

# THE THREE-DIMENSIONAL STRUCTURE OF PROTEINS

- 4.1 Overview of Protein Structure** 116
- 4.2 Protein Secondary Structure** 120
- 4.3 Protein Tertiary and Quaternary Structures** 125
- 4.4 Protein Denaturation and Folding** 147

Perhaps the more remarkable features of [myoglobin] are its complexity and its lack of symmetry. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates, and it is more complicated than has been predicted by any theory of protein structure.

—John Kendrew, article in *Nature*, 1958

The covalent backbone of a typical protein contains hundreds of individual bonds. Because free rotation is possible around many of these bonds, the protein can assume an unlimited number of conformations. However, each protein has a specific chemical or structural function, strongly suggesting that each has a unique three-dimensional structure (Fig. 4–1). By the late 1920s, several proteins had been crystallized, including hemoglobin ( $M_r$  64,500) and the enzyme urease ( $M_r$  483,000). Given that the ordered array of molecules in a crystal can generally form only if the molecular units are identical, the simple fact that many proteins can be crystallized provides strong evidence that even very large proteins are discrete chemical entities with unique structures. This conclusion revolutionized thinking about proteins and their functions.

In this chapter, we explore the three-dimensional structure of proteins, emphasizing five themes. First, the three-dimensional structure of a protein is determined by its amino acid sequence. Second, the function

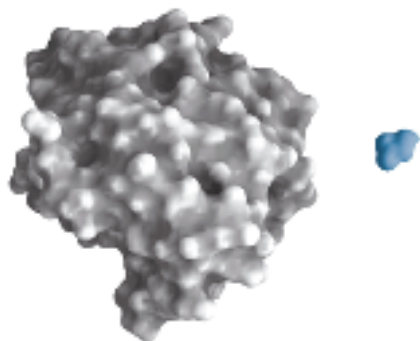
of a protein depends on its structure. Third, an isolated protein usually exists in one or a small number of stable structural forms. Fourth, the most important forces stabilizing the specific structures maintained by a given protein are noncovalent interactions. Finally, amid the huge number of unique protein structures, we can recognize some common structural patterns that help us organize our understanding of protein architecture.

These themes should not be taken to imply that proteins have static, unchanging three-dimensional structures. Protein function often entails an interconversion between two or more structural forms. The dynamic aspects of protein structure will be explored in Chapters 5 and 6.

The relationship between the amino acid sequence of a protein and its three-dimensional structure is an intricate puzzle that is gradually yielding to techniques used in modern biochemistry. An understanding of structure, in turn, is essential to the discussion of function in succeeding chapters. We can find and understand the patterns within the biochemical labyrinth of protein structure by applying fundamental principles of chemistry and physics.

## 4.1 Overview of Protein Structure

The spatial arrangement of atoms in a protein is called its **conformation**. The possible conformations of a protein include any structural state that can be achieved without breaking covalent bonds. A change in conformation could occur, for example, by rotation about single bonds. Of the numerous conformations that are theoretically possible in a protein containing hundreds of single bonds, one or (more commonly) a few generally predominate under biological conditions. The need for multiple stable conformations reflects the changes that must occur in most proteins as they bind to other



**FIGURE 4-1 Structure of the enzyme chymotrypsin, a globular protein.** Proteins are large molecules and, as we shall see, each has a unique structure. A molecule of glycine (blue) is shown for size comparison. The known three-dimensional structures of proteins are archived in the Protein Data Bank, or PDB ([www.rcsb.org/pdb](http://www.rcsb.org/pdb)). Each structure is assigned a unique four-character identifier, or PDB ID. Where appropriate, we will provide the PDB IDs for molecular graphic images in the figure captions. The image shown here was made using data from the PDB file 6GCH. The data from the PDB files provide only a series of coordinates detailing the location of atoms and their connectivity. Viewing the images requires easy-to-use graphics programs such as RasMol and Chime that convert the coordinates into an image and allow the viewer to manipulate the structure in three dimensions. You will find instructions for downloading Chime with the structure tutorials on the textbook website ([www.whfreeman.com/lehninger](http://www.whfreeman.com/lehninger)). The PDB website has instructions for downloading other viewers. We encourage all students to take advantage of the resources of the PDB and the free molecular graphics programs.

molecules or catalyze reactions. The conformations existing under a given set of conditions are usually the ones that are thermodynamically the most stable, having the lowest Gibbs free energy ( $G$ ). Proteins in any of their functional, folded conformations are called **native proteins**.

What principles determine the most stable conformations of a protein? An understanding of protein conformation can be built stepwise from the discussion of primary structure in Chapter 3 through a consideration of secondary, tertiary, and quaternary structures. To this traditional approach must be added a new emphasis on supersecondary structures, a growing set of known and classifiable protein folding patterns that provides an important organizational context to this complex endeavor. We begin by introducing some guiding principles.

### A Protein's Conformation Is Stabilized Largely by Weak Interactions

In the context of protein structure, the term **stability** can be defined as the tendency to maintain a native conformation. Native proteins are only marginally stable; the  $\Delta G$  separating the folded and unfolded states in typical proteins under physiological conditions is in the range of only 20 to 65 kJ/mol. A given polypeptide chain

can theoretically assume countless different conformations, and as a result the unfolded state of a protein is characterized by a high degree of conformational entropy. This entropy, and the hydrogen-bonding interactions of many groups in the polypeptide chain with solvent (water), tend to maintain the unfolded state. The chemical interactions that counteract these effects and stabilize the native conformation include disulfide bonds and the weak (noncovalent) interactions described in Chapter 2: hydrogen bonds, and hydrophobic and ionic interactions. An appreciation of the role of these weak interactions is especially important to our understanding of how polypeptide chains fold into specific secondary and tertiary structures, and how they combine with other polypeptides to form quaternary structures.

About 200 to 460 kJ/mol are required to break a single covalent bond, whereas weak interactions can be disrupted by a mere 4 to 30 kJ/mol. Individual covalent bonds that contribute to the native conformations of proteins, such as disulfide bonds linking separate parts of a single polypeptide chain, are clearly much stronger than individual weak interactions. Yet, because they are so numerous, it is weak interactions that predominate as a stabilizing force in protein structure. In general, the protein conformation with the lowest free energy (that is, the most stable conformation) is the one with the maximum number of weak interactions.

The stability of a protein is not simply the sum of the free energies of formation of the many weak interactions within it. Every hydrogen-bonding group in a folded polypeptide chain was hydrogen-bonded to water prior to folding, and for every hydrogen bond formed in a protein, a hydrogen bond (of similar strength) between the same group and water was broken. The net stability contributed by a given weak interaction, or the *difference* in free energies of the folded and unfolded states, may be close to zero. We must therefore look elsewhere to explain why the native conformation of a protein is favored.

We find that the contribution of weak interactions to protein stability can be understood in terms of the properties of water (Chapter 2). Pure water contains a network of hydrogen-bonded  $H_2O$  molecules. No other molecule has the hydrogen-bonding potential of water, and other molecules present in an aqueous solution disrupt the hydrogen bonding of water. When water surrounds a hydrophobic molecule, the optimal arrangement of hydrogen bonds results in a highly structured shell, or **solvation layer**, of water in the immediate vicinity. The increased order of the water molecules in the solvation layer correlates with an unfavorable decrease in the entropy of the water. However, when nonpolar groups are clustered together, there is a decrease in the extent of the solvation layer because each group no longer presents its entire surface to the solution. The result is a favorable increase in entropy. As described in

Chapter 2, this entropy term is the major thermodynamic driving force for the association of hydrophobic groups in aqueous solution. Hydrophobic amino acid side chains therefore tend to be clustered in a protein's interior, away from water.


Under physiological conditions, the formation of hydrogen bonds and ionic interactions in a protein is driven largely by this same entropic effect. Polar groups can generally form hydrogen bonds with water and hence are soluble in water. However, the number of hydrogen bonds per unit mass is generally greater for pure water than for any other liquid or solution, and there are limits to the solubility of even the most polar molecules as their presence causes a net decrease in hydrogen bonding per unit mass. Therefore, a solvation shell of structured water will also form to some extent around polar molecules. Even though the energy of formation of an intramolecular hydrogen bond or ionic interaction between two polar groups in a macromolecule is largely canceled out by the elimination of such interactions between the same groups and water, the release of structured water when the intramolecular interaction is formed provides an entropic driving force for folding. Most of the net change in free energy that occurs when weak interactions are formed within a protein is therefore derived from the increased entropy in the surrounding aqueous solution resulting from the burial of hydrophobic surfaces. This more than counterbalances the large loss of conformational entropy as a polypeptide is constrained into a single folded conformation.

Hydrophobic interactions are clearly important in stabilizing a protein conformation; the interior of a protein is generally a densely packed core of hydrophobic amino acid side chains. It is also important that any polar or charged groups in the protein interior have suitable partners for hydrogen bonding or ionic interactions. One hydrogen bond seems to contribute little to the stability of a native structure, but the presence of hydrogen-bonding or charged groups without partners in the hydrophobic core of a protein can be so *destabilizing* that conformations containing these groups are often thermodynamically untenable. The favorable free-energy change realized by combining such a group with a partner in the surrounding solution can be greater than the difference in free energy between the folded and unfolded states. In addition, hydrogen bonds between groups in proteins form cooperatively. Formation of one hydrogen bond facilitates the formation of additional hydrogen bonds. The overall contribution of hydrogen bonds and other noncovalent interactions to the stabilization of protein conformation is still being evaluated. The interaction of oppositely charged groups that form an ion pair (salt bridge) may also have a stabilizing effect on one or more native conformations of some proteins.

