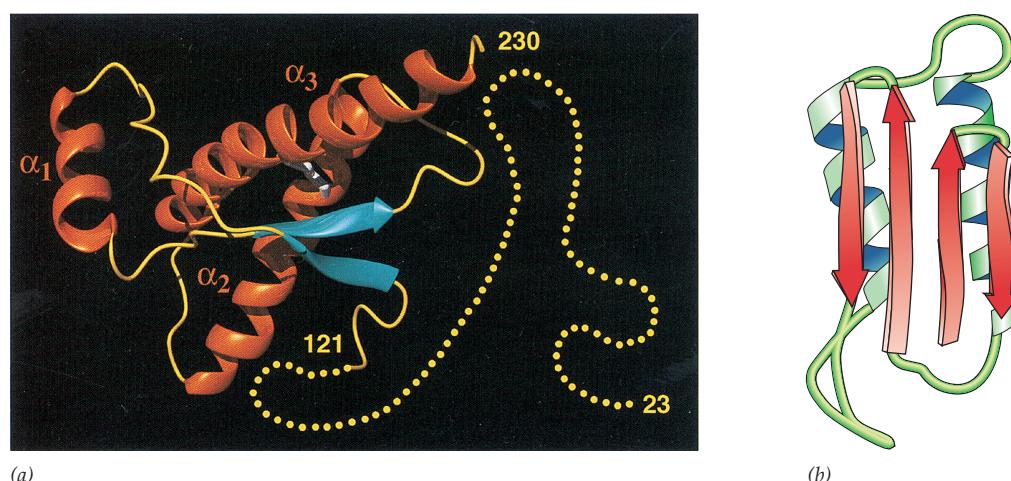
<sup>1</sup>On the surface, this would suggest that the epidemic has run its course, but there are several reasons for public health officials to remain concerned. For one, studies of tissues that had been removed during surgeries in England indicate that thousands of people are likely to be infected with the disease without exhibiting symptoms (discussed in *Science* 335:411, 2012). Even if these individuals never develop clinical disease, they remain potential carriers who could pass CJD on to others through blood transfusions. In fact, at least two individuals are believed to have contracted CJD after receiving blood from a donor harboring the disease. These findings underscore the need to test blood for the presence of the responsible agent (whose nature is discussed below).

from a biopsy of the brain of a person who had died from CJD were injected into a suitable laboratory animal, that animal did indeed develop a spongiform encephalopathy similar to that of kuru or CJD. Clearly, the extracts contained an infectious agent, which at the time was presumed to be a virus.

In 1982, Stanley Prusiner of the University of California, San Francisco, published a paper suggesting that, unlike viruses, the infectious agent responsible for CJD lacked nucleic acid and instead was composed solely of protein. He called the protein a prion. This "protein only" hypothesis, as it is called, was originally met with considerable skepticism, but subsequent studies by Prusiner and others have provided overwhelming support for the proposal. It was presumed initially that the prion protein was an external agent—some type of virus-like particle lacking nucleic acid. Contrary to this expectation, the prion protein was soon shown to be encoded by a gene (called *PRNP*) within the cell's own chromosomes. The gene is expressed within *normal* brain tissue and encodes a protein designated  $\text{PrP}^C$  (standing for prion protein cellular) that resides at the surface of nerve cells. The precise function of  $\text{PrP}^C$  remains a mystery. A modified version of the protein (designated  $\text{PrP}^{Sc}$ , standing for prion protein scrapie) is present in the brains of humans with CJD. Unlike the normal  $\text{PrP}^C$ , the modified version of the protein accumulates within nerve cells, forming aggregates that kill the cells.

In their purified states,  $\text{PrP}^C$  and  $\text{PrP}^{Sc}$  have very different physical properties.  $\text{PrP}^C$  remains as a monomeric molecule that is soluble in salt solutions and is readily destroyed by protein-digesting enzymes. In contrast,  $\text{PrP}^{Sc}$  molecules interact with one another to form insoluble fibrils that are resistant to enzymatic digestion. Based on these differences, one might expect these two forms of the PrP protein to be composed of distinctly different sequences of amino acids, but this is not the case. The two forms can have identical amino acid sequences, but they differ in the way the polypeptide chain folds to form the three-dimensional protein molecule (FIGURE 1). Whereas a  $\text{PrP}^C$  molecule consists largely of  $\alpha$ -helical segments and interconnecting coils, the core of a  $\text{PrP}^{Sc}$  molecule consists largely of  $\beta$  sheet.

It is not hard to understand how a mutant polypeptide might be less stable and more likely to fold into the abnormal  $\text{PrP}^{Sc}$  conformation, but how is such a protein able to act as an infectious



**FIGURE 1** A contrast in structure. (a) Tertiary structure of the normal ( $\text{PrP}^C$ ) protein as determined by NMR spectroscopy. The orange portions represent  $\alpha$ -helical segments, and the blue portions are short  $\beta$  strands. The yellow dotted line represents the N-terminal portion of the polypeptide, which lacks defined structure. (b) A proposed model of the abnormal, infectious ( $\text{PrP}^{Sc}$ ) prion protein, which consists largely of  $\beta$ -sheet. The actual tertiary structure of the prion protein has not been determined. The two molecules shown in this figure are formed by polypeptide chains that can be identical in amino acid sequence but fold very differently. As a result of the differences in folding,  $\text{PrP}^C$  remains soluble, whereas  $\text{PrP}^{Sc}$  produces aggregates that kill the cell. (The two molecules shown in this figure are called conformers because they differ only in conformation.).

SOURCE: (a) From Adriana Verschoor, et al., *J. Cell Biol.*, Vol. 133 (cover 3), 1996; by copyright permission of Rockefeller University Press; (b) Reprinted from S.B. Prusiner, *Trends Biochem. Sci.* 21:483, 1996 Copyright 1996, with permission from Elsevier.

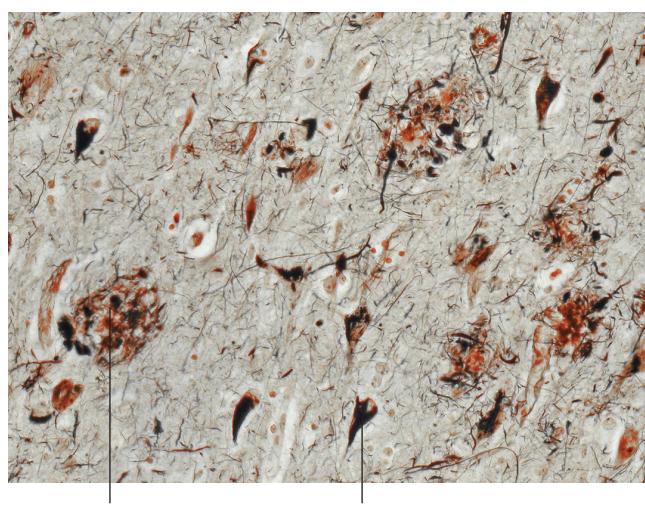
continued

agent? According to the prevailing hypothesis, an abnormal prion molecule ( $\text{PrP}^{\text{Sc}}$ ) can bind to a normal protein molecule ( $\text{PrP}^{\text{C}}$ ) and cause the normal protein to fold into the abnormal form. This conversion can be shown to occur in the test tube: Addition of  $\text{PrP}^{\text{Sc}}$  to a preparation of  $\text{PrP}^{\text{C}}$  can convert the  $\text{PrP}^{\text{C}}$  molecules into the  $\text{PrP}^{\text{Sc}}$  conformation. According to this hypothesis, the appearance of the abnormal protein in the body—whether as a result of a rare misfolding event in the case of sporadic disease or by exposure to contaminated beef—starts a chain reaction in which normal protein molecules in the cells are gradually converted to the misshapen prion form as they are recruited into growing insoluble fibrils. The precise mechanism by which prions lead to neurodegeneration remains unclear.

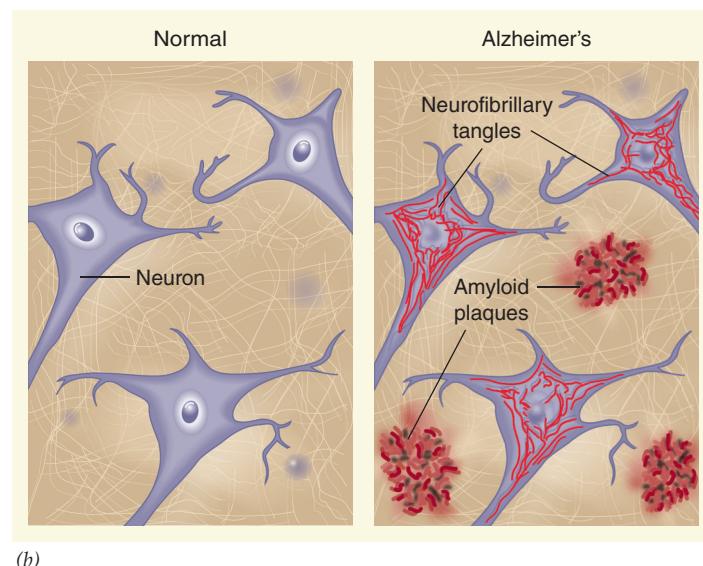
CJD is a rare disease caused by a protein with unique infective properties. Alzheimer's disease (AD), on the other hand, is a common disorder that strikes as many as 10 percent of individuals who are at least 65 years of age and perhaps 40 percent of individuals who are 80 years or older. Persons with AD exhibit memory loss, confusion, and a loss of reasoning ability. CJD and AD share a number of important features. Both are fatal neurodegenerative diseases that can occur in either an inherited or sporadic form. Like CJD, the brain of a person with Alzheimer's disease contains fibrillar deposits of an insoluble material referred to as *amyloid* (FIGURE 2).<sup>2</sup> In both diseases, the fibrillar deposits result from the self-association of a polypeptide composed predominantly of  $\beta$  sheet. There are also many basic differences between the two diseases: the proteins that form the disease-causing aggregates are unrelated, the parts of the brain that are affected are distinct, and the protein responsible for AD is not considered to be an infectious agent (i.e., it does not spread in a contagious pattern from one person to another, although it may spread from cell to cell within the brain).

<sup>2</sup>It should be noted that the term *amyloid* is not restricted to the abnormal protein found in AD. Many different proteins are capable of assuming an abnormal conformation that is rich in  $\beta$ -sheet, which causes the protein monomers to aggregate into characteristic amyloid fibrils that bind certain dyes. Amyloid fibrils are defined by their molecular structure in which the  $\beta$ -strands are oriented perpendicular to the long axis of the fibrils. The  $\text{PrP}^{\text{Sc}}$ -forming fibrils of prion diseases are also described as amyloid.

Over the past two decades, research on AD has been dominated by the *amyloid hypothesis*, which contends that the disease is caused by the production of a molecule, called the *amyloid  $\beta$ -peptide* ( $\text{A}\beta$ ).  $\text{A}\beta$  is originally part of a larger protein called the *amyloid precursor protein* (APP), which spans the nerve cell membrane. The  $\text{A}\beta$  peptide is released from the APP molecule following cleavage by two specific enzymes,  $\beta$ -secretase and  $\gamma$ -secretase (FIGURE 3). The length of the  $\text{A}\beta$  peptide is somewhat variable. The predominant species has a length of 40 amino acids (designated as  $\text{A}\beta 40$ ), but a minor species with two additional hydrophobic residues (designated as  $\text{A}\beta 42$ ) is also produced. Both of these peptides can exist in a soluble form that consists predominantly of  $\alpha$  helices, but  $\text{A}\beta 42$  has a tendency to spontaneously refold into a very different conformation that contains considerable  $\beta$  sheet. It is the misfolded  $\text{A}\beta 42$  version of the molecule that has the greatest potential to cause damage to the brain.  $\text{A}\beta 42$  tends to self-associate to form small complexes (oligomers) as well as large aggregates that are visible as fibrils in the electron microscope. These amyloid fibrils are deposited outside of the nerve cells in the form of extracellular *amyloid plaques* (Figure 2). Although the issue is far from settled, a body of evidence suggests that it is the soluble oligomers that are most toxic to nerve cells, rather than the insoluble aggregates. Cultured nerve cells, for example, are much more likely to be damaged by the presence of soluble intracellular  $\text{A}\beta$  oligomers than by either  $\text{A}\beta$  monomers or extracellular fibrillar aggregates. In the brain, the  $\text{A}\beta$  oligomers appear to attack the synapses that connect one nerve cell to another and eventually lead to the death of the nerve cells. Persons who suffer from an inherited form of AD carry a mutation that leads to an increased production of the  $\text{A}\beta 42$  peptide. Overproduction of  $\text{A}\beta 42$  can be caused by possession of extra copies (duplications) of the APP gene, by mutations in the APP gene, or by mutations in genes (*PSEN1*, *PSEN2*) that encode subunits of  $\gamma$ -secretase. Individuals with such mutations exhibit symptoms of the disease at an early age, typically in their 50s. The fact that all mutations associated with these inherited, early-onset forms of AD lead to increased production of  $\text{A}\beta 42$  is the strongest argument favoring amyloid formation as the underlying basis of the disease. The strongest argument against the



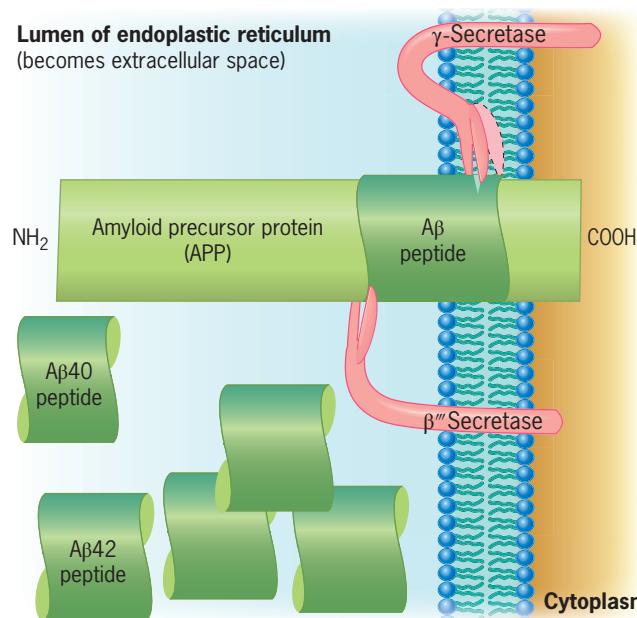
(a)



(b)

**FIGURE 2 Alzheimer's disease.** (a) The defining characteristics of brain tissue from a person who died of Alzheimer's disease. (b) Amyloid plaques containing aggregates of the  $\text{A}\beta$  peptide appear extracellularly (between nerve cells), whereas neurofibrillary tangles (NFTs) appear within the cells themselves. NFTs, which are discussed at the end of the Human Perspective, are composed of misfolded tangles of a protein called tau that is involved in maintaining the microtubule organization of the nerve cell. Both the plaques and tangles have been implicated as a cause of the disease.