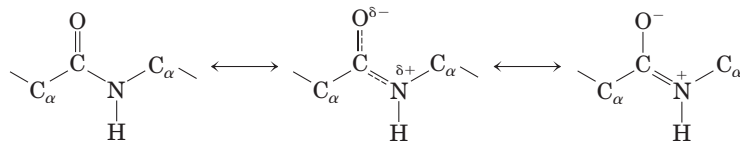
Most of the structural patterns outlined in this chapter reflect two simple rules: (1) hydrophobic residues

are largely buried in the protein interior, away from water; and (2) the number of hydrogen bonds within the protein is maximized. Insoluble proteins and proteins within membranes (which we examine in Chapter 11) follow somewhat different rules because of their function or their environment, but weak interactions are still critical structural elements.

### The Peptide Bond Is Rigid and Planar

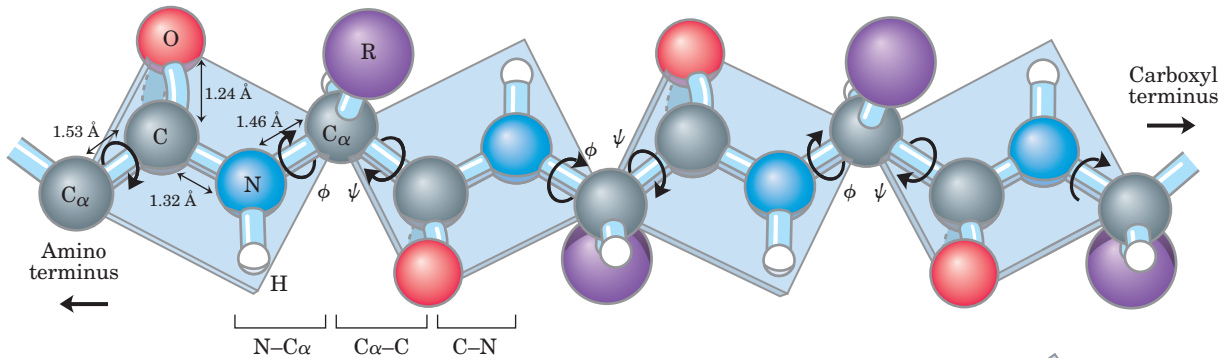
 **Protein Architecture—Primary Structure** Covalent bonds also place important constraints on the conformation of a polypeptide. In the late 1930s, Linus Pauling and Robert Corey embarked on a series of studies that laid the foundation for our present understanding of protein structure. They began with a careful analysis of the peptide bond. The  $\alpha$  carbons of adjacent amino acid residues are separated by three covalent bonds, arranged as  $C_\alpha-C-N-C_\alpha$ . X-ray diffraction studies of crystals of amino acids and of simple dipeptides and tripeptides demonstrated that the peptide C—N bond is somewhat shorter than the C—N bond in a simple amine and that the atoms associated with the peptide bond are coplanar. This indicated a resonance or partial sharing of two pairs of electrons between the carbonyl oxygen and the amide nitrogen (Fig. 4-2a). The oxygen has a partial negative charge and the nitrogen a partial positive charge, setting up a small electric dipole. The six atoms of the **peptide group** lie in a single plane, with the oxygen atom of the carbonyl group and the hydrogen atom of the amide nitrogen trans to each other. From these findings Pauling and Corey concluded that the peptide C—N bonds are unable to rotate freely because of their partial double-bond character. Rotation is permitted about the N— $C_\alpha$  and the  $C_\alpha$ —C bonds. The backbone of a polypeptide chain can thus be pictured as a series of rigid planes with consecutive planes sharing a common point of rotation at  $C_\alpha$  (Fig. 4-2b). The rigid peptide bonds limit the range of conformations that can be assumed by a polypeptide chain.

By convention, the bond angles resulting from rotations at  $C_\alpha$  are labeled  $\phi$  (phi) for the N— $C_\alpha$  bond and  $\psi$  (psi) for the  $C_\alpha$ —C bond. Again by convention, both  $\phi$  and  $\psi$  are defined as  $180^\circ$  when the polypeptide is in its fully extended conformation and all peptide groups are in the same plane (Fig. 4-2b). In principle,  $\phi$  and  $\psi$  can have any value between  $-180^\circ$  and  $+180^\circ$ , but many values are prohibited by steric interference between atoms in the polypeptide backbone and amino acid side chains. The conformation in which both  $\phi$  and  $\psi$  are  $0^\circ$  (Fig. 4-2c) is prohibited for this reason; this conformation is used merely as a reference point for describing the angles of rotation. Allowed values for  $\phi$  and  $\psi$  are graphically revealed when  $\psi$  is plotted versus  $\phi$  in a **Ramachandran plot** (Fig. 4-3), introduced by G. N. Ramachandran.



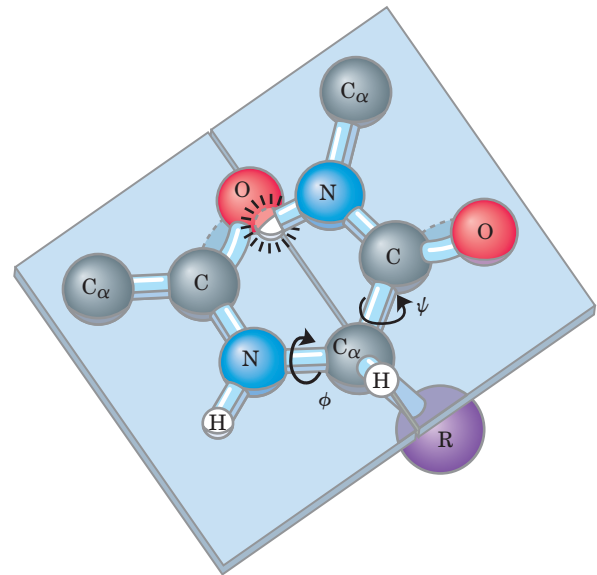
The carbonyl oxygen has a partial negative charge and the amide nitrogen a partial positive charge, setting up a small electric dipole. Virtually all peptide bonds in proteins occur in this trans configuration; an exception is noted in Figure 4–8b.

(a)



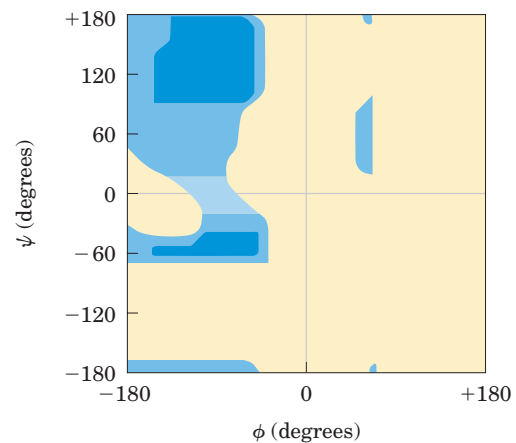
(b)

**FIGURE 4–2 The planar peptide group.** (a) Each peptide bond has some double-bond character due to resonance and cannot rotate. (b) Three bonds separate sequential  $\alpha$  carbons in a polypeptide chain. The  $N-C_\alpha$  and  $C_\alpha-C$  bonds can rotate, with bond angles designated  $\phi$  and  $\psi$ , respectively. The peptide  $C-N$  bond is not free to rotate. Other single bonds in the backbone may also be rotationally hindered, depending on the size and charge of the R groups. In the conformation shown,  $\phi$  and  $\psi$  are  $180^\circ$  (or  $-180^\circ$ ). As one looks out from the  $\alpha$  carbon, the  $\psi$  and  $\phi$  angles increase as the carbonyl or amide nitrogens (respectively) rotate clockwise. (c) By convention, both  $\phi$  and  $\psi$  are defined as  $0^\circ$  when the two peptide bonds flanking that  $\alpha$  carbon are in the same plane and positioned as shown. In a protein, this conformation is prohibited by steric overlap between an  $\alpha$ -carbonyl oxygen and an  $\alpha$ -amino hydrogen atom. To illustrate the bonds between atoms, the balls representing each atom are smaller than the van der Waals radii for this scale.  $1 \text{ \AA} = 0.1 \text{ nm}$ .

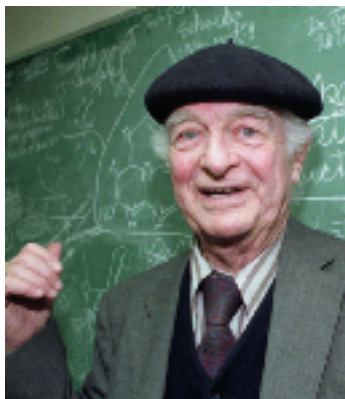


(c)

**FIGURE 4–3 Ramachandran plot for L-Ala residues.** The conformations of peptides are defined by the values of  $\phi$  and  $\psi$ . Conformations deemed possible are those that involve little or no steric interference, based on calculations using known van der Waals radii and bond angles. The areas shaded dark blue reflect conformations that involve no steric overlap and thus are fully allowed; medium blue indicates conformations allowed at the extreme limits for unfavorable atomic contacts; the lightest blue area reflects conformations that are permissible if a little flexibility is allowed in the bond angles. The asymmetry of the plot results from the L stereochemistry of the amino acid residues. The plots for other L-amino acid residues with unbranched side chains are nearly identical. The allowed ranges for branched amino acid residues such as Val, Ile, and Thr are somewhat smaller than for Ala. The Gly residue, which is less sterically hindered, exhibits a much broader range of allowed conformations. The range for Pro residues is greatly restricted because  $\phi$  is limited by the cyclic side chain to the range of  $-35^\circ$  to  $-85^\circ$ .