SOURCE: (a) © Thomas Deerinck, NCMIR/Photo Researchers, Inc.; (b) © American Health Assistance Foundation.



**FIGURE 3 Formation of the A<sub>β</sub> peptide.** The A<sub>β</sub> peptide is carved from the amyloid precursor protein (APP) as the result of cleavage by two enzymes,  $\beta$ -secretase (also called BACE1) and  $\gamma$ -secretase. It is interesting that APP and the two secretases are all proteins that span the membrane. Cleavage of APP occurs inside the cell (probably in the endoplasmic reticulum), and the A<sub>β</sub> product is ultimately secreted into the space outside of the cell. The  $\gamma$ -secretase can cut at either of two sites in the APP molecule, producing either A<sub>β</sub>40 or A<sub>β</sub>42 peptides, the latter of which is primarily responsible for production of the amyloid plaques seen in Figure 2.  $\gamma$ -Secretase is a multisubunit enzyme that cleaves its substrate at a site within the membrane.

amyloid hypothesis is the weak correlation that can exist between the number and size of amyloid plaques in the brain and the severity of the disease. Elderly persons who show little or no sign of memory loss or dementia can have relatively high levels of amyloid deposits in their brain and those with severe disease can have little or no amyloid deposition.

All of the drugs currently on the market for the treatment of AD are aimed only at management of symptoms; none has any effect on stopping disease progression. With the amyloid hypothesis as the guiding influence, researchers have followed three basic strategies in the pursuit of new drugs for the prevention and/or reversal of mental decline associated with AD. These strategies are (1) to prevent the formation of the A<sub>β</sub>42 peptide in the first place; (2) to remove the A<sub>β</sub>42 peptide (or the amyloid deposits it produces) once it has been formed; and (3) to prevent the interaction between A<sub>β</sub> molecules, thereby preventing the formation of both oligomers and fibrillar aggregates. Before examining each of these strategies, we can consider how investigators can determine what type of drugs might be successful in the prevention or treatment of AD.

One of the best approaches to the development of treatments for human diseases is to find laboratory animals, particularly mice, that develop similar diseases, and use these animals to test the effectiveness of potential therapies. Animals that exhibit a disease that mimics a human disease are termed *animal models*. For whatever reason, the brains of aging mice show no evidence of the amyloid deposits found in humans, and, up until 1995, there was no animal model for AD. Then, in that year, researchers found that they could create a strain of mice that developed amyloid plaques in their brain and performed poorly at tasks that required

memory. They created this strain by genetically engineering the mice to carry a mutant human APP gene, one responsible for causing AD in families. These genetically engineered (*transgenic*) mice have proven invaluable for testing potential therapies for AD. The greatest excitement in the field of AD therapeutics has centered on the second strategy mentioned above, and we can use these investigations to illustrate some of the steps required in the development of a new drug.

In 1999, Dale Schenk and his colleagues at Elan Pharmaceuticals published an extraordinary finding. They had discovered that the formation of amyloid plaques in mice carrying the mutant human APP gene could be blocked by repeatedly injecting the animals with the very same substance that causes the problem, the aggregated A<sub>β</sub>42 peptide. In effect, the researchers had immunized (i.e., vaccinated) the mice against the disease. When young (6-week-old) mice were immunized with A<sub>β</sub>42, they failed to develop the amyloid brain deposits as they grew older. When older (13-month-old) mice whose brains already contained extensive amyloid deposits were immunized with the A<sub>β</sub>42, a significant fraction of the fibrillar deposits was cleared out of the nervous system. Even more importantly, the immunized mice performed better than their nonimmunized littermates on memory-based tests.

The dramatic success of these experiments on mice, combined with the fact that the animals showed no ill effects from the immunization procedure, led government regulators to quickly approve a Phase I clinical trial of the A<sub>β</sub>42 vaccine. A Phase I clinical trial is the first step in testing a new drug or procedure in humans and usually comes after years of preclinical testing on cultured cells and animal models. Phase I tests are carried out on a small number of subjects and are designed to monitor the safety of the therapy and the optimal dose of the drug rather than its effectiveness against the disease.

None of the subjects in two separate Phase I trials of the A<sub>β</sub> vaccine showed any ill-effects from the injection of the amyloid peptide. As a result, the investigators were allowed to proceed to a Phase II clinical trial, which involves a larger group of subjects and is designed to obtain a measure of the effectiveness of the procedure (or drug). This particular Phase II trial was carried out as a randomized, double-blind, placebo-controlled study. In this type of study:

1. The patients are *randomly* divided into two groups that are treated similarly except that one group is given the curative factor (protein, antibodies, drugs, etc.) being investigated and the other group is given a *placebo* (an inactive substance that has no therapeutic value).
2. The study is *double-blinded*, which means that neither the researchers nor patients know who is receiving treatment and who is receiving the placebo.

The Phase II trial for the A<sub>β</sub> vaccine began in 2001 and enrolled more than 350 individuals in the United States and Europe who had been diagnosed with mild to moderate AD. After receiving two injections of synthetic  $\beta$ -amyloid (or a placebo), 6 percent of the subjects experienced a potentially life-threatening inflammation of the brain. The inflammation in most of these patients was successfully treated with steroids, but the trial was discontinued. More recently, vaccination trials have been conducted using fragments of A<sub>β</sub> protein that do not induce inflammation but the results of these trials are not yet known.

Once it had become apparent that vaccination of patients with A<sub>β</sub>42 had inherent risks, it was decided to pursue a safer form of immunotherapy, which is to administer antibodies directed against A<sub>β</sub> that have been produced outside the body. This type of approach is known as *passive immunization* because the person does not produce the therapeutic antibodies. Passive

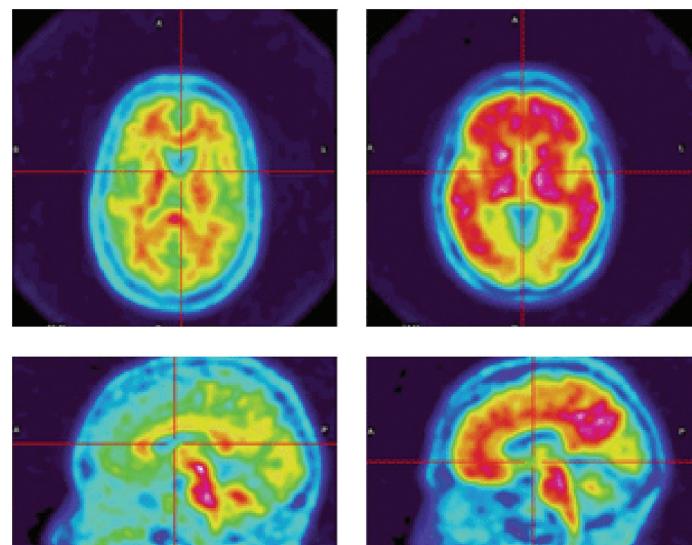
*continued*

immunization with an anti-A $\beta$ 42 antibody (called bapineuzumab) had already proven capable of restoring memory function in transgenic mice and was quickly shown to be safe, and apparently effective, in Phase I and II clinical trials. The last step before government approval is a Phase III trial, which typically employs large numbers of subjects (a thousand or more at several research centers) and compares the effectiveness of the new treatment against standard approaches. The first results of the Phase III trials on bapineuzumab were reported in 2008 and were disappointing: There was little or no evidence that the antibody provided benefits in preventing the progression of the disease. Another antibody known as Solanezumab, which recognizes a different part of the A $\beta$  protein, has been the subject of ongoing Phase III trials and so far the results indicate effectiveness in restoring cognitive function in patients with mild Alzheimer's symptoms, but no effect in patients with more severe symptoms. Given the impact of AD on human health and the large amount of money that could be earned from this type of drug, pharmaceutical companies are willing to take the risk that one of these immunologic strategies will exhibit some therapeutic value.

Meanwhile, a comprehensive analysis of some of the patients who had been vaccinated with A $\beta$ 42 in the original immunization trial from 2001 was also reported in 2008. Analysis of this patient group indicated that the A $\beta$ 42 vaccination had had no effect on preventing disease progression. It was particularly striking that in several of these patients who had died of severe dementia, there were virtually no amyloid plaques left in their brains. This finding strongly suggests that removal of amyloid deposits in a patient already suffering the symptoms of mild-to-moderate dementia does not stop disease progression. These results can be interpreted in more than one way. One interpretation is that the amyloid deposits are not the cause of the symptoms of dementia. An alternate interpretation is that irreversible toxic effects of the deposits had already occurred by the time immunization had begun and it was too late to reverse the disease course using treatments that remove existing amyloid deposits. It is important to note, in this regard, that the formation of amyloid deposits in the brain begins 10 or more years before any clinical symptoms of AD are reported. It is possible that, if these treatments had started earlier, the symptoms of the disease might never have appeared. The first clinical trial of this type (known as a "preventive trial") was begun in 2012 as several hundred individuals who would normally be destined to develop early-onset AD (due to mutations in the *PSEN1* gene) were treated with an anti-A $\beta$  antibody in the hopes of blocking the future buildup of amyloid and preventing the disease. Additional trials using larger numbers of patients who do not carry any known AD predisposing mutations have also been started, using the new generation antibody Solanezumab. Preventive trials which test the ability of a treatment to slow the onset of a disease that normally can take decades to manifest itself naturally take a long time to carry out, and so the results are not yet known.

Clearly, preventive treatments make the most sense for patients who might be on the verge of developing disease. Recent advances in brain-imaging procedures now allow clinicians to observe amyloid deposits in the brains of individuals long before any symptoms of AD have developed (**FIGURE 4**). Based on these studies, it may be possible to begin preventive treatments in persons who are at very high risk of developing AD before they develop symptoms.

Drugs have also been developed that follow the other two strategies outlined above. Alzhemed and scyllo-inositol are two small molecules that bind to A $\beta$  peptides and block molecular aggregation and fibril formation. Clinical trials have failed



**FIGURE 4** A neuroimaging technique that reveals the presence of amyloid in the brain. These PET (positron emission tomography) scans show the brains of two individuals that have ingested a radioactive compound, called flutemetamol, that binds to amyloid deposits and appears red in the image. The left panels show a healthy brain and the right panels show a brain from a patient with AD, revealing extensive amyloid build-up. Amyloid deposits in the brain can be detected with this technique in persons who show no evidence of cognitive dysfunction. Such symptom-free individuals are presumed to be at high risk of going on to develop AD. Those who lack such deposits can be considered at very low risk of the disease in the near future.

SOURCE: Courtesy of Ken Garber, *Nature Biotechnology* 30, 575 (2012) doi:10.1038/nbt0712-575 Published online 10 July 2012.

to demonstrate that either drug is effective in stopping disease progression in patients with mild to moderate AD. The third strategy outlined above is to stop production of A $\beta$  peptides. This can be accomplished by inhibiting either  $\beta$ - or  $\gamma$ -secretase, because both enzymes are required in the pathway that cleaves the APP precursor to release the internal peptide (Figure 3). Pharmaceutical companies have had great difficulty developing a  $\beta$ -secretase inhibitor that is both potent and small enough to enter the brain from the bloodstream, however one such inhibitor, known as MK-8931, is currently in Phase III clinical trials. A number of potent  $\gamma$ -secretase inhibitors have been developed that block the production of all A $\beta$  peptides, both in cultured nerve cells and in transgenic AD mice. But there is a biological problem that has to be overcome with this class of inhibitors. In addition to cleaving APP,  $\gamma$ -secretase activity is also required in a key signaling pathway involving a protein called Notch. Two of the most promising  $\gamma$ -secretase inhibitors, flurizan and semagacestat, have both failed to show any benefit in stopping AD progression. In addition to its lack of efficacy, semagacestat caused adverse side effects that were probably a result of blockade of the Notch pathway. The goal of drug designers is to develop a compound (e.g., begacestat) that blocks APP cleavage but does not interfere with cleavage of Notch.

Taken collectively, the apparent failure of all of these drugs, aimed at various steps in the formation of A $\beta$ -containing aggregates and amyloid deposition, has left the field of AD therapeutics without a clear plan for the future. Some pharmaceutical companies are continuing to develop new drugs aimed at blocking the

formation of amyloid aggregates, whereas others are moving in different directions. These findings also raise a more basic question: Is the A $\beta$  peptide even part of the underlying mechanism that leads to AD? It hasn't been mentioned, but A $\beta$  is not the only misfolded protein found in the brains of persons with AD. Another protein called tau, which functions as part of a nerve cell's cytoskeleton (Section 9.2), can develop into bundles of tangled cellular filaments called neurofibrillary tangles (or NFTs) (Figure 2) that interfere with the movement of substances down the length of the nerve cell. NFTs form when the tau molecules in nerve cells become excessively phosphorylated. Mutations in the gene that encodes tau have been found to cause a rare form of dementia (called FTD), which is characterized by the formation of NFTs. Thus, NFTs have been linked to dementia, but they have been largely ignored as a causative factor in AD pathogenesis, due primarily to the fact that the transgenic AD mouse models discussed above do not develop NFTs. If one extrapolates the results of these mouse studies to humans, they suggest that NFTs are not required for the cognitive decline that occurs in patients with AD. At the same time, however, autopsies of the brains of humans who died of AD suggest that NFT burden correlates much better with cognitive dysfunction and neuronal loss than does the concentration of amyloid plaques. Given that mutations in genes in the A $\beta$  pathway are clearly a cause of AD, and yet it is the NFT burden that correlates with cognitive decline, it would appear that both A $\beta$  and NFTs must be involved in AD etiology. Many researchers believe that A $\beta$  deposition somehow leads

to NFT formation, an idea known as the "amyloid cascade hypothesis," but the mechanism by which this might occur remains unknown. In fact, one of the most promising new drugs at the time of this writing is one that acts on NFTs rather than  $\beta$ -amyloid. In this case, the drug methylthioninium chloride, which dissolves NFTs, was tested on a group of more than 300 patients with mild to moderate AD in a Phase II trial. The drug was found to reduce mental decline over a period of one year by an average of 81 percent compared to patients receiving a placebo. A modified version of the drug, known as leuco methylthioninium, is now being tested in larger Phase III studies for both AD and FTD, but results are not yet available. Other compounds that inhibit one of the enzymes (GSK-3) that adds phosphate groups to the tau protein are also being investigated as AD therapeutics. Clinical trials of one GSK-3 inhibitor, valproate, have been stopped due to adverse effects. (The results of studies on these and other treatments can be examined at [www.alzforum.org/dis/tre/drc](http://www.alzforum.org/dis/tre/drc))

It is evident from this discussion that a great deal of work on AD has been based on transgenic mice carrying human AD genes. These animals have served as the primary preclinical subjects for testing AD drugs, and they have been used extensively in basic research that aims to understand the disease mechanisms responsible for the development of AD. But many questions have been raised as to how accurately these animal models mimic the disease in humans, particularly the sporadic human cases in which affected individuals lack the mutant genes that cause the animals to develop the corresponding disorder.

## 2.14 EXPERIMENTAL PATHWAYS

### Chaperones—Helping Proteins Reach Their Proper Folded State

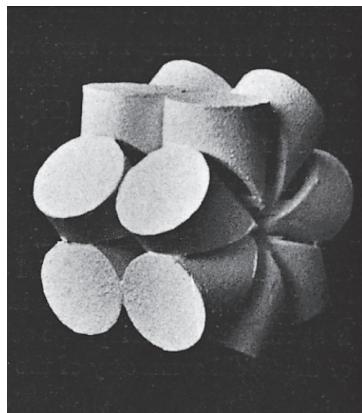
In 1962, F. M. Ritossa, an Italian biologist studying the development of the fruit-fly *Drosophila*, reported a curious finding.<sup>1</sup> When the temperature at which fruit-fly larvae were developing was raised from the normal 25°C to 32°C, a number of new sites on the giant chromosomes of the larval cells became activated. As we will see in Chapter 10, the giant chromosomes of these insect larvae provide a visual exhibit of gene expression (see Figure 10.8). The results suggested that increased temperature induced the expression of new genes, a finding that was confirmed a decade later with the characterization of several proteins that appeared in larvae following temperature elevation.<sup>2</sup> It was soon found that this response, called the **heat-shock response**, was not confined to fruit flies, but can be initiated in many different cells from virtually every type of organism—from bacteria to plants and mammals. Closer examination revealed that the proteins produced during the response were found not only in heat-shocked cells, but also at lower concentration in cells under normal conditions. What is the function of these so-called **heat-shock proteins (hsp)**? The answer to this question was gradually revealed by a series of seemingly unrelated studies.

We will see on page 79 that some complex, multisubunit structures, such as a bacterial ribosome or a tobacco mosaic virus particle, can self-assemble from purified subunits. It was demonstrated in the 1960s that the proteins that make up bacteriophage particles (see Figure 1.22c) also possess a remarkable ability to

self-assemble, but they are generally unable to form a complete, functional virus particle by themselves *in vitro*. Experiments on phage assembly in bacterial cells confirmed that phages require bacterial help. It was shown in 1973, for example, that a certain mutant strain of bacteria, called *GroE*, could not support the assembly of normal phages. Depending on the type of phage, the head or the tail of the phage particle was assembled incorrectly.<sup>3,4</sup> These studies suggested that a protein encoded by the bacterial chromosome participated in the assembly of viruses, even though this host protein was not a component of the final virus particles. Because it obviously did not evolve as an aid for virus assembly, the bacterial protein required for phage assembly had to play some role in the cell's normal activities, but the precise role remained obscure. Subsequent studies revealed that the *GroE* site on the bacterial chromosome actually contains two separate genes, *GroEL* and *GroES*, that encode two separate proteins *GroEL* and *GroES*. Under the electron microscope, the purified *GroEL* protein appeared as a cylindrical assembly consisting of two disks. Each disk was composed of seven subunits arranged symmetrically around the central axis (**FIGURE 1**).<sup>5,6</sup>

Several years later, a study on pea plants hinted at the existence of a similar assembly-promoting protein in the chloroplasts of plants.<sup>7</sup> Rubisco is a large protein in chloroplasts that catalyzes the reaction in which CO<sub>2</sub> molecules taken up from the atmosphere are covalently linked to organic molecules during

*continued*



**FIGURE 1** An early model of the GroEL complex built according to data from electron microscopy and molecular-weight determination. The complex is seen to consist of two disks, each composed of seven identical subunits arranged symmetrically around a central axis. Subsequent studies showed the complex contains two internal chambers.

SOURCE: From T. Hohn et al., *J. Mol. Biol.* 129:371, © 1979, with permission of Elsevier.

photosynthesis (Section 6.9). Rubisco comprises 16 subunits: 8 small subunits (molecular mass of 14,000 daltons) and 8 large subunits (55,000 daltons). It was found that large Rubisco subunits, synthesized inside the chloroplast, are not present in an independent state, but are associated with a huge protein assembly consisting of identical subunits of 60,000 daltons (60 kDa) molecular mass. In their paper, the researchers considered the possibility that the complex formed by the large Rubisco subunits and the 60-kDa polypeptides was an intermediate in the assembly of a complete Rubisco molecule.

A separate line of investigation on mammalian cells also revealed the existence of proteins that appeared to assist the assembly of multisubunit proteins. Like Rubisco, antibody molecules consist of a complex of two different types of subunits, smaller light chains and larger heavy chains. Just as the large subunits of Rubisco become associated with another protein not found in the final complex, so too do the heavy chains of an antibody complex.<sup>8</sup> This protein, which associates with newly synthesized heavy chains, but not with heavy chains that are already bound to light chains, was named *binding protein*, or BiP. BiP was subsequently found to have a molecular mass of 70,000 daltons (70 kDa).

To this point, we have been discussing two lines of investigation: one concerned with the heat-shock response and the other with proteins that promote protein assembly. These two fields came together in 1986, when it was shown that one of the proteins that figured most prominently in the heat-shock response, a protein that had been named *heat-shock protein 70* (*hsp70*) because of its molecular mass, was identical to BiP, the protein implicated in the assembly of antibody molecules.<sup>9</sup>

Even before the discovery of the heat-shock response, the structure of proteins was known to be sensitive to temperature, with a small rise in temperature causing these delicate molecules to begin to unfold. Unfolding exposes hydrophobic residues that were previously buried in the protein's core. Just as fat molecules in a bowl of soup are pushed together into droplets, so too are proteins with hydrophobic patches on their surface. Consequently, when a cell is heat shocked, soluble proteins become denatured and form aggregates. A report in 1985 demonstrated that, following temperature elevation, newly synthesized hsp70 molecules enter cell nuclei and bind to aggregates of nuclear proteins, where they act like molecular crowbars to promote disaggregation.<sup>10</sup> Because of their role in assisting the assembly of proteins

by preventing undesirable interactions, hsp70 and related molecules were named **molecular chaperones**.<sup>11</sup>

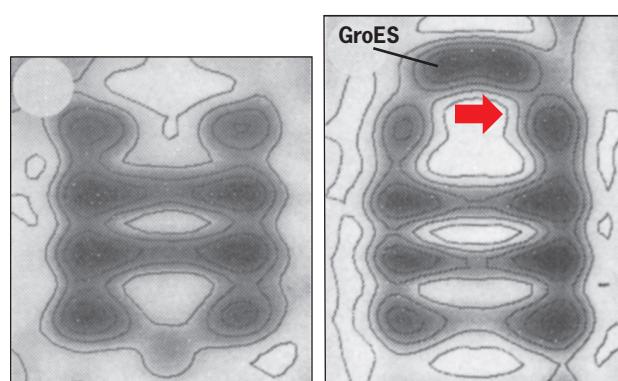
It was soon demonstrated that the bacterial heat-shock protein GroEL and the Rubisco assembly protein in plants were homologous proteins. In fact, the two proteins share the same amino acids at nearly half of the more than 500 residues in their respective molecules.<sup>12</sup> The fact that the two proteins—both members of the *Hsp60 chaperone family*—have retained so many of the same amino acids reflects their similar and essential function in the two types of cells. But what was that essential function? At this point it was thought that their primary function was to mediate the assembly of multisubunit complexes, such as Rubisco. This view was changed in 1989 by experiments studying molecular chaperones in mitochondria by Arthur Horwich of Yale University and F.-Ulrich Hartl, Walter Neupert, and their colleagues at the University of Munich.<sup>13,14</sup> It was known that newly made mitochondrial proteins produced in the cytosol had to cross the outer mitochondrial membranes in an unfolded, extended, monomeric form. A mutant was found that altered the activity of another member of the Hsp60 chaperone family that resided inside mitochondria. In cells containing this mutant chaperone, proteins that were transported into mitochondria failed to fold into their active forms. Even proteins that consisted of a single polypeptide chain failed to fold into their native conformation. This finding changed the perception of chaperone function from a notion that they assist assembly of already folded subunits into larger complexes, to our current understanding that they assist polypeptide chain folding within the crowded confines of the cell.

The results of these and other studies indicated the presence in cells of at least two major families of molecular chaperones: the Hsp70 chaperones, such as BiP, and the Hsp60 chaperones (which are also called *chaperonins*), such as Hsp60, GroEL, and the Rubisco assembly protein. We will focus on the Hsp60 chaperonins, such as GroEL, which are best understood.

As first revealed in 1979, GroEL is a huge molecular complex of 14 polypeptide subunits arranged in two stacked rings resembling a double doughnut.<sup>5,6</sup> Fifteen years after these first electron micrographs were taken, the three-dimensional structure of the GroEL complex was determined by X-ray crystallography.<sup>15</sup> The study revealed the presence of a central cavity within the GroEL cylinder. Subsequent studies demonstrated that this cavity was divided into two separate chambers. Each chamber was situated within the center of one of the rings of the GroEL complex and was large enough to enclose a polypeptide undergoing folding.

Electron microscopic studies also provided information about the structure and function of a second protein, GroES, which acts in conjunction with GroEL. Like GroEL, GroES is a ring-like protein with seven subunits arrayed symmetrically around a central axis. GroES, however, consists of only one ring, and its subunits are much smaller (10,000 daltons) than those of GroEL (60,000 daltons). GroES is seen as a cap or dome that fits on top of either end of a GroEL cylinder (FIGURE 2). The attachment of GroES to one end of GroEL causes a dramatic conformational change in the GroEL protein that markedly increases the volume of the enclosed chamber at that end of the complex.<sup>16</sup>

The importance of this conformational change has been revealed in remarkable detail by X-ray crystallographic studies in the laboratories of Arthur Horwich and Paul Sigler at Yale University.<sup>17</sup> As shown in FIGURE 3, the binding of the GroES cap is accompanied by a 60° rotation of the apical (red) domain of the subunits that make up the GroEL ring at that end of the GroEL cylinder. The attachment of GroES does more than trigger a conformational change that enlarges the GroEL chamber. Before attachment of GroES, the inner wall of the GroEL chamber has exposed hydrophobic residues that give the lining a hydrophobic character.



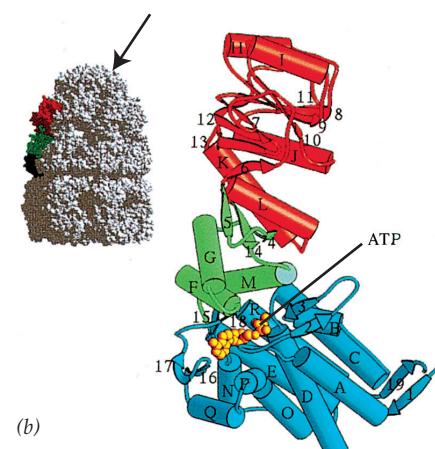
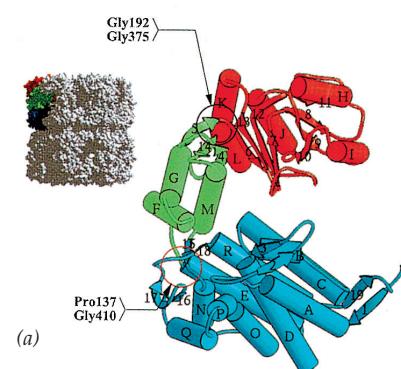
**FIGURE 2** Reconstructions of GroEL based on high-resolution electron micrographs taken of specimens that had been frozen in liquid ethane and examined at  $-170^{\circ}\text{C}$ . The image on the left shows the GroEL complex, and that on the right shows the GroEL complex with GroES, which appears as a dome on one end of the cylinder. It is evident that the binding of the GroES is accompanied by a marked change in conformation of the apical end of the proteins that make up the top GroEL ring (arrow), which results in a marked enlargement of the upper chamber.

SOURCE: From S. Chen, et al., Courtesy of Helen R. Saibil, *Nature* 371:263, © 1994, reprinted by permission from Macmillan Publishers Limited.