Linus Pauling, 1901–1994



Robert Corey, 1897–1971

### SUMMARY 4.1 Overview of Protein Structure

- Every protein has a three-dimensional structure that reflects its function.
- Protein structure is stabilized by multiple weak interactions. Hydrophobic interactions are the major contributors to stabilizing the globular form of most soluble proteins; hydrogen bonds and ionic interactions are optimized in the specific structures that are thermodynamically most stable.
- The nature of the covalent bonds in the polypeptide backbone places constraints on structure. The peptide bond has a partial double-bond character that keeps the entire six-atom peptide group in a rigid planar configuration. The N—C<sub>α</sub> and C<sub>α</sub>—C bonds can rotate to assume bond angles of  $\phi$  and  $\psi$ , respectively.

## 4.2 Protein Secondary Structure

The term **secondary structure** refers to the local conformation of some part of a polypeptide. The discussion of secondary structure most usefully focuses on common regular folding patterns of the polypeptide backbone. A few types of secondary structure are particularly stable and occur widely in proteins. The most prominent are the  $\alpha$  helix and  $\beta$  conformations described below. Using fundamental chemical principles and a few experimental observations, Pauling and Corey predicted the existence of these secondary structures in 1951, several years before the first complete protein structure was elucidated.

### The $\alpha$ Helix Is a Common Protein Secondary Structure

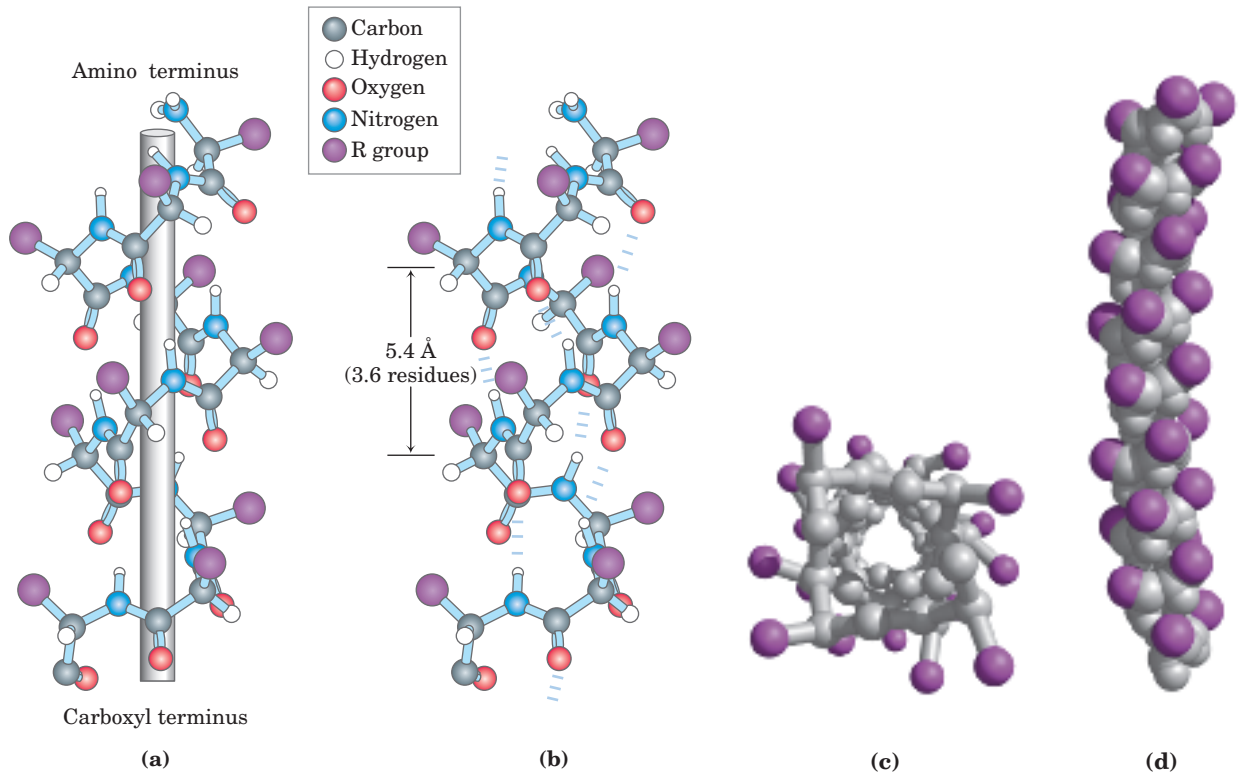
 **Protein Architecture— $\alpha$  Helix** Pauling and Corey were aware of the importance of hydrogen bonds in orient-

ing polar chemical groups such as the C=O and N—H groups of the peptide bond. They also had the experimental results of William Astbury, who in the 1930s had conducted pioneering x-ray studies of proteins. Astbury demonstrated that the protein that makes up hair and porcupine quills (the fibrous protein  $\alpha$ -keratin) has a regular structure that repeats every 5.15 to 5.2 Å. (The angstrom, Å, named after the physicist Anders J. Ångström, is equal to 0.1 nm. Although not an SI unit, it is used universally by structural biologists to describe atomic distances.) With this information and their data on the peptide bond, and with the help of precisely constructed models, Pauling and Corey set out to determine the likely conformations of protein molecules.

The simplest arrangement the polypeptide chain could assume with its rigid peptide bonds (but other single bonds free to rotate) is a helical structure, which Pauling and Corey called the  **$\alpha$  helix** (Fig. 4–4). In this structure the polypeptide backbone is tightly wound around an imaginary axis drawn longitudinally through the middle of the helix, and the R groups of the amino acid residues protrude outward from the helical backbone. The repeating unit is a single turn of the helix, which extends about 5.4 Å along the long axis, slightly greater than the periodicity Astbury observed on x-ray analysis of hair keratin. The amino acid residues in an  $\alpha$  helix have conformations with  $\psi = -45^\circ$  to  $-50^\circ$  and  $\phi = -60^\circ$ , and each helical turn includes 3.6 amino acid residues. The helical twist of the  $\alpha$  helix found in all proteins is right-handed (Box 4–1). The  $\alpha$  helix proved to be the predominant structure in  $\alpha$ -keratins. More generally, about one-fourth of all amino acid residues in polypeptides are found in  $\alpha$  helices, the exact fraction varying greatly from one protein to the next.

Why does the  $\alpha$  helix form more readily than many other possible conformations? The answer is, in part, that an  $\alpha$  helix makes optimal use of internal hydrogen bonds. The structure is stabilized by a hydrogen bond between the hydrogen atom attached to the electronegative nitrogen atom of a peptide linkage and the electronegative carbonyl oxygen atom of the fourth amino acid on the amino-terminal side of that peptide bond (Fig. 4–4b). Within the  $\alpha$  helix, every peptide bond (except those close to each end of the helix) participates in such hydrogen bonding. Each successive turn of the  $\alpha$  helix is held to adjacent turns by three to four hydrogen bonds. All the hydrogen bonds combined give the entire helical structure considerable stability.

Further model-building experiments have shown that an  $\alpha$  helix can form in polypeptides consisting of either L- or D-amino acids. However, all residues must be of one stereoisomeric series; a D-amino acid will disrupt a regular structure consisting of L-amino acids, and vice versa. Naturally occurring L-amino acids can form either right- or left-handed  $\alpha$  helices, but extended left-handed helices have not been observed in proteins.



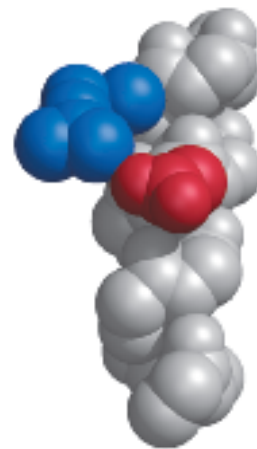
**FIGURE 4-4** Four models of the  $\alpha$  helix, showing different aspects of its structure. (a) Formation of a right-handed  $\alpha$  helix. The planes of the rigid peptide bonds are parallel to the long axis of the helix, depicted here as a vertical rod. (b) Ball-and-stick model of a right-handed  $\alpha$  helix, showing the intrachain hydrogen bonds. The repeat unit is a single turn of the helix, 3.6 residues. (c) The  $\alpha$  helix as viewed from one end, looking down the longitudinal axis (derived from PDB

ID 4TNC). Note the positions of the R groups, represented by purple spheres. This ball-and-stick model, used to emphasize the helical arrangement, gives the false impression that the helix is hollow, because the balls do not represent the van der Waals radii of the individual atoms. As the space-filling model (d) shows, the atoms in the center of the  $\alpha$  helix are in very close contact.

### Amino Acid Sequence Affects $\alpha$ Helix Stability

Not all polypeptides can form a stable  $\alpha$  helix. Interactions between amino acid side chains can stabilize or destabilize this structure. For example, if a polypeptide chain has a long block of Glu residues, this segment of the chain will not form an  $\alpha$  helix at pH 7.0. The negatively charged carboxyl groups of adjacent Glu residues repel each other so strongly that they prevent formation of the  $\alpha$  helix. For the same reason, if there are many adjacent Lys and/or Arg residues, which have positively charged R groups at pH 7.0, they will also repel each other and prevent formation of the  $\alpha$  helix. The bulk and shape of Asn, Ser, Thr, and Cys residues can also destabilize an  $\alpha$  helix if they are close together in the chain.

The twist of an  $\alpha$  helix ensures that critical interactions occur between an amino acid side chain and the side chain three (and sometimes four) residues away on either side of it (Fig. 4-5). Positively charged amino acids are often found three residues away from negatively charged amino acids, permitting the formation of an ion pair. Two aromatic amino acid residues are often similarly spaced, resulting in a hydrophobic interaction.



**FIGURE 4-5** Interactions between R groups of amino acids three residues apart in an  $\alpha$  helix. An ionic interaction between Asp<sup>100</sup> and Arg<sup>103</sup> in an  $\alpha$ -helical region of the protein troponin C, a calcium-binding protein associated with muscle, is shown in this space-filling model (derived from PDB ID 4TNC). The polypeptide backbone (carbons,  $\alpha$ -amino nitrogens, and  $\alpha$ -carbonyl oxygens) is shown in gray for a helix segment 13 residues long. The only side chains represented here are the interacting Asp (red) and Arg (blue) side chains.

## BOX 4-1 WORKING IN BIOCHEMISTRY

## Knowing the Right Hand from the Left

There is a simple method for determining whether a helical structure is right-handed or left-handed. Make fists of your two hands with thumbs outstretched and pointing straight up. Looking at your right hand, think of a helix spiraling up your right thumb in the direction in which the other four fingers are curled as shown (counterclockwise). The resulting helix is right-handed. Your left hand will demonstrate a left-handed helix, which rotates in the clockwise direction as it spirals up your thumb.

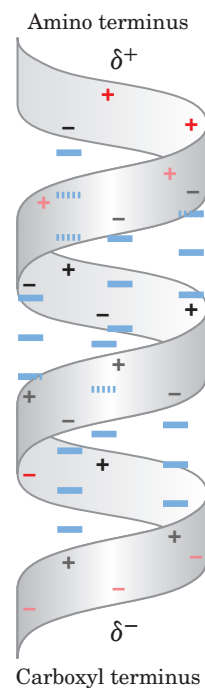


A constraint on the formation of the  $\alpha$  helix is the presence of Pro or Gly residues. In proline, the nitrogen atom is part of a rigid ring (see Fig. 4-8b), and rotation about the N—C $_{\alpha}$  bond is not possible. Thus, a Pro residue introduces a destabilizing kink in an  $\alpha$  helix. In addition, the nitrogen atom of a Pro residue in peptide linkage has no substituent hydrogen to participate in hydrogen bonds with other residues. For these reasons, proline is only rarely found within an  $\alpha$  helix. Glycine occurs infrequently in  $\alpha$  helices for a different reason: it has more conformational flexibility than the other amino acid residues. Polymers of glycine tend to take up coiled structures quite different from an  $\alpha$  helix.

A final factor affecting the stability of an  $\alpha$  helix in a polypeptide is the identity of the amino acid residues near the ends of the  $\alpha$ -helical segment. A small electric dipole exists in each peptide bond (Fig. 4-2a). These dipoles are connected through the hydrogen bonds of the helix, resulting in a net dipole extending along the helix that increases with helix length (Fig. 4-6). The four amino acid residues at each end of the helix do not participate fully in the helix hydrogen bonds. The partial positive and negative charges of the helix dipole actually reside on the peptide amino and carbonyl groups near the amino-terminal and carboxyl-terminal ends of the helix, respectively. For this reason, negatively charged amino acids are often found near the amino terminus of the helical segment, where they have a stabilizing interaction with the positive charge of the helix dipole; a positively charged amino acid at the amino-terminal end is destabilizing. The opposite is true at the carboxyl-terminal end of the helical segment.

Thus, five different kinds of constraints affect the stability of an  $\alpha$  helix: (1) the electrostatic repulsion (or attraction) between successive amino acid residues with charged R groups, (2) the bulkiness of adjacent R groups, (3) the interactions between R groups spaced

three (or four) residues apart, (4) the occurrence of Pro and Gly residues, and (5) the interaction between amino acid residues at the ends of the helical segment and the electric dipole inherent to the  $\alpha$  helix. The tendency of a given segment of a polypeptide chain to fold up as an  $\alpha$  helix therefore depends on the identity and sequence of amino acid residues within the segment.



**FIGURE 4-6 Helix dipole.** The electric dipole of a peptide bond (see Fig. 4-2a) is transmitted along an  $\alpha$ -helical segment through the intrachain hydrogen bonds, resulting in an overall helix dipole. In this illustration, the amino and carbonyl constituents of each peptide bond are indicated by + and - symbols, respectively. Non-hydrogen-bonded amino and carbonyl constituents in the peptide bonds near each end of the  $\alpha$ -helical region are shown in red.

## The $\beta$ Conformation Organizes Polypeptide Chains into Sheets

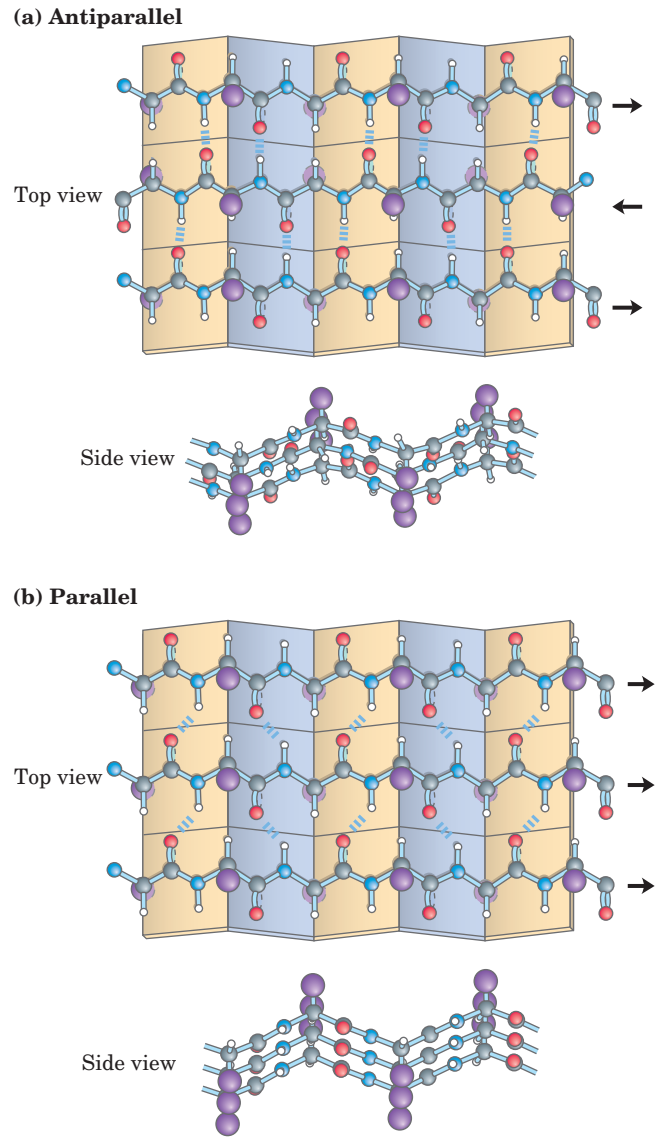
**Protein Architecture— $\beta$  Sheet** Pauling and Corey predicted a second type of repetitive structure, the  **$\beta$  conformation**. This is a more extended conformation of polypeptide chains, and its structure has been confirmed by x-ray analysis. In the  $\beta$  conformation, the backbone of the polypeptide chain is extended into a zigzag rather than helical structure (Fig. 4–7). The zigzag polypeptide chains can be arranged side by side to form a structure resembling a series of pleats. In this arrangement, called a  **$\beta$  sheet**, hydrogen bonds are formed between adjacent segments of polypeptide chain. The individual segments that form a  $\beta$  sheet are usually nearby on the polypeptide chain, but can also be quite distant from each other in the linear sequence of the polypeptide; they may even be segments in different polypeptide chains. The R groups of adjacent amino acids protrude from the zigzag structure in opposite directions, creating the alternating pattern seen in the side views in Figure 4–7.

The adjacent polypeptide chains in a  $\beta$  sheet can be either parallel or antiparallel (having the same or opposite amino-to-carboxyl orientations, respectively). The structures are somewhat similar, although the repeat period is shorter for the parallel conformation (6.5 Å, versus 7 Å for antiparallel) and the hydrogen-bonding patterns are different.

Some protein structures limit the kinds of amino acids that can occur in the  $\beta$  sheet. When two or more  $\beta$  sheets are layered close together within a protein, the R groups of the amino acid residues on the touching surfaces must be relatively small.  $\beta$ -Keratins such as silk fibroin and the fibroin of spider webs have a very high content of Gly and Ala residues, the two amino acids with the smallest R groups. Indeed, in silk fibroin Gly and Ala alternate over large parts of the sequence.