Nonnative polypeptides also have exposed hydrophobic residues that become buried in the interior of the native polypeptide. Because hydrophobic surfaces tend to interact, the hydrophobic lining of the GroEL cavity binds to the surface of nonnative polypeptides. Binding of GroES to GroEL buries the hydrophobic residues of the GroEL wall and exposes a number of polar residues, thereby changing the character of the chamber wall. As a result of this change, a nonnative polypeptide that had been bound to the GroEL wall by hydrophobic interactions is displaced into the space within the chamber. Once freed from its attachment to the chamber wall, the polypeptide is given the opportunity to continue its folding in a protected environment. After about 15 seconds, the GroES cap dissociates from the GroEL ring, and the polypeptide is ejected from the chamber. If the polypeptide has not reached its native conformation by the time it is ejected, it can rebind to the same or another GroEL, and the process is repeated. A model depicting some of the steps thought to occur during GroEL-GroES-assisted folding is shown in **FIGURE 4**.

Approximately 250 of the roughly 2400 proteins present in the cytosol of an *E. coli* cell normally interact with GroEL.<sup>18</sup> How is it possible for a chaperone to bind so many different polypeptides? The GroEL binding site consists of a hydrophobic surface formed largely by two  $\alpha$  helices of the apical domain that is capable of binding virtually any sequence of hydrophobic residues that might be accessible in a partially folded or misfolded polypeptide.<sup>19</sup> A comparison of the crystal structure of the unbound GroEL molecule with that of GroEL bound to several different peptides revealed that the binding site on the apical domain of a GroEL subunit can locally adjust its positioning when bound to different partners. This finding indicates that the binding site has structural flexibility that allows it to adjust its shape to fit the shape of the particular polypeptide with which it has to interact.

A number of studies have also suggested that GroEL does more than simply provide a passive chamber in which proteins can fold without outside interference. In one study, site-directed mutagenesis was utilized to modify a key residue, Tyr71 of GroES, whose side chain hangs from the ceiling of the folding chamber.<sup>20</sup> Because of its aromatic ring, tyrosine is a modestly hydrophobic residue (Figure 2.26). When Tyr71 was replaced by a positively or negatively charged amino acid, the resulting GroEL-GroES

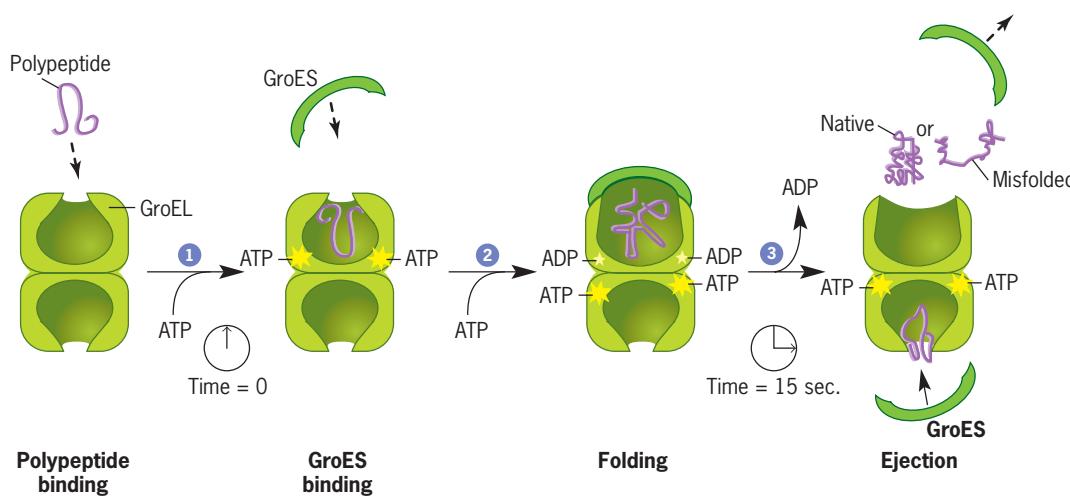


**FIGURE 3** Conformational change in GroEL. (a) The model on the left shows a surface view of the two rings that make up the GroEL chaperonin. The drawing on the right shows the tertiary structure of one of the subunits of the top GroEL ring. The polypeptide chain can be seen to fold into three domains. (b) When a GroES ring (arrow) binds to the GroEL cylinder, the apical domain of each GroEL subunit of the adjacent ring undergoes a dramatic rotation of approximately  $60^{\circ}$  with the intermediate domain (shown in green) acting like a hinge. The effect of this shift in parts of the polypeptide is a marked elevation of the GroEL wall and enlargement of the enclosed chamber.

SOURCE: From Z. Xu, A.L. Horwitz, and P. B. Sigler, *Nature* 388:744, 1997. © 1997, reprinted with permission Macmillan Publishers, Limited.

variant exhibited an increased ability to assist the folding of one specific foreign polypeptide, the green fluorescent protein (GFP). However, substitutions for Tyr71 that improved the ability of GroES-GroEL to increase GFP folding made the chaperonin less competent to help its natural substrates fold. Thus as the chaperonin became more and more specialized to interact with GFP, it lost its general ability to assist folding of proteins having an unrelated structure. This finding suggests that individual amino acids in the wall of the folding chamber may participate somehow in the folding reaction. Data from another study has suggested that binding of a nonnative protein to GroEL is followed by a forced unfolding of the substrate protein.<sup>21</sup> FRET (fluorescence resonance energy transfer) is a technique (discussed in Section 18.3) that allows researchers to determine the distance between different parts of a protein molecule at different times during a given process. In this study, investigators found that the protein undergoing folding, in this case Rubisco, bound to the apical domain of the GroEL ring in a relatively compact state. The compact nature of the bound protein was revealed by the close proximity to one another of the FRET tags, which were attached to amino

*continued*



**FIGURE 4** A schematic illustration of the proposed steps that occur during the GroEL-GroES-assisted folding of a polypeptide. The GroEL is seen to consist of two chambers that have equivalent structures and functions and that alternate in activity. Each chamber is located within one of the two rings that make up the GroEL complex. The nonnative polypeptide enters one of the chambers (step 1) and binds to hydrophobic sites on the chamber wall. Binding of the GroES cap produces a conformational change in the wall of the top chamber, causing the enlargement of the chamber and release of the nonnative polypeptide from the wall into the encapsulated space (step 2). After about 15 seconds have elapsed, the GroES dissociates from the complex and the polypeptide is ejected from the chamber (step 3). If the polypeptide has achieved its native conformation, as has the molecule on the left, the folding process is complete. If, however, the polypeptide is only partially folded, or is misfolded, it will rebind the GroEL chamber for another round of folding. (Note: As indicated, the mechanism of GroEL action is driven by the binding and hydrolysis of ATP, an energy-rich molecule whose function is discussed at length in the following chapter.)

SOURCE: A. L. Horwitz, et al., *Proc. Nat'l. Acad. Sci. U.S.A.* 96:11037, 1999.

acids located at opposite ends of the Rubisco chain. Then, during the conformational change that enlarges the volume of the GroEL cavity (Figure 3), the bound Rubisco protein was forcibly unfolded, as evidenced by the increased distance between the two tagged ends of the molecule. This study suggests that the Rubisco polypeptide is taken completely back to the unfolded state, where it is given the opportunity to refold from scratch. This action should help prevent the nonnative protein from becoming trapped permanently in a misfolded state. In other words, each individual visit to a GroEL-GroES chamber provides an all-or-none attempt to reach the native state, rather than just one stage in a series of steps in which the protein moves closer to the native state with each round of folding. Recent reviews of molecular chaperones can be found in References 22–23.

Keep in mind that molecular chaperones do not convey information for the folding process but instead prevent proteins from veering off their correct folding pathway and finding themselves in misfolded or aggregated states. Just as Anfinsen discovered decades ago, the three-dimensional structure of a protein is determined by its amino acid sequence.

## References

- Ritossa, F. 1962. A new puffing pattern induced by temperature shock and DNP in *Drosophila*. *Experientia* 18:571–573.
- Tissieres, A., Mitchell, H. K., & Tracy, U. M. 1974. Protein synthesis in salivary glands of *Drosophila melanogaster*: Relation to chromosomal puffs. *J. Mol. Biol.* 84:389–398.
- Sternberg, N. 1973. Properties of a mutant of *Escherichia coli* defective in bacteriophage lambda head formation (*groE*). *J. Mol. Biol.* 76:1–23.
- Georgopoulos, C. P. et al. 1973. Host participation in bacteriophage lambda head assembly. *J. Mol. Biol.* 76:45–60.
- Hohn, T. et al. 1979. Isolation and characterization of the host protein *groE* involved in bacteriophage lambda assembly. *J. Mol. Biol.* 129: 359–373.
- Hendrix, R. W. 1979. Purification and properties of *groE*, a host protein involved in bacteriophage assembly. *J. Mol. Biol.* 129:375–392.
- Barraclough, R. & Ellis, R. J. 1980. Protein synthesis in chloroplasts. IX. *Biochim. Biophys. Acta* 608:19–31.
- Haas, I. G. & Wabl, M. 1983. Immunoglobulin heavy chain binding protein. *Nature* 306:387–389.
- Munro, S. & Pelham, H.R.B. 1986. An Hsp70-like protein in the ER: Identity with the 78 kD glucose-regulated protein and immunoglobulin heavy chain binding protein. *Cell* 46:291–300.
- Lewis, M. J. & Pelham, H.R.B. 1985. Involvement of ATP in the nuclear and nucleolar functions of the 70kD heat-shock protein. *EMBO J.* 4: 3137–3143.
- Ellis, J. 1987. Proteins as molecular chaperones. *Nature* 328:378–379.
- Hemmingsen, S. M. et al. 1988. Homologous plant and bacterial proteins chaperone oligomeric protein assembly. *Nature* 333:330–334.
- Cheng, M. Y. et al. 1989. Mitochondrial heat-shock protein Hsp60 is essential for assembly of proteins imported into yeast mitochondria. *Nature* 337:620–625.
- Ostermann, J. et al. 1989. Protein folding in mitochondria requires complex formation with hsp60 and ATP hydrolysis. *Nature* 341: 125–130.
- Braig, K. et al. 1994. The crystal structure of the bacterial chaperonin GroEL at 2.8A. *Nature* 371:578–586.
- Chen, S. et al. 1994. Location of a folding protein and shape changes in GroEL-GroES complexes. *Nature* 371:261–264.
- Xu, Z., Horwitz, A. L., & Sigler, P. B. 1997. The crystal structure of the asymmetric GroEL-GroES-(ADP)<sub>7</sub> chaperonin complex. *Nature* 388: 741–750.
- Kerner, M. J. et al. 2005. Proteome-wide analysis of chaperonin-independent protein folding in *Escherichia coli*. *Cell* 122:209–220.
- Chen, L. & Sigler, P. 1999. The crystal structure of a GroEL/peptide complex: plasticity as a basis for substrate diversity. *Cell* 99:757–768.
- Wang, J. D. et al. 2002. Directed evolution of substrate-optimized GroEL/S chaperonins. *Cell* 111:1027–1039.
- Lin, Z. et al. 2008. GroEL stimulates protein folding through forced unfolding. *Nature Struct. Mol. Biol.* 15:303–311.
- Rothman, J. E. & Schekman, R. 2011. Molecular mechanisms of protein folding in the cell. *Cell* 146:851–854.
- Sabil, H. 2013. Chaperone machines for protein folding, unfolding, and disaggregation. *Nature Reviews Molecular Cell Biology* 14: 630–642.

## 2.15 Proteomics and Interactomics

With all of the attention on genome sequencing in recent years, it is easy to lose sight of the fact that genes are primarily information storage units, whereas proteins orchestrate cellular activities. Genome sequencing provides a kind of “parts list.” The human genome probably contains between 20,000 and 22,000 genes, each of which can potentially give rise to a variety of different proteins.<sup>5</sup> To date, only a fraction of these molecules have been characterized.

### Proteomics

The entire inventory of proteins that is produced by an organism, whether human or otherwise, is known as that organism's **proteome**. The term *proteome* is also applied to the inventory of all proteins that are present in a particular tissue, cell, or cellular organelle. Because of the sheer numbers of proteins that are currently being studied, investigators have sought to develop techniques that allow them to determine the properties or activities of a large number of proteins in a single experiment. A new term—**proteomics**—was coined to describe the expanding field of protein biochemistry. This term carries with it the concept that advanced technologies and high-speed computers are used to perform large-scale studies on diverse arrays of proteins. This is the same basic approach that has proven so successful over the past decade in the study of genomes. But the study of proteomics is inherently more difficult than the study of genomics because proteins are more difficult to work with than DNA. In physical terms, one gene is pretty much the same as all other genes, whereas each protein has unique chemical properties and handling requirements. In addition, small quantities of a particular DNA segment can be expanded greatly using readily available enzymes, whereas protein quantities cannot be increased. This is particularly troublesome when one considers that many of the proteins regulating important cellular processes are present in only a handful of copies per cell.

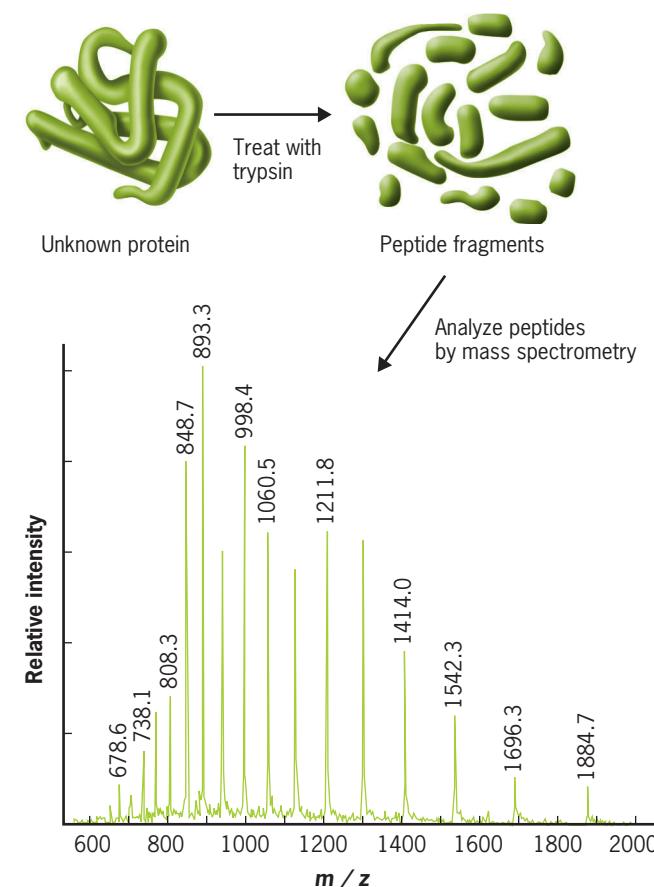
Traditionally, protein biochemists have sought to answer a number of questions about particular proteins. These include: What specific activity does the protein demonstrate in vitro, and how does this activity help a cell carry out a particular function such as cell locomotion or DNA replication? What is the protein's three-dimensional structure? When does the protein appear in the development of the organism and in which types of cells? Where in the cell is it localized? Is the protein modified after synthesis by the addition of chemical groups (e.g., phosphates or sugars) and, if so, how does this modify its activity? How much of the protein is present, and how long does it survive before being degraded? Does the level of the protein change during physiologic activities or as the result of disease? Which other proteins in the cell does it interact with? Biologists have been attempting to answer these questions for decades but, for the most part, they've been doing it one protein at a time. Proteomics researchers attempt to answer similar questions on a more comprehensive scale using large-scale (or *high-throughput*) techniques to catalog the vast array of proteins produced by a particular cell.

The key technology in proteomics is mass spectrometry, a method to probe the chemical structure of an unknown sample. As

<sup>5</sup>There are a number of ways that a single gene can give rise to more than one polypeptide. Two of the most prominent mechanisms, alternative splicing and posttranslational modification, are discussed in other sections of the text. It can also be noted that many proteins have more than one distinct function. Even myoglobin, which has long been studied as an oxygen-storage protein, has recently been shown to be involved in the conversion of nitric oxide (NO) to nitrate (NO<sub>3</sub><sup>-</sup>).

discussed in further detail in Section 18.15, mass spectrometry is a technique to determine the precise mass of a molecule or fragment of a molecule, which can then be used to identify that molecule. Suppose that we wanted to identify an unknown protein that we have in a test tube. The protein is first digested into peptides with the enzyme trypsin. When these peptides are introduced into a mass spectrometer, they are converted into gaseous ions and separated according to their mass/charge (m/z) ratio. The results are displayed as a series of peaks of known m/z ratio, such as that shown in **FIGURE 2.47**. The pattern of peaks constitutes a highly characteristic *peptide mass fingerprint* of that protein. But how can a protein be identified based on its peptide mass fingerprint?

The answer comes from genomics. Once a genome has been sequenced, the amino acid sequences of encoded proteins can be predicted. This list of “virtual proteins” can then be subjected to a theoretical trypsin digestion and the masses of the resulting virtual peptides calculated and entered into a database. Once this has been done, the actual peptide masses of a purified protein obtained by the mass spectrometer can be compared using a computer to the masses predicted by theoretical digests of all polypeptides encoded by the



**FIGURE 2.47** Identifying proteins by mass spectrometry. A protein is isolated from a cell or other sample and subjected to digestion by the enzyme trypsin. The peptide fragments are then introduced into a mass spectrometer where they are ionized and separated according to their mass/charge (m/z) ratio. The separated peptides appear as a pattern of peaks whose precise m/z ratio is indicated. A comparison of these ratios to those obtained by a theoretical digest of virtual proteins encoded by the genome allows researchers to identify the protein being studied. In this case, the MS spectrum is that of horse myoglobin lacking its heme group. SOURCE: Data reprinted from J. R. Yates, *Methods Enzymol.* 271:353, 1996. Copyright 1996, with permission from Elsevier.

**72** genome. In most cases, the protein that has been isolated and subjected to mass spectrometry can be directly identified based on this type of database search. Mass spectrometers are not restricted to handling one purified protein at a time, but are also capable of analyzing proteins present in complex mixtures (Section 18.15). By analyzing the total protein extracted from a cell by mass spectrometry, it is possible to determine a list of all the proteins present in the cell. This list is known as the proteome of the cell. Proteomic studies are particularly useful when two different samples are compared to see how the protein composition changes over time. For example, the proteome can be analyzed before and after the secretion of a hormone within the body, after taking a drug, or during a particular disease. The changes in the abundance of different proteins during such transitions can provide valuable clues about how cell function is changing in a given situation.

Proteomics is playing an increasingly important role in advancing the practice of medicine. It is thought that most human diseases leave telltale patterns (or *biomarkers*) among the thousands of proteins present in the blood or other bodily fluids. The simplest way to determine whether a particular biomarker protein characteristic of a disease is present in a blood or urine sample, and how much of that protein is present, is to measure the protein's interaction with a specific antibody. This is the basis, for example, of the PSA test used in routine screening of men for prostate cancer. PSA is a protein that is found in the blood of normal men, but is present at elevated levels in individuals with prostate cancer. PSA levels are determined by measuring the amount of protein in the blood that binds to anti-PSA antibodies. The most challenging hurdle in developing this type of diagnostic test is knowing what protein will act as the most reliable biomarker. This is where proteomics comes into play. Many efforts have been made to compare the proteins present in the blood of healthy individuals with those present in the blood of persons suffering from various diseases, especially cancer. Initially, the results of these biomarker searches were unreliable in that the findings of one research group could not be duplicated by the efforts of other groups. The primary difficulty stems from the fact that human blood serum is such a complex solution containing thousands of proteins that range in abundance over 9 or 10 orders of magnitude. But as proteomic and mass spectrometry technology improves, the information about complex samples becomes richer and richer. Already, proteomics has been used to reveal important biomarkers for disease. For example, the OVA1 blood test for ovarian cancer, which detects a collection of biomarkers using antibody-based tests, was invented using data from proteomic analysis of a large number of patient samples. The OVA1 test is mainly used for cancers that have already been detected, in order to provide more information about the tumor before surgery. It is hoped that, one day, it will be possible to use a single blood test to reveal the existence of early-stage heart, liver, or kidney disease that can be treated before it becomes a life-threatening condition.

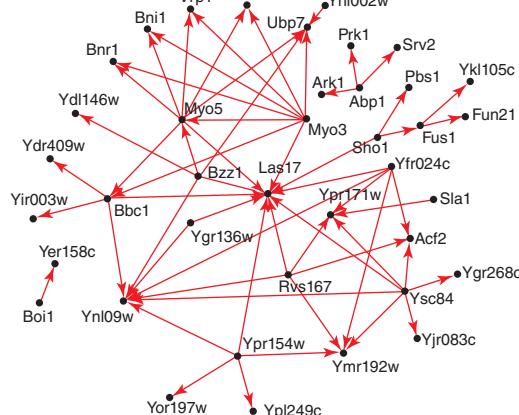
Protein separation and mass spectrometric techniques don't tell us anything about a protein's function. Researchers have been working to devise techniques that allow protein function to be determined on a large scale, rather than one protein at a time. Several new technologies have been developed to accomplish this mission; we will consider only one—the use of genome-wide RNA interference (RNAi) screens. RNAi is a cellular process by which cells produce small RNAs (called siRNAs) that bind to specific mRNAs and inhibit the translation of these mRNAs into proteins. This phenomenon and its use are discussed in detail in Sections 11.10 and 18.25. For the present purpose we will simply note that researchers can synthesize a

collection (library) of siRNAs that are capable of inhibiting the translation of virtually any mRNA that is produced by a genome. Each mRNA represents the expression of a specific gene encoding a particular protein; therefore, one can find out which proteins are involved in a particular cellular process by determining which siRNAs interfere with that process. RNAi can be used in large-scale screens of the whole genome and can also be used to test the function of individual genes or sets of genes. The combination of RNAi with proteomics is extremely powerful. As soon as a proteomic analysis suggests that the level of a particular protein changes during a process of interest, the expression of that protein can be blocked by an appropriate siRNA to ask whether its function is relevant for that process.

## Interactomics

Most investigators who study protein–protein interactions want to know whether one protein they are working with, call it protein X, interacts physically with another protein, call it protein Y. Several techniques can be used to test for an interaction between any particular pair of proteins, as discussed in Section 18.12. However, in recent years, a number of research teams have set out to study protein–protein interactions on a global scale. For example, one might want to know all of the interactions that occur among the 6000 or so proteins encoded by the genome of the budding yeast *Saccharomyces cerevisiae*. As is true for an increasing number of organisms, the entire genome of this yeast has been sequenced, and virtually every gene within the genome is available as an individual DNA segment that can be cloned and used as desired. Testing for potential interactions between all of these proteins requires methods that can be easily automated and performed by robots, but that are also reliable and sensitive enough to detect most of the true interactions without producing too many false interactions in the process. One such method that has been widely adopted is TAP-tag mass spectrometry. In this approach, DNA from a gene of interest is fused to DNA encoding a protein tag called a TAP tag that is easily purified using affinity chromatography methods (See the [Experimental Walkthrough Video: Purification of proteins using the TAP tag](#) and Chapter 18). This TAP-tagged gene is expressed in a cell and then the cell is broken open and the TAP-tagged protein is purified, carrying any interacting proteins along with it. The set of proteins that are co-purified with the TAP-tagged protein is then identified by mass spectrometry as in Figure 2.47. This process is repeated for each gene in the genome, eventually producing a map that shows all of the proteins that co-purify with each other and therefore presumably interact inside the cell. This complete set of interactions is called the “interactome” of the cell.

The results from large-scale protein–protein interaction studies can be presented in the form of a network, such as that shown in [FIGURE 2.48](#). This figure displays the potential binding partners of the various yeast proteins that contain an SH3 domain (see Figure 2.42a) and illustrates the complexities of such interactions at the level of an entire organism. Those proteins that have multiple binding partners, such as Las17 (situated near the center of Figure 2.48), are referred to as *hubs* of the protein interaction network. Hub proteins are more likely than non-hub proteins to be essential proteins, that is, proteins that the organism cannot survive without. Some hub proteins have several different binding interfaces and are capable of binding a number of different binding partners at the same time. In contrast, other hubs have a single binding interface, which is capable of binding several different partners, but only one at a time. Examples of each of these types of hub proteins are illustrated in [FIGURE 2.49](#). The hub protein depicted in Figure 2.49a plays a central role in the



**FIGURE 2.48** A network of protein–protein interactions. Each red line represents an interaction between two yeast proteins, which are indicated by the named black dots. In each case, the arrow points from an SH3 domain protein to a target protein with which it can bind. The 59 interactions depicted here were detected using two different types of techniques that measure protein–protein interactions. (See *Nature* 417:399–403, for discussion of the validity of protein–protein interaction studies.)

SOURCE: From A.H.Y. Tong, et al., *Science* 295:323, 2002, Copyright © 2002. Reprinted with permission from AAAS.

process of gene expression, while that shown in Figure 2.49b plays an equally important role in the process of cell division. Overall, it is estimated that, on average, each protein encoded in the genome of a eukaryotic organism interacts with about five different protein partners. According to this estimate, human proteins would engage in roughly 100,000 different interactions.

It is important to remember, when viewing these network diagrams, that any large-scale automated experiment will contain some errors. For interactome data, some true interactions will be missed, for example because the interaction is too weak to survive the purification process, or because the TAP tag disrupts an interacting domain. Likewise, some of the interactions reported in the analysis may not actually represent true interactions in the cell. For example if a protein is present as a contaminant when purifying a particular

protein complex, it would be incorrectly reported as interacting with the TAP-tagged protein. It is therefore best to view interactome datasets as showing potential interactions that can serve as a guide for future experiments using more direct means.

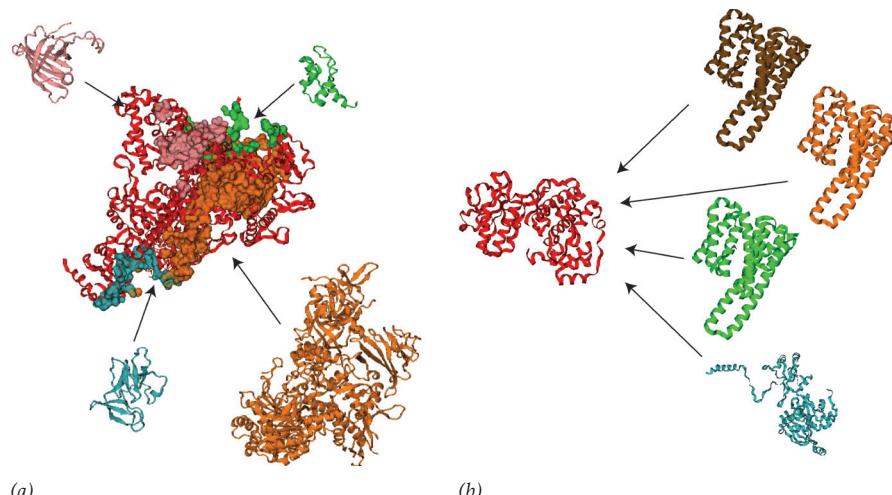
Aside from obtaining a lengthy list of *potential* interactions, what do we learn about cellular activities from these types of large-scale studies? Genome-sequencing projects have provided scientists with the amino acid sequences of a huge number of proteins whose very existence was previously unknown. What do these proteins do? One approach to determining a protein's function is to identify the proteins with which it associates. If, for example, a known protein has been shown to be involved in DNA replication, and an unknown protein is found to interact with the known protein, then it is likely that the unknown protein is also part of the cell's DNA-replication machinery. Thus, regardless of their limitations, these large-scale protein interaction studies provide the starting point to infer the function of unknown genes based on their possible interaction partners.

### REVIEW |

1. Which of the two methods discussed in this section, proteomics or interactomics, gives us information about the primary structure of proteins? Which gives information about quaternary structure of protein complexes?

## 2.16 Protein Engineering

Advances in molecular biology have created the opportunity to design and mass-produce novel proteins that are different from those made by living organisms. It is possible with current DNA-synthesizing techniques to create an artificial gene that can be used in the production of a protein having any desired sequence of amino acids. Polypeptides can also be synthesized from “scratch” using chemical techniques. This latter strategy allows researchers to incorporate building blocks other than the 20 amino acids that normally occur in nature.