## $\beta$ Turns Are Common in Proteins

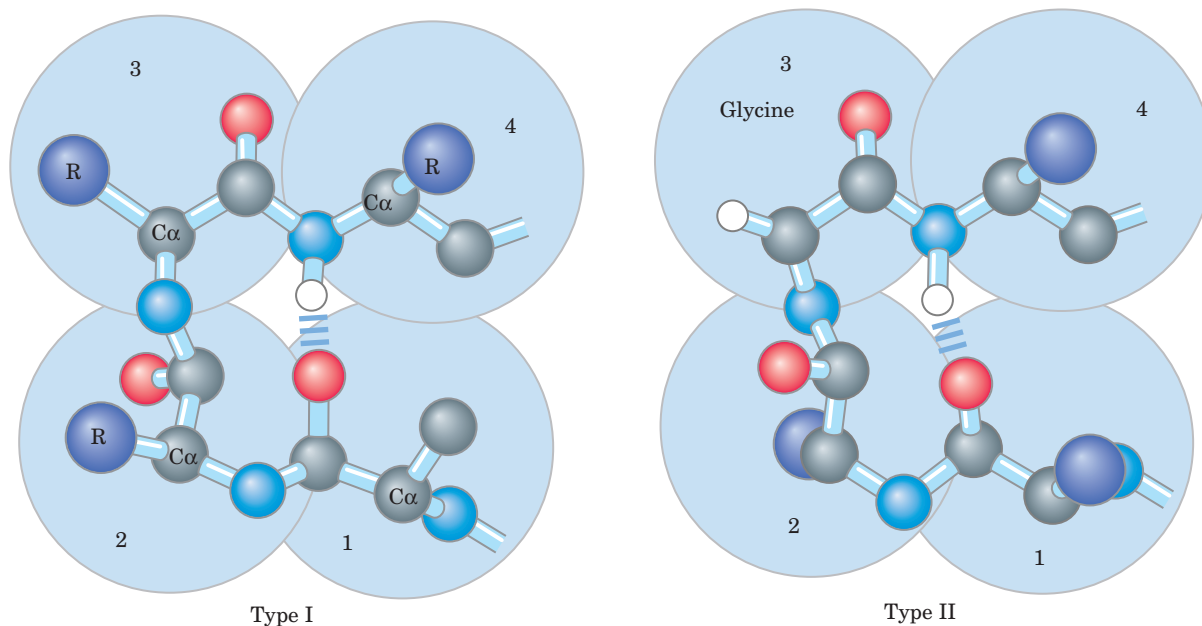
**Protein Architecture— $\beta$  Turn** In globular proteins, which have a compact folded structure, nearly one-third of the amino acid residues are in turns or loops where the polypeptide chain reverses direction (Fig. 4–8). These are the connecting elements that link successive runs of  $\alpha$  helix or  $\beta$  conformation. Particularly common are  **$\beta$  turns** that connect the ends of two adjacent segments of an antiparallel  $\beta$  sheet. The structure is a 180° turn involving four amino acid residues, with the carbonyl oxygen of the first residue forming a hydrogen bond with the amino-group hydrogen of the fourth. The peptide groups of the central two residues do not participate in any interresidue hydrogen bonding. Gly and Pro residues often occur in  $\beta$  turns, the former because it is small and flexible, the latter because peptide bonds



**FIGURE 4–7** The  $\beta$  conformation of polypeptide chains. These top and side views reveal the R groups extending out from the  $\beta$  sheet and emphasize the pleated shape described by the planes of the peptide bonds. (An alternative name for this structure is  $\beta$ -pleated sheet.) Hydrogen-bond cross-links between adjacent chains are also shown. (a) Antiparallel  $\beta$  sheet, in which the amino-terminal to carboxyl-terminal orientation of adjacent chains (arrows) is inverse. (b) Parallel  $\beta$  sheet.

involving the imino nitrogen of proline readily assume the cis configuration (Fig. 4–8b), a form that is particularly amenable to a tight turn. Of the several types of  $\beta$  turns, the two shown in Figure 4–8a are the most common. Beta turns are often found near the surface of a protein, where the peptide groups of the central two amino acid residues in the turn can hydrogen-bond with water. Considerably less common is the  $\gamma$  turn, a three-residue turn with a hydrogen bond between the first and third residues.



(a)  $\beta$  Turns

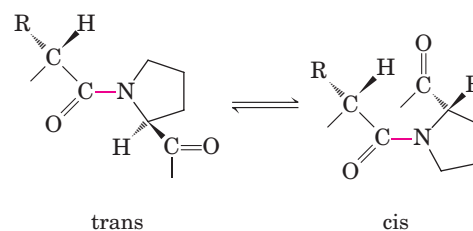
**FIGURE 4-8** Structures of  $\beta$  turns. (a) Type I and type II  $\beta$  turns are most common; type I turns occur more than twice as frequently as type II. Type II  $\beta$  turns always have Gly as the third residue. Note the hydrogen bond between the peptide groups of the first and fourth residues of the bends. (Individual amino acid residues are framed by large blue circles.) (b) The trans and cis isomers of a peptide bond involving the imino nitrogen of proline. Of the peptide bonds between amino acid residues other than Pro, over 99.95% are in the trans configuration. For peptide bonds involving the imino nitrogen of proline, however, about 6% are in the cis configuration; many of these occur at  $\beta$  turns.

### Common Secondary Structures Have Characteristic Bond Angles and Amino Acid Content

The  $\alpha$  helix and the  $\beta$  conformation are the major repetitive secondary structures in a wide variety of proteins, although other repetitive structures do exist in some specialized proteins (an example is collagen; see Fig. 4-13 on page 128). Every type of secondary structure can be completely described by the bond angles  $\phi$  and  $\psi$  at each residue. As shown by a Ramachandran plot, the  $\alpha$  helix and  $\beta$  conformation fall within a relatively restricted range of sterically allowed structures (Fig. 4-9a). Most values of  $\phi$  and  $\psi$  taken from known protein structures fall into the expected regions, with high concentrations near the  $\alpha$  helix and  $\beta$  conformation values as predicted (Fig. 4-9b). The only amino acid residue often found in a conformation outside these regions is glycine. Because its side chain, a single hydrogen atom, is small, a Gly residue can take part in many conformations that are sterically forbidden for other amino acids.

Some amino acids are accommodated better than others in the different types of secondary structures. An overall summary is presented in Figure 4-10. Some

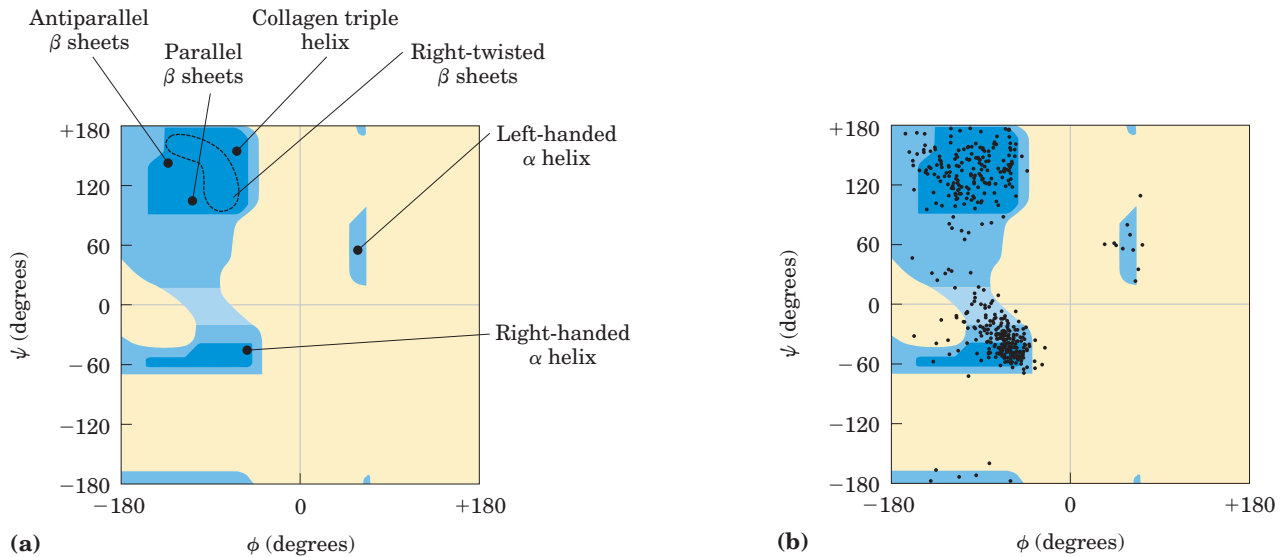
### (b) Proline isomers



biases, such as the common presence of Pro and Gly residues in  $\beta$  turns and their relative absence in  $\alpha$  helices, are readily explained by the known constraints on the different secondary structures. Other evident biases may be explained by taking into account the sizes or charges of side chains, but not all the trends shown in Figure 4-10 are understood.