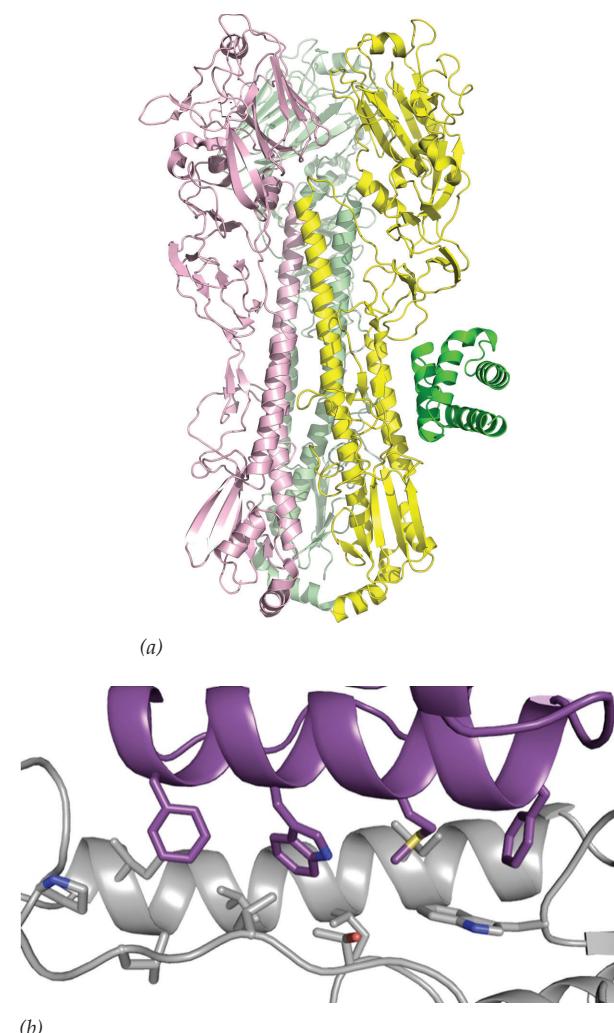
**FIGURE 2.49** Protein–protein interactions of hub proteins. (a) The enzyme RNA polymerase II, which synthesizes messenger RNAs in the cell, binds a multitude of other proteins simultaneously using multiple interfaces. (b) The enzyme Cdc28, which phosphorylates other proteins as it regulates the cell division cycle of budding yeast. Cdc28 binds a number of different proteins (Cln1–Cln3) at the same interface, which allows only one of these partners to bind at a time.

SOURCE: From Damien Devos and Robert B. Russell, *Curr. Opin. Struct. Biol.* 17:373, © 2007, with permission of Elsevier.

The problem with these types of engineering efforts is in knowing which of the virtually infinite variety of possible proteins one could manufacture might have some useful function. Consider, for example, a pharmaceutical company that wanted to manufacture a therapeutic protein that would bind to the surface of the AIDS or influenza virus. Assume that computer simulation programs could predict the shape such a protein should have to bind to the viral surface. What sequence of amino acids strung together would produce such a protein? The answer requires detailed insight into the rules governing the complex relationship between a protein's primary structure and its tertiary structure. **FIGURE 2.50** illustrates that protein biochemists now have the knowledge that allows them to construct a protein capable of binding to the surface of another protein, in this case the hemagglutinin (HA) protein that was present in the reconstructed 1918 influenza virus (page 24). HA protein is used by the virus to gain entry to a human cell, so that inhibition of HA could, in theory, prevent viral infection. Figure 2.50a shows the HA protein in proximity to a small engineered protein (in green). This engineered protein is capable of binding to a hydrophobic patch on the surface of the HA protein with high affinity. Figure 2.50b shows a closer view of the interface between the targeted portion of the HA protein (gray) and the binding surface of the designed protein (purple). It can be seen that side chains from the designed protein interact in highly specific ways with sites on the  $\alpha$  helix of HA. This peptide was found to inhibit the function of HA protein, potentially pointing the way to future efforts to developing HA binding peptides as antiviral drugs.

## Production of Novel Proteins

You might think that designing a protein from "scratch" that is capable of catalyzing a given chemical reaction—that is, designing an enzyme—would be far beyond the capability of present-day biotechnology. Given the magnitude of the task, it came as a surprise when researchers reported in 2008 that they had successfully designed and produced artificial proteins that were capable of catalyzing two different organic reactions, neither of which was catalyzed by any known natural enzyme. One of the reactions involved breaking a carbon–carbon bond, and the other the transfer of a proton from a carbon atom. These protein architects began by choosing a catalytic mechanism that might accelerate each chosen reaction and then used computer-based calculations to construct an idealized space in which amino acid side chains were positioned (forming an *active site*) to accomplish the task. They then searched among known protein structures to find ones that might serve as a framework or scaffold that could hold the active site they had designed. To transform the computer models into an actual protein, they used computational techniques to generate DNA sequences that had the potential to encode such a protein. The proposed DNA molecules were synthesized and introduced into bacterial cells where the proteins were manufactured. The catalytic activities of the proteins were then tested. Those proteins that showed the greatest promise were then subjected to a process of test-tube evolution; the proteins were mutated to create a new generation of altered proteins, which could in turn be screened for enhanced activity. Eventually, the team obtained proteins that could accelerate the rates of reaction as much as one million times that of the uncatalyzed reaction. While this is not a rate of enhancement that would fill a natural enzyme with pride, it is a remarkable accomplishment for a team of biochemists. It suggests, in fact, that



**FIGURE 2.50** The computational design of a protein that is capable of binding specifically to the surface of another protein. (a) The computationally designed protein is shown in green and its target protein (the HA protein from the H1N1 1918 influenza virus) is shown on the left in multiple colors. The predicted structure of the designed protein fits closely with that of the actual protein that was generated from the predicted sequence. (b) The actual interface of the targeted hydrophobic helix of the HA protein (gray) and the designed protein (purple). Side chains of the designed protein are seen to interact with sites on the HA helix.

SOURCE: (a) From Sarel J. Fleishman, et al., *Science* 332:820, 2011, image courtesy of David Baker; (b) From Bryan S. Der and Brian Kuhlman, *Science* 332:801, 2011, both © 2011, reprinted with permission of AAAs.

scientists will ultimately be able to construct proteins from scratch that will be capable of catalyzing virtually any chemical reaction.

An alternate approach to the production of novel proteins has been to modify those that are already produced by cells. Recent advances in DNA technology have allowed investigators to isolate an individual gene from human chromosomes, to alter its information content in a precisely determined way, and to synthesize the modified protein with its altered amino acid sequence. This technique, which is called **site-directed mutagenesis** (Section 18.25), has many different uses, both in basic research and in applied biology. If, for example, an investigator wants to know about the role of a particular residue in the folding or function of a polypeptide, the gene can be mutated in a way that substitutes an amino acid with different charge, hydrophobic character, or hydrogen-bonding properties. The effect of the substitution on the structure and function of the

modified protein can then be determined. As we will see throughout this textbook, site-directed mutagenesis has proven invaluable in the analysis of the specific functions of minute parts of virtually every protein of interest to biologists.

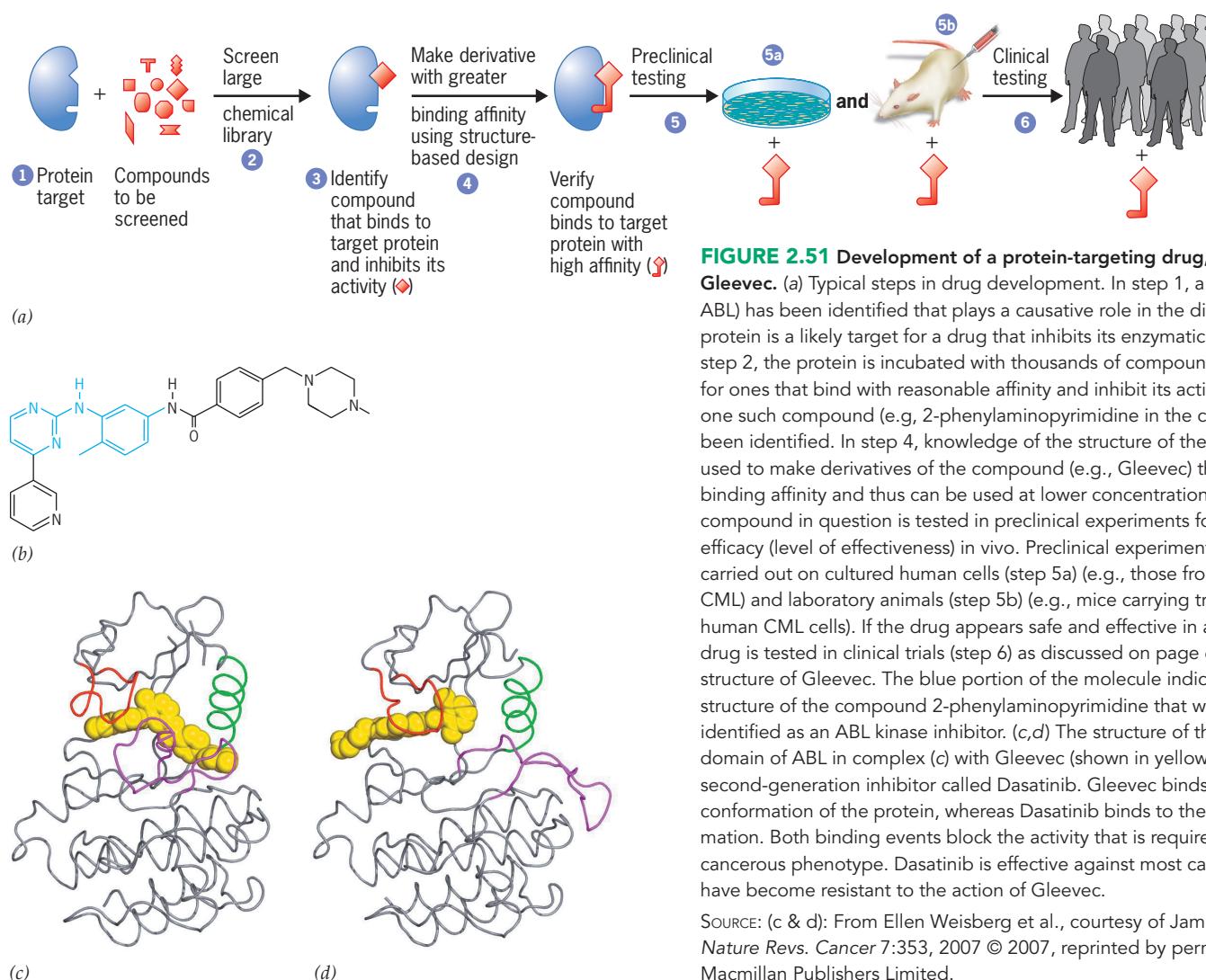
Site-directed mutagenesis is also used to modify the structure of clinically useful proteins to bring about various physiological effects. For example, the drug Somavert, which was approved by the FDA in 2003, is a modified version of human growth hormone (GH) containing several alterations. GH normally acts by binding to a receptor on the surface of target cells, which triggers a physiological response. Somavert competes with GH in binding to the GH receptor, but interaction between drug and receptor fails to trigger the cellular response. Somavert is prescribed for the treatment of acromegaly, a disorder that results from excess production of growth hormone.

## Structure-Based Drug Design

The production of new proteins is one clinical application of recent advances in molecular biology; another is the development of new drugs that act by binding to known proteins, thereby inhibiting their activity. Drug companies have access to chemical “libraries” that contain millions of different organic compounds that have been

either isolated from plants or microorganisms or chemically synthesized. One way to search for potential drugs is to expose the protein being targeted to combinations of these compounds and determine which compounds, if any, happen to bind to the protein with reasonable affinity. An alternate approach, called *structure-based drug design*, relies upon knowledge of the structure of the protein target. If the tertiary structure of a protein has been determined, researchers can use computers to design “virtual” drug molecules whose size and shape might allow them to fit into the apparent cracks and crevices of the protein, rendering it inactive.

We can illustrate both of these approaches by considering the development of the drug Gleevec (generic name Imatinib), as depicted in **FIGURE 2.51a**. Introduction of Gleevec into the clinic has revolutionized the treatment of a number of relatively rare cancers, most notably that of chronic myelogenous leukemia (CML). As discussed at length in Chapters 15 and 16, a group of enzymes called tyrosine kinases are often involved in the transformation of normal cells into cancer cells. Tyrosine kinases catalyze a reaction in which a phosphate group is added to specific tyrosine residues within a target protein, an event that may activate or inhibit the target protein. The development of CML is driven almost single-handedly by the presence of an overactive tyrosine kinase called ABL.



**FIGURE 2.51 Development of a protein-targeting drug, such as Gleevec.** (a) Typical steps in drug development. In step 1, a protein (e.g., ABL) has been identified that plays a causative role in the disease. This protein is a likely target for a drug that inhibits its enzymatic activity. In step 2, the protein is incubated with thousands of compounds in a search for ones that bind with reasonable affinity and inhibit its activity. In step 3, one such compound (e.g., 2-phenylaminopyrimidine in the case of ABL) has been identified. In step 4, knowledge of the structure of the target protein is used to make derivatives of the compound (e.g., Gleevec) that have greater binding affinity and thus can be used at lower concentrations. In step 5, the compound in question is tested in preclinical experiments for toxicity and efficacy (level of effectiveness) *in vivo*. Preclinical experiments are typically carried out on cultured human cells (step 5a) (e.g., those from patients with CML) and laboratory animals (step 5b) (e.g., mice carrying transplants of human CML cells). If the drug appears safe and effective in animals, the drug is tested in clinical trials (step 6) as discussed on page 65. (b) The structure of Gleevec. The blue portion of the molecule indicates the structure of the compound 2-phenylaminopyrimidine that was initially identified as an ABL kinase inhibitor. (c,d) The structure of the catalytic domain of ABL in complex (c) with Gleevec (shown in yellow) and (d) with a second-generation inhibitor called Dasatinib. Gleevec binds to the inactive conformation of the protein, whereas Dasatinib binds to the active conformation. Both binding events block the activity that is required for the cell's cancerous phenotype. Dasatinib is effective against most cancer cells that have become resistant to the action of Gleevec.