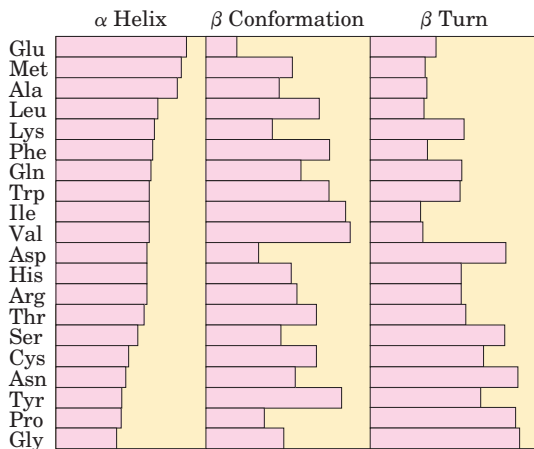
### SUMMARY 4.2 Protein Secondary Structure

- Secondary structure is the regular arrangement of amino acid residues in a segment of a polypeptide chain, in which each residue is spatially related to its neighbors in the same way.
- The most common secondary structures are the  $\alpha$  helix, the  $\beta$  conformation, and  $\beta$  turns.
- The secondary structure of a polypeptide segment can be completely defined if the  $\phi$  and  $\psi$  angles are known for all amino acid residues in that segment.



**FIGURE 4-9** Ramachandran plots for a variety of structures. (a) The values of  $\phi$  and  $\psi$  for various allowed secondary structures are overlaid on the plot from Figure 4-3. Although left-handed  $\alpha$  helices extending over several amino acid residues are theoretically possible, they have not been observed in proteins. (b) The values of  $\phi$  and  $\psi$

for all the amino acid residues except Gly in the enzyme pyruvate kinase (isolated from rabbit) are overlaid on the plot of theoretically allowed conformations (Fig. 4-3). The small, flexible Gly residues were excluded because they frequently fall outside the expected ranges (blue).



**FIGURE 4-10** Relative probabilities that a given amino acid will occur in the three common types of secondary structure.

## 4.3 Protein Tertiary and Quaternary Structures

**Protein Architecture—Introduction to Tertiary Structure** The overall three-dimensional arrangement of all atoms in a protein is referred to as the protein's **tertiary structure**. Whereas the term “secondary structure” refers to the spatial arrangement of amino acid residues that are adjacent in the primary structure, tertiary structure includes *longer-range* aspects of amino acid sequence. Amino acids that are far apart in the polypeptide sequence and that reside in different types of secondary structure may interact within the completely folded structure of a protein. The location of bends (including

$\beta$  turns) in the polypeptide chain and the direction and angle of these bends are determined by the number and location of specific bend-producing residues, such as Pro, Thr, Ser, and Gly. Interacting segments of polypeptide chains are held in their characteristic tertiary positions by different kinds of weak bonding interactions (and sometimes by covalent bonds such as disulfide cross-links) between the segments.

Some proteins contain two or more separate polypeptide chains, or subunits, which may be identical or different. The arrangement of these protein subunits in three-dimensional complexes constitutes **quaternary structure**.

In considering these higher levels of structure, it is useful to classify proteins into two major groups: **fibrous proteins**, having polypeptide chains arranged in long strands or sheets, and **globular proteins**, having polypeptide chains folded into a spherical or globular shape. The two groups are structurally distinct: fibrous proteins usually consist largely of a single type of secondary structure; globular proteins often contain several types of secondary structure. The two groups differ functionally in that the structures that provide support, shape, and external protection to vertebrates are made of fibrous proteins, whereas most enzymes and regulatory proteins are globular proteins. Certain fibrous proteins played a key role in the development of our modern understanding of protein structure and provide particularly clear examples of the relationship between structure and function. We begin our discussion with fibrous proteins, before turning to the more complex folding patterns observed in globular proteins.

## Fibrous Proteins Are Adapted for a Structural Function

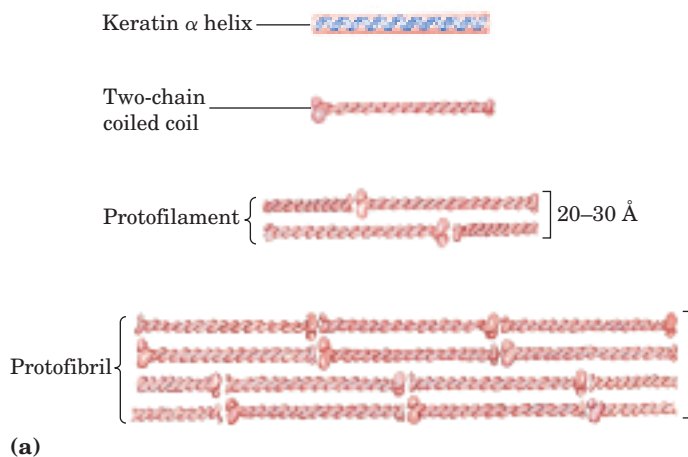
**Protein Architecture—Tertiary Structure of Fibrous Proteins**  
 $\alpha$ -Keratin, collagen, and silk fibroin nicely illustrate the relationship between protein structure and biological function (Table 4–1). Fibrous proteins share properties that give strength and/or flexibility to the structures in which they occur. In each case, the fundamental structural unit is a simple repeating element of secondary structure. All fibrous proteins are insoluble in water, a property conferred by a high concentration of hydrophobic amino acid residues both in the interior of the protein and on its surface. These hydrophobic surfaces are largely buried by packing many similar polypeptide chains together to form elaborate supramolecular complexes. The underlying structural simplicity of fibrous proteins makes them particularly useful for illustrating some of the fundamental principles of protein structure discussed above.

**$\alpha$ -Keratin** The  $\alpha$ -keratins have evolved for strength. Found in mammals, these proteins constitute almost the entire dry weight of hair, wool, nails, claws, quills, horns, hooves, and much of the outer layer of skin. The  $\alpha$ -keratins are part of a broader family of proteins called intermediate filament (IF) proteins. Other IF proteins are found in the cytoskeletons of animal cells. All IF proteins have a structural function and share structural features exemplified by the  $\alpha$ -keratins.

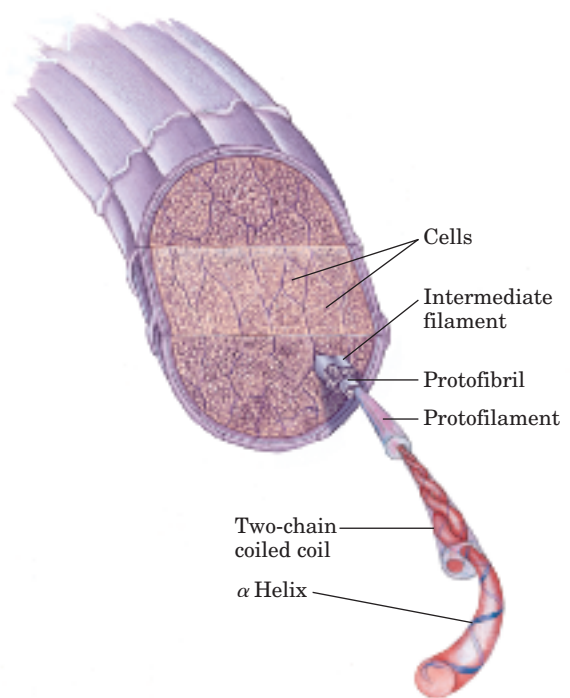
The  $\alpha$ -keratin helix is a right-handed  $\alpha$  helix, the same helix found in many other proteins. Francis Crick

and Linus Pauling in the early 1950s independently suggested that the  $\alpha$  helices of keratin were arranged as a coiled coil. Two strands of  $\alpha$ -keratin, oriented in parallel (with their amino termini at the same end), are wrapped about each other to form a supertwisted coiled coil. The supertwisting amplifies the strength of the overall structure, just as strands are twisted to make a strong rope (Fig. 4–11). The twisting of the axis of an  $\alpha$  helix to form a coiled coil explains the discrepancy between the 5.4 Å per turn predicted for an  $\alpha$  helix by Pauling and Corey and the 5.15 to 5.2 Å repeating structure observed in the x-ray diffraction of hair (p. 120). The helical path of the supertwists is left-handed, opposite in sense to the  $\alpha$  helix. The surfaces where the two  $\alpha$  helices touch are made up of hydrophobic amino acid residues, their R groups meshed together in a regular interlocking pattern. This permits a close packing of the polypeptide chains within the left-handed supertwist. Not surprisingly,  $\alpha$ -keratin is rich in the hydrophobic residues Ala, Val, Leu, Ile, Met, and Phe.

An individual polypeptide in the  $\alpha$ -keratin coiled coil has a relatively simple tertiary structure, dominated by an  $\alpha$ -helical secondary structure with its helical axis twisted in a left-handed superhelix. The intertwining of the two  $\alpha$ -helical polypeptides is an example of quaternary structure. Coiled coils of this type are common structural elements in filamentous proteins and in the muscle protein myosin (see Fig. 5–29). The quaternary structure of  $\alpha$ -keratin can be quite complex. Many coiled coils can be assembled into large supramolecular complexes, such as the arrangement of  $\alpha$ -keratin to form the intermediate filament of hair (Fig. 4–11b).



**FIGURE 4–11 Structure of hair.** (a) Hair  $\alpha$ -keratin is an elongated  $\alpha$  helix with somewhat thicker elements near the amino and carboxyl termini. Pairs of these helices are interwound in a left-handed sense to form two-chain coiled coils. These then combine in higher-order structures called protofilaments and microfibrils. About four microfibrils—32 strands of  $\alpha$ -keratin altogether—combine to form an intermediate filament. The individual two-chain coiled coils in the various substructures also appear to be interwound, but the handedness of the interwinding and other structural details are unknown. (b) A hair is an array of many  $\alpha$ -keratin filaments, made up of the substructures shown in (a).



**(b) Cross section of a hair**

**TABLE 4-1** Secondary Structures and Properties of Fibrous Proteins

Structure	Characteristics	Examples of occurrence
$\alpha$ Helix, cross-linked by disulfide bonds	Tough, insoluble protective structures of varying hardness and flexibility	$\alpha$ -Keratin of hair, feathers, and nails
$\beta$ Conformation	Soft, flexible filaments	Silk fibroin
Collagen triple helix	High tensile strength, without stretch	Collagen of tendons, bone matrix

The strength of fibrous proteins is enhanced by covalent cross-links between polypeptide chains within the multihelical “ropes” and between adjacent chains in a supramolecular assembly. In  $\alpha$ -keratins, the cross-links stabilizing quaternary structure are disulfide bonds (Box 4-2). In the hardest and toughest  $\alpha$ -keratins, such as those of rhinoceros horn, up to 18% of the residues are cysteines involved in disulfide bonds.

**Collagen** Like the  $\alpha$ -keratins, collagen has evolved to provide strength. It is found in connective tissue such as tendons, cartilage, the organic matrix of bone, and the cornea of the eye. The collagen helix is a unique

secondary structure quite distinct from the  $\alpha$  helix. It is left-handed and has three amino acid residues per turn (Fig. 4-12). Collagen is also a coiled coil, but one with distinct tertiary and quaternary structures: three separate polypeptides, called  $\alpha$  chains (not to be confused with  $\alpha$  helices), are supertwisted about each other (Fig. 4-12c). The superhelical twisting is right-handed in collagen, opposite in sense to the left-handed helix of the  $\alpha$  chains.

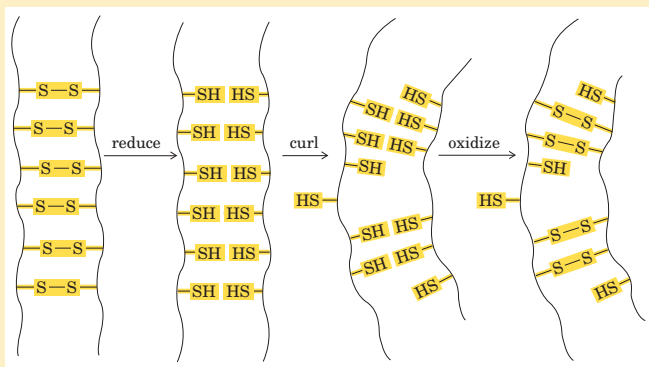
There are many types of vertebrate collagen. Typically they contain about 35% Gly, 11% Ala, and 21% Pro and 4-Hyp (4-hydroxyproline, an uncommon amino acid; see Fig. 3-8a). The food product gelatin is derived

## BOX 4-2 THE WORLD OF BIOCHEMISTRY

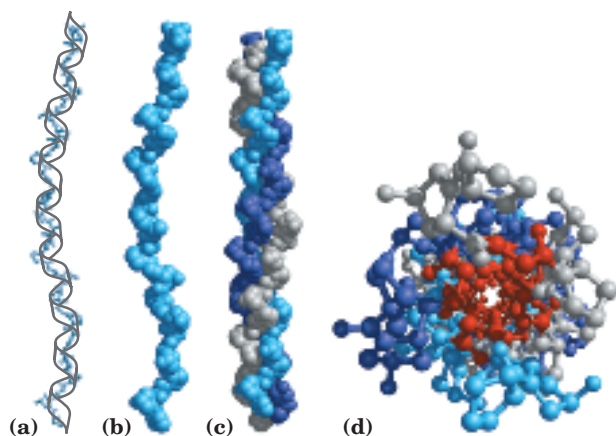
### Permanent Waving Is Biochemical Engineering

When hair is exposed to moist heat, it can be stretched. At the molecular level, the  $\alpha$  helices in the  $\alpha$ -keratin of hair are stretched out until they arrive at the fully extended  $\beta$  conformation. On cooling they spontaneously revert to the  $\alpha$ -helical conformation. The characteristic “stretchability” of  $\alpha$ -keratins, and their numerous disulfide cross-linkages, are the basis of permanent waving. The hair to be waved or curled is first bent around a form of appropriate shape. A solution of a reducing agent, usually a compound containing a thiol or sulfhydryl group ( $-\text{SH}$ ), is then applied with heat. The reducing agent cleaves the cross-linkages by reducing each disulfide bond to form two Cys residues. The moist heat breaks hydrogen

bonds and causes the  $\alpha$ -helical structure of the polypeptide chains to uncoil. After a time the reducing solution is removed, and an oxidizing agent is added to establish *new* disulfide bonds between pairs of Cys residues of adjacent polypeptide chains, but not the same pairs as before the treatment. After the hair is washed and cooled, the polypeptide chains revert to their  $\alpha$ -helical conformation. The hair fibers now curl in the desired fashion because the new disulfide cross-linkages exert some torsion or twist on the bundles of  $\alpha$ -helical coils in the hair fibers. A permanent wave is not truly permanent, because the hair grows; in the new hair replacing the old, the  $\alpha$ -keratin has the natural, nonwavy pattern of disulfide bonds.





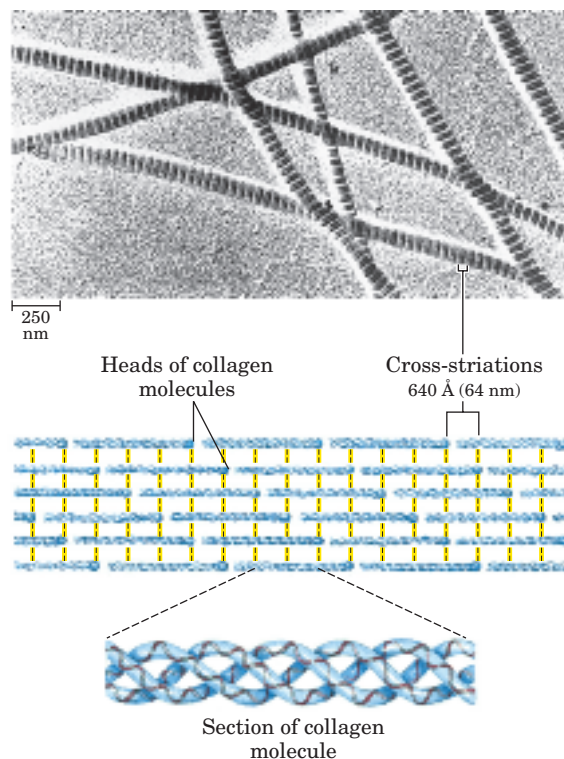


**FIGURE 4-12 Structure of collagen.** (Derived from PDB ID 1CGD.)

(a) The  $\alpha$  chain of collagen has a repeating secondary structure unique to this protein. The repeating tripeptide sequence Gly-X-Pro or Gly-X-4-Hyp adopts a left-handed helical structure with three residues per turn. The repeating sequence used to generate this model is Gly-Pro-4-Hyp. (b) Space-filling model of the same  $\alpha$  chain. (c) Three of these helices (shown here in gray, blue, and purple) wrap around one another with a right-handed twist. (d) The three-stranded collagen superhelix shown from one end, in a ball-and-stick representation. Gly residues are shown in red. Glycine, because of its small size, is required at the tight junction where the three chains are in contact. The balls in this illustration do not represent the van der Waals radii of the individual atoms. The center of the three-stranded superhelix is not hollow, as it appears here, but is very tightly packed.

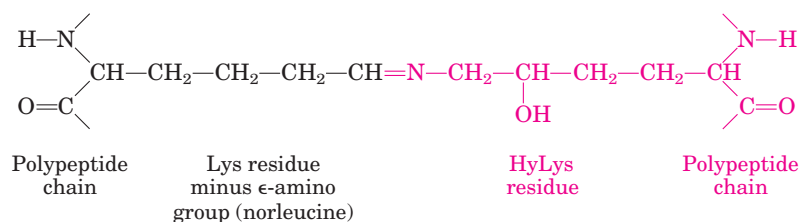
from collagen; it has little nutritional value as a protein, because collagen is extremely low in many amino acids that are essential in the human diet. The unusual amino acid content of collagen is related to structural constraints unique to the collagen helix. The amino acid sequence in collagen is generally a repeating tripeptide unit, Gly-X-Y, where X is often Pro, and Y is often 4-Hyp. Only Gly residues can be accommodated at the very tight junctions between the individual  $\alpha$  chains (Fig. 4-12d); The Pro and 4-Hyp residues permit the sharp twisting of the collagen helix. The amino acid sequence and the supertwisted quaternary structure of collagen allow a very close packing of its three polypeptides. 4-Hydroxyproline has a special role in the structure of collagen—and in human history (Box 4-3).