SOURCE: (c & d): From Ellen Weisberg et al., courtesy of James D. Griffin, *Nature Revs. Cancer* 7:353, 2007 © 2007, reprinted by permission from Macmillan Publishers Limited.

During the 1980s researchers identified a compound called 2-phenylaminopyrimidine that was capable of inhibiting tyrosine kinases. This compound was discovered by randomly screening a large chemical library for compounds that exhibited this particular activity (Figure 2.51a). As is usually the case in these types of blind screening experiments, 2-phenylaminopyrimidine would not have made a very effective drug. For one reason, it was only a weak enzyme inhibitor, which meant it would have had to be used in very large quantities. 2-Phenylaminopyrimidine is described as a *lead* compound, a starting point from which usable drugs might be developed. Beginning with this lead molecule, compounds of greater potency and specificity were synthesized using structure-based drug design. One of the compounds to emerge from this process was Gleevec (Figure 2.51b), which was found to bind tightly to the inactive form of the ABL tyrosine kinase and prevent the enzyme from becoming activated, which is a necessary step if the cell is to become cancerous. The complementary nature of the interaction between the drug and its enzyme target is shown in Figure 2.51c. Preclinical studies demonstrated that Gleevec strongly inhibited the growth in the laboratory of cells from CML patients and that the compound showed no harmful effects in tests in animals. In the very first clinical trial of Gleevec, virtually all of the CML patients went into remission after taking once-daily doses of the compound. Gleevec has gone on to become the primary drug prescribed for treatment of CML, but this is not the end of the story. Many patients taking Gleevec eventually experience a recurrence of their cancer when the ABL kinase mutates to become resistant to the drug. In such cases, the cancer can continue to be suppressed by treatment with more recently designed drugs that are capable of inhibiting Gleevec-resistant forms of the ABL kinase. One of these newer (second-generation) ABL kinase inhibitors is shown bound to the protein in Figure 2.51d.

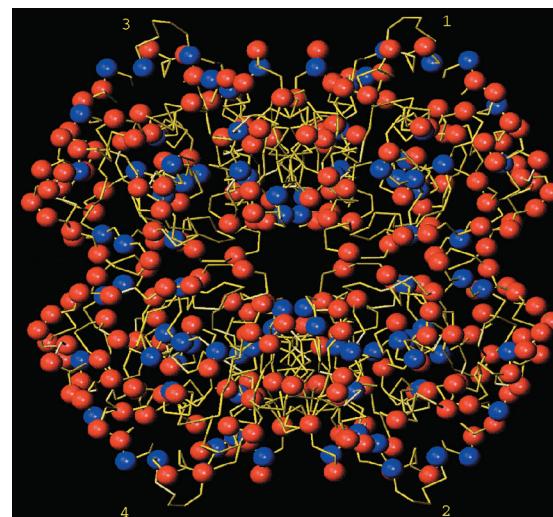
### REVIEW |

- What are some of the ways that knowledge of protein structure can help with developing new drugs?

## 2.17 Protein Adaptation and Evolution

Adaptations are traits that improve the likelihood that an organism will survive in a particular environment. Proteins are biochemical adaptations that are subject to natural selection and evolutionary change in the same way as other types of characteristics, such as eyes or skeletons. This is best revealed by comparing evolutionarily related (*homologous*) proteins in organisms living in very different environments. For example, the proteins of halophilic (salt-loving) archaeabacteria possess amino acid substitutions that allow them to maintain their solubility and function at very high cytosolic salt concentrations (up to 4 M KCl). Unlike its counterpart in other organisms, the surface of the halophilic version of the protein malate dehydrogenase, for example, is coated with aspartic and glutamic acid residues whose carboxyl groups can compete with the salt for water molecules (FIGURE 2.52).

Homologous proteins isolated from different organisms can exhibit virtually identical shapes and folding patterns, but show strikingly divergent amino acid sequences. The greater the evolutionary distance between two organisms, the greater the difference in the



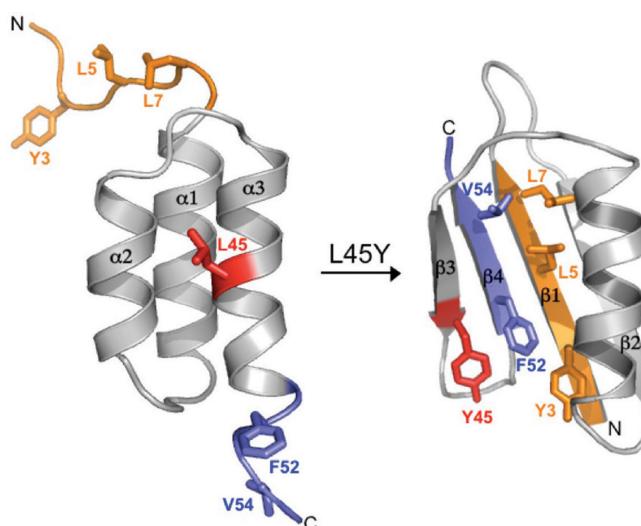
**FIGURE 2.52** Distribution of polar, charged amino acid residues in the enzyme malate dehydrogenase from a halophilic archaeabacterium. Red balls represent acidic residues, and blue balls represent basic residues. The surface of the enzyme is seen to be covered with acidic residues, which gives the protein a net charge of  $-156$ , and promotes its solubility in extremely salty environments. For comparison, a homologous protein from the dogfish, an ocean-dwelling shark, has a net charge of  $+16$ .

SOURCE: From O. Dym, M. Mevarech, and J. L. Sussman, *Science* 267:1345, © 1995, reprinted with permission of AAAS.

amino acid sequences of their proteins. In some cases, only a few key amino acids located in a critical portion of the protein will be present in all of the organisms from which that protein has been studied. In one comparison of 226 globin sequences, only two residues were found to be absolutely conserved in all of these polypeptides; one is a histidine residue that plays a key role in the binding and release of  $O_2$ . These observations indicate that the secondary and tertiary structures of proteins change much more slowly during evolution than their primary structures. This does not mean that the conformation of a protein cannot be affected in a major way by simple changes in primary structure. An example of such a change is shown in FIGURE 2.53. In this case, an amino acid substitution was experimentally introduced into a protein that completely altered the conformation of a small domain within a large protein molecule. The polypeptide on the left, which has a leucine at position 45, has a conformation consisting of a bundle of three  $\alpha$  helices, whereas the polypeptide on the right, which has a tyrosine at this position, has a conformation that contains a single  $\alpha$  helix and a four-stranded  $\beta$ -sheet. If a mutation having an effect of this magnitude happened to occur in nature, it might result in the formation of a protein with new functional properties and thus could be responsible for generating the ancestral form of an entirely new family of proteins.

We have seen how evolution has produced different versions of proteins in different organisms, but it has also produced different versions of proteins in individual organisms. Take a particular protein with a given function, such as globin or collagen. Several different versions of each of these proteins are encoded by the human genome. In most cases, different versions of a protein, which are known as **isoforms**, are adapted to function in different tissues or at different stages of development. For example, humans possess six different genes encoding isoforms of the cytoskeletal protein actin. Two of these isoforms are found in smooth muscle, one in skeletal muscle, one in heart muscle, and two in virtually all other types of cells.

Now that a large number of amino acid sequences and tertiary structures of proteins have been reported, it is clear that most proteins



**FIGURE 2.53** The dramatic effect on conformation that can result from a single amino acid substitution. In this case the switch between a leucine and a tyrosine at a critical position within this 56-amino acid polypeptide chain results in a transformation of the entire fold of the backbone of this polypeptide. This single substitution causes 85 percent of the amino acid residues to change their secondary structure. The spatial disposition of the two alternate side chains, which brings about this conformational shift, is shown in red in the model structures. The N-terminal amino acids are shown in orange and the C-terminal amino acids in blue.

SOURCE: From Patrick A. Alexander et al., Courtesy of Philip N. Bryan, Proc. Nat'l Acad. Sci. U.S.A. 106:21153, 2009, Fig. 6. © 2009 National Academy of Sciences.

are members of much larger **families** (or *superfamilies*) of related molecules. The genes that encode the various members of a protein family are thought to have arisen from a single ancestral gene that underwent a series of duplications during the course of evolution (see Figure 10.23). Over long periods of time, the nucleotide sequences of the various copies diverge from one another to generate proteins with related structures. Many protein families contain a remarkable variety of proteins that have evolved diverse functions. The expansion of protein families is responsible for much of the protein diversity encoded in the genomes of today's complex plants and animals.

### REVIEW |

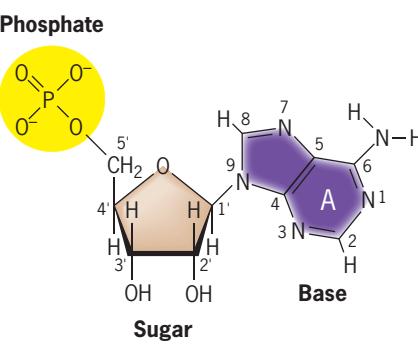
- Can a single amino acid change alter the secondary or tertiary structure of a protein?

## 2.18 Nucleic Acids

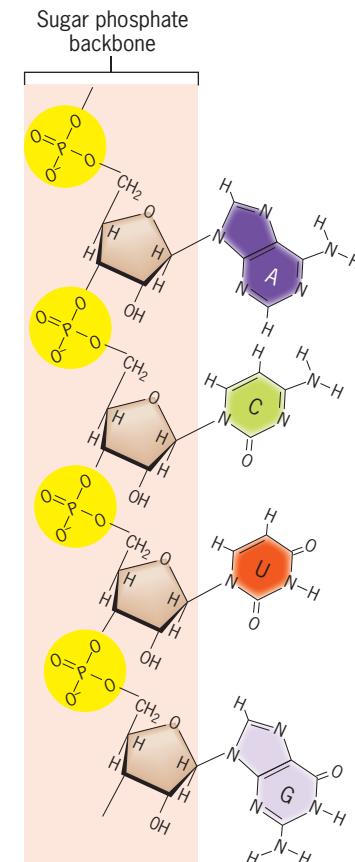
Nucleic acids are macromolecules constructed out of long chains (strands) of monomers called **nucleotides**. Nucleic acids function primarily in the **storage and transmission of genetic information**, but they may also have structural or catalytic roles. There are two types of nucleic acids found in living organisms, **deoxyribonucleic acid (DNA)** and **ribonucleic acid (RNA)**. DNA serves as the genetic material of all cellular organisms, though RNA carries out that role for many viruses. In cells, information stored in DNA is used to govern cellular activities through the formation of RNA messages. In the present discussion, we will examine the basic structure of nucleic acids using single-stranded RNA as the representative molecule.

We will look at the structure of DNA in Chapter 10, where it can be tied to its central role in the chemical basis of life.

Each nucleotide in a strand of RNA consists of three parts (**FIGURE 2.54a**): (1) a five-carbon sugar, ribose; (2) a nitrogenous base (so called because nitrogen atoms form part of the rings of the molecule); and (3) a phosphate group. The sugar and nitrogenous base together form a *nucleoside*, so that the nucleotides of an RNA strand are also known as ribonucleoside monophosphates. The phosphate is



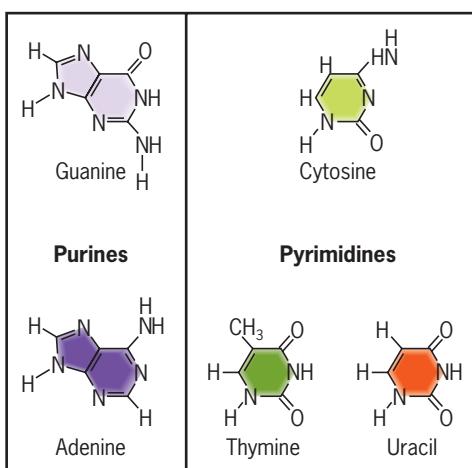
(a)



(b)

**FIGURE 2.54** Nucleotides and nucleotide strands of RNA.

(a) Nucleotides are the monomers from which strands of nucleic acid are constructed. A nucleotide consists of three parts: a sugar, a nitrogenous base, and a phosphate. The nucleotides of RNA contain the sugar ribose, which has a hydroxyl group bonded to the second carbon atom. In contrast, the nucleotides of DNA contain the sugar deoxyribose, which has a hydrogen atom rather than a hydroxyl group attached to the second carbon atom. Each nucleotide is polarized, having a 5' end (corresponding to the 5' side of the sugar) and a 3' end. (b) Nucleotides are joined together to form a strand by covalent bonds that link the 3' hydroxyl group of one sugar with the 5' phosphate group of the adjoining sugar.