The tight wrapping of the  $\alpha$  chains in the collagen triple helix provides tensile strength greater than that

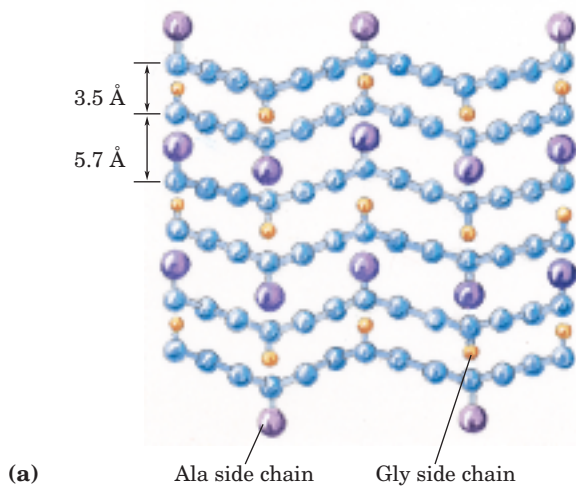


**FIGURE 4-13 Structure of collagen fibrils.** Collagen ( $M_r$  300,000) is a rod-shaped molecule, about 3,000 Å long and only 15 Å thick. Its three helically intertwined  $\alpha$  chains may have different sequences, but each has about 1,000 amino acid residues. Collagen fibrils are made up of collagen molecules aligned in a staggered fashion and cross-linked for strength. The specific alignment and degree of cross-linking vary with the tissue and produce characteristic cross-striations in an electron micrograph. In the example shown here, alignment of the head groups of every fourth molecule produces striations 640 Å apart.


of a steel wire of equal cross section. Collagen fibrils (Fig. 4-13) are supramolecular assemblies consisting of triple-helical collagen molecules (sometimes referred to as tropocollagen molecules) associated in a variety of ways to provide different degrees of tensile strength. The  $\alpha$  chains of collagen molecules and the collagen molecules of fibrils are cross-linked by unusual types of covalent bonds involving Lys, HyLys (5-hydroxylysine; see Fig. 3-8a), or His residues that are present at a few of the X and Y positions in collagens. These links create uncommon amino acid residues such as dehydrohydroxylysinonorleucine. The increasingly rigid and brittle character of aging connective tissue results from accumulated covalent cross-links in collagen fibrils.



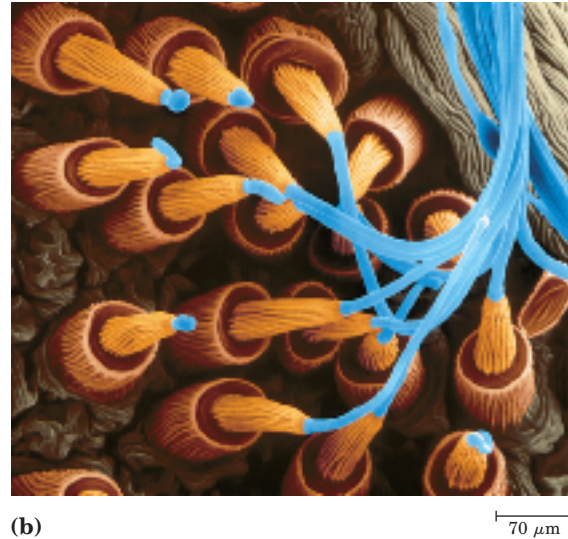
Dehydrohydroxylysinonorleucine



**FIGURE 4-14 Structure of silk.** The fibers used to make silk cloth or a spider web are made up of the protein fibroin. (a) Fibroin consists of layers of antiparallel  $\beta$  sheets rich in Ala (purple) and Gly (yellow) residues. The small side chains interdigitate and allow close packing

 A typical mammal has more than 30 structural variants of collagen, particular to certain tissues and each somewhat different in sequence and function. Some human genetic defects in collagen structure illustrate the close relationship between amino acid sequence and three-dimensional structure in this protein. Osteogenesis imperfecta is characterized by abnormal bone formation in babies; Ehlers-Danlos syndrome is characterized by loose joints. Both conditions can be lethal, and both result from the substitution of an amino acid residue with a larger R group (such as Cys or Ser) for a single Gly residue in each  $\alpha$  chain (a different Gly residue in each disorder). These single-residue substitutions have a catastrophic effect on collagen function because they disrupt the Gly-X-Y repeat that gives collagen its unique helical structure. Given its role in the collagen triple helix (Fig. 4-12d), Gly cannot be replaced by another amino acid residue without substantial deleterious effects on collagen structure. ■

**Silk Fibroin** Fibroin, the protein of silk, is produced by insects and spiders. Its polypeptide chains are predominantly in the  $\beta$  conformation. Fibroin is rich in Ala and Gly residues, permitting a close packing of  $\beta$  sheets and an interlocking arrangement of R groups (Fig. 4-14). The overall structure is stabilized by extensive hydrogen bonding between all peptide linkages in the polypeptides of each  $\beta$  sheet and by the optimization of van der Waals interactions between sheets. Silk does not stretch, because the  $\beta$  conformation is already highly extended (Fig. 4-7; see also Fig. 4-15). However, the structure is flexible because the sheets are held together by numerous weak interactions rather than by covalent bonds such as the disulfide bonds in  $\alpha$ -keratins.



of each layered sheet, as shown in this side view. (b) Strands of fibroin (blue) emerge from the spinnerets of a spider in this colored electron micrograph.

### Structural Diversity Reflects Functional Diversity in Globular Proteins

In a globular protein, different segments of a polypeptide chain (or multiple polypeptide chains) fold back on each other. As illustrated in Figure 4-15, this folding generates a compact form relative to polypeptides in a fully extended conformation. The folding also provides the structural diversity necessary for proteins to carry out a wide array of biological functions. Globular proteins include enzymes, transport proteins, motor proteins, regulatory proteins, immunoglobulins, and proteins with many other functions.

As a new millennium begins, the number of known three-dimensional protein structures is in the thousands and more than doubles every two years. This wealth of structural information is revolutionizing our understanding of protein structure, the relation of structure

$\beta$  Conformation  
2,000  $\times$  5 Å

$\alpha$  Helix  
900  $\times$  11 Å

Native globular form  
100  $\times$  60 Å

**FIGURE 4-15** Globular protein structures are compact and varied. Human serum albumin ( $M_r$  64,500) has 585 residues in a single chain. Given here are the approximate dimensions its single polypeptide chain would have if it occurred entirely in extended  $\beta$  conformation or as an  $\alpha$  helix. Also shown is the size of the protein in its native globular form, as determined by X-ray crystallography; the polypeptide chain must be very compactly folded to fit into these dimensions.



## BOX 4-3 BIOCHEMISTRY IN MEDICINE

### Why Sailors, Explorers, and College Students Should Eat Their Fresh Fruits and Vegetables

... from this misfortune, together with the unhealthiness of the country, where there never falls a drop of rain, we were stricken with the “camp-sickness,” which was such that the flesh of our limbs all shrivelled up, and the skin of our legs became all blotched with black, mouldy patches, like an old jack-boot, and proud flesh came upon the gums of those of us who had the sickness, and none escaped from this sickness save through the jaws of death. The signal was this: when the nose began to bleed, then death was at hand . . .

—from *The Memoirs of the Lord of Joinville*, ca. 1300

This excerpt describes the plight of Louis IX’s army toward the end of the Seventh Crusade (1248–1254), immediately preceding the battle of Fariskur, where the scurvy-weakened Crusader army was destroyed by the Egyptians. What was the nature of the malady afflicting these thirteenth-century soldiers?

Scurvy is caused by lack of vitamin C, or ascorbic acid (ascorbate). Vitamin C is required for, among other things, the hydroxylation of proline and lysine in collagen; scurvy is a deficiency disease characterized by general degeneration of connective tissue. Manifestations of advanced scurvy include numerous small hemorrhages caused by fragile blood vessels, tooth loss, poor wound healing and the reopening of old wounds, bone pain and degeneration, and eventually heart failure. Despondency and oversensitivity to stimuli of many kinds are also observed. Milder cases of vitamin

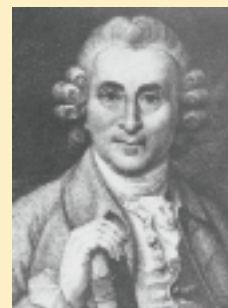


**FIGURE 1** Iroquois showing Jacques Cartier how to make cedar tea as a remedy for scurvy.

C deficiency are accompanied by fatigue, irritability, and an increased severity of respiratory tract infections. Most animals make large amounts of vitamin C, converting glucose to ascorbate in four enzymatic steps. But in the course of evolution, humans and some other animals—gorillas, guinea pigs, and fruit bats—have lost the last enzyme in this pathway and must obtain ascorbate in their diet. Vitamin C is available in a wide range of fruits and vegetables. Until 1800, however, it was often absent in the dried foods and other food supplies stored for winter or for extended travel.

Scurvy was recorded by the Egyptians in 1500 BCE, and it is described in the fifth century BCE writings of Hippocrates. Although scurvy played a critical role in medieval wars and made regular winter appearances in northern climates, it did not come to wide public notice until the European voyages of discovery from 1500 to 1800. The first circumnavigation of the globe, led by Ferdinand Magellan (1520), was accomplished only with the loss of more than 80% of his crew to scurvy. Vasco da Gama lost two-thirds of his crew to the same fate during his first exploration of trade routes to India (1499). During Jacques Cartier’s second voyage to explore the St. Lawrence River (1535–1536), his band suffered numerous fatalities and was threatened with complete disaster until the native Americans taught the men to make a cedar tea that cured and prevented scurvy (it contained vitamin C) (Fig. 1). It is estimated that a million sailors died of scurvy in the years 1600 to 1800. Winter outbreaks of scurvy in Europe were gradually eliminated in the nineteenth century as the cultivation of the potato, introduced from South America, became widespread