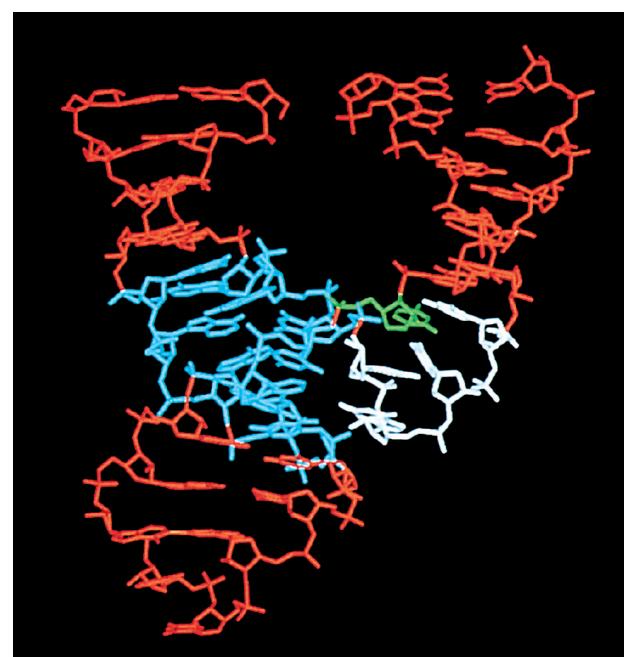
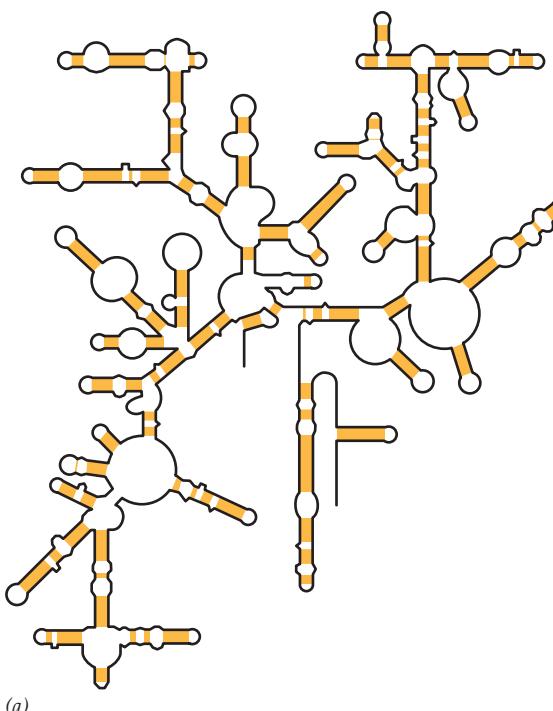
**FIGURE 2.55** Nitrogenous bases in nucleic acids. Of the four standard bases found in RNA, adenine and guanine are purines, and uracil and cytosine are pyrimidines. In DNA, the pyrimidines are cytosine and thymine, which differs from uracil by a methyl group attached to the ring.

linked to the 5' carbon of the sugar, and the nitrogenous base is attached to the sugar's 1' carbon. During the assembly of a nucleic acid strand, the hydroxyl group attached to the 3' carbon of the sugar of one nucleotide becomes linked by an ester bond to the phosphate group attached to the 5' carbon of the next nucleotide in the chain. Thus the nucleotides of an RNA (or DNA) strand are connected by sugar-phosphate linkages (Figure 2.54b), which are described as *3'-5'-phosphodiester bonds* because the phosphate atom is esterified to two oxygen atoms, one from each of the two adjoining sugars.

A strand of RNA (or DNA) contains four different types of nucleotides distinguished by their nitrogenous base. Two types of bases occur in nucleic acids: pyrimidines and purines (FIGURE 2.55). **Pyrimidines** are smaller molecules, consisting of a single ring; **purines** are larger, consisting of two rings. RNAs contain two different purines, **adenine** and **guanine**, and two different pyrimidines, **cytosine** and **uracil**. In DNA, uracil is replaced by **thymine**, a pyrimidine with an extra methyl group attached to the ring (Figure 2.55).

Although RNAs consist of a continuous single strand, they often fold back on themselves to produce molecules having extensive double-stranded segments and complex three-dimensional structures. This is illustrated by the two RNAs shown in FIGURE 2.56. The RNA whose secondary structure is shown in Figure 2.56a is a component of the small subunit of the bacterial ribosome (see FIGURE 2.57). Ribosomal RNAs are not molecules that carry genetic information; rather, they serve as structural scaffolds on which the proteins of the ribosome can be attached and as elements that recognize and bind various soluble components required for protein synthesis. One of the ribosomal RNAs of the large subunit acts as the catalyst for the reaction by which amino acids are covalently joined during protein synthesis. **RNAs having a catalytic role are called RNA enzymes, or ribozymes.** Figure 2.56b depicts the tertiary structure of the so-called hammerhead ribozyme, which is able to cleave its own RNA strand. In both examples shown in Figure 2.56, the double-stranded regions are held together by hydrogen bonds between the bases. This same principle is responsible for holding together the two strands of a DNA molecule.

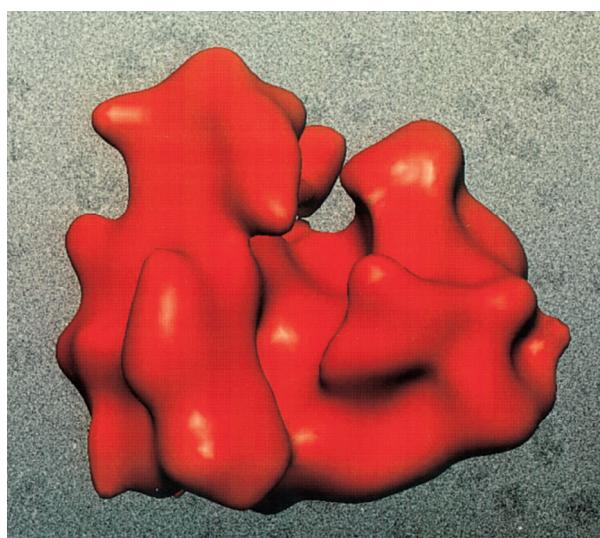
Nucleotides are not only important as building blocks of nucleic acids, they also have important functions in their own right. Most of the energy being put to use at any given moment in any living organism is derived from the nucleotide **adenosine triphosphate (ATP)**. The structure of ATP and its key role in cellular metabolism are



**FIGURE 2.56** RNAs can assume complex shapes. (a) This ribosomal RNA is an integral component of the small ribosomal subunit of a bacterium. In this two-dimensional profile, the RNA strand is seen to be folded back on itself in a highly ordered pattern so that most of the molecule is double-stranded. (b) This hammerhead ribozyme, as it is called, is a small RNA molecule from a viroid (page 25). The helical nature of the double-stranded portions of this RNA can be appreciated in this three-dimensional model of the molecule.

SOURCE: From William G. Scott, et al., *Cell* 81:993, © 1995, with permission from Elsevier.

discussed in the following chapter. **Guanosine triphosphate (GTP)** is another nucleotide of enormous importance in cellular activities. GTP binds to a variety of proteins (called G proteins) and acts as a switch to turn on their activities (see Figure 11.49 for an example).



**FIGURE 2.57** Reconstruction of a ribosome from the cytoplasm of a wheat germ cell. This reconstruction is based on high-resolution electron micrographs and shows the two subunits of this eukaryotic ribosome, the small (40S) subunit on the left and the large (60S) subunit on the right. The internal structure of a ribosome is discussed in Sections 11.16 and 11.17.

SOURCE: From Adriana Verschoor, et al., *J. Cell Biol.*, Vol. 133 (cover 3), 1996; by copyright permission of Rockefeller University Press.

#### REVIEW

1. Describe the structure of nucleotides and the manner in which these monomers are joined to form a polynucleotide strand. Why would it be overly simplistic to describe RNA as a single-stranded nucleic acid?
2. Which macromolecules are polymers? What is the basic structure of each type of monomer? How do the various monomers of each type of macromolecule vary among themselves?

## 2.19 The Formation of Complex Macromolecular Structures

To what degree can the lessons learned from the study of protein architecture be applied to more complex structures in the cell? Can structures, such as membranes, ribosomes, and cytoskeletal elements, which consist of different types of subunits, also assemble by themselves? How far can subcellular organization be explained simply by having the pieces fit together to form the most stable arrangement? The assembly of cellular organelles is poorly understood, but it is apparent from the following examples that different types of subunits can self-assemble to form higher-order arrangements.

### The Assembly of Tobacco Mosaic Virus Particles

The most convincing evidence that a particular assembly process is self-directed is the demonstration that the assembly can occur outside the cell (*in vitro*) under physiological conditions when the only macromolecules present are those that make up the final structure. In 1955, Heinz Fraenkel-Conrat and Robley Williams of the University

of California, Berkeley, demonstrated that TMV particles, which consist of one long RNA molecule (approximately 6600 nucleotides) wound within a helical capsule made of 2130 identical protein subunits (see Figure 1.21), were capable of self-assembly. In their experiments, they purified TMV RNA and protein separately, mixed them together under suitable conditions, and recovered mature, infective particles after a short period of incubation. Clearly the two components contain all the information necessary for particle formation.

### The Assembly of Ribosomal Subunits

Ribosomes, like TMV particles, are made of RNA and protein. Unlike the simpler TMV, ribosomes contain several different types of RNA and a considerable collection of different proteins. All ribosomes, regardless of their source, are composed of two subunits of different size. Although ribosomal subunits are often depicted in drawings as symmetric structures, in fact they have a highly irregular shape, as indicated in Figure 2.57. The large (or 50S) ribosomal subunit of bacteria contains two molecules of RNA and approximately 32 different proteins. The small (or 30S) ribosomal subunit of bacteria contains one molecule of RNA and 21 different proteins. The structure and function of the ribosome are discussed in detail in Sections 11.16 and 11.17.

One of the milestones in the study of ribosomes came in the mid-1960s, when Masayasu Nomura and his co-workers at the University of Wisconsin succeeded in reconstituting complete, fully functional 30S bacterial subunits by mixing the 21 purified proteins of the small subunit with purified small-subunit ribosomal RNA. Apparently, the components of the small subunit contain all the information necessary for the assembly of the entire particle. Analysis of the intermediates that form at different stages during reconstitution *in vitro* indicates that subunit assembly occurs in a sequential step-by-step manner that closely parallels the process *in vivo*. At least one of the proteins of the small subunit (S16) appears to function solely in ribosome assembly; deletion of this protein from the reconstitution mixture greatly slowed the assembly process but did not block the formation of fully functional ribosomes. Reconstitution of the large subunit of the bacterial ribosome was accomplished in the following decade. It should be kept in mind that although it takes approximately 2 hours at 50°C to reconstitute the ribosome *in vitro*, the bacterium can assemble the same structure in a few minutes at temperatures as low as 10°C. It may be that the bacterium uses something that is not available to the investigator who begins with purified components. Assembly of the ribosome within the cell, for example, may include the participation of accessory factors that function in protein folding, such as the chaperones described in the Experimental Pathways. In fact, the formation of ribosomes within a *eukaryotic* cell requires the transient association of many proteins that do not end up in the final particle, as well as the removal of approximately half the nucleotides of the large ribosomal RNA precursor (Section 11.4). As a result, the components of the mature eukaryotic ribosome no longer possess the information to reconstitute themselves *in vitro*.

#### REVIEW

1. What type of evidence suggests that bacterial ribosomal subunits are capable of self-assembly, but eukaryotic subunits are not?
2. What evidence would indicate that a particular ribosomal protein had a role in ribosome function but not assembly?

1. The pH of cola beverages is around 3, while the pH of battery acid is around 1. Which has a higher concentration of protons? By how much? If the concentration of protons in wine is approximately one-tenth that in cola, what is the pH of wine? (See Quantitative Tutorial Video).
2. Sickle cell anemia results from a substitution of a valine for a glutamic acid. Would you expect a similar effect if the mutation were to have placed a leucine at that site? An aspartic acid?
3. Of the following amino acids, glycine, isoleucine, and lysine, which would you expect to be the most soluble in an acidic aqueous solution? Which the least?
4. How many structural isomers could be formed from a molecule with the formula  $C_5H_{12}$ ?  $C_4H_8$ ?
5. Glyceraldehyde is the only three-carbon aldotetrose, and it can exist as two stereoisomers. What is the structure of dihydroxyacetone, the only ketotriose? How many stereoisomers does it form?
6. Bacteria are known to change the kinds of fatty acids they produce as the temperature of their environment changes. What types of changes in fatty acids would you expect as the temperature drops? Why would this be adaptive?
7. In the polypeptide backbone  $-C-C-N-C-C-N-C-C-$   $NH_2$ , identify the  $\alpha$ -carbons.
8. Which of the following are true? Increasing the pH of a solution would (1) suppress the dissociation of a carboxylic acid, (2) increase the charge on an amino group, (3) increase the dissociation of a carboxylic acid, (4) suppress the charge on an amino group.
9. Which of the four classes of amino acids has side chains with the greatest hydrogen-bond-forming potential? Which has the greatest potential to form ionic bonds? Hydrophobic interactions?
10. If the three enzymes of the pyruvate dehydrogenase complex existed as physically separate proteins rather than as a complex, what effect might this have on the rate of reactions catalyzed by these enzymes?
11. Would you agree that neither ribonuclease nor myoglobin had quaternary structure? Why or why not?
12. How many different tripeptides are possible? How many carboxyl terminals of polypeptide chains are present in a molecule of hemoglobin? (See Quantitative Tutorial Video).
13. You have isolated a pentapeptide composed of four glycine residues and one lysine residue that resides at the C-terminus of the peptide. Using the information provided in the legend of Figure 2.27, if the pK of the side chain of lysine is 10 and the pK of the terminal carboxyl group is 4, what is the structure of the peptide at pH 7? At pH 12?
14. The side chains of glutamic acid (pK 4.3) and arginine (pK 12.5) can form an ionic bond under certain conditions. Draw the relevant portions of the side chains and indicate whether an ionic bond could form at the following: (a) pH 4; (b) pH 7; (c) pH 12; (d) pH 13.
15. Would you expect a solution of high salt to be able to denature ribonuclease? Why or why not?
16. You have read in the Human Perspective that (1) mutations in the *PRNP* gene can make a polypeptide more likely to fold into the  $PrP^{Sc}$  conformation, thus causing CJD and (2) exposure to the  $PrP^{Sc}$  prion can lead to an infection that also causes CJD. How can you explain the occurrence of rare sporadic cases of the disease in persons who have no genetic propensity for it?
17. Persons who are born with Down syndrome have an extra (third) copy of chromosome #21 in their cells. Chromosome #21 contains the gene that encodes the APP protein. Why do you suppose that individuals with Down syndrome typically develop Alzheimer's disease at an early age?
18. We saw on page 77 how evolution has led to the existence of protein families composed of related molecules with similar functions. A few examples are also known where proteins with very similar functions have primary and tertiary structures that show no evidence of evolutionary relationship. Subtilisin and trypsin, for example, are two protein-digesting enzymes (proteases) that show no evidence they are homologous despite the fact that they utilize the same mechanism for attacking their substrates. How can this coincidence be explained?
19. Would you agree with the statement that many different amino acid sequences can fold into the same basic tertiary structure? What data can you cite as evidence for your position.
20. In the words of one scientist: "The first question any structural biologist asks upon being told that a new [protein] structure has been solved is no longer 'What does it look like?'; it is now 'What does it look like?'" What do you suppose he meant by this statement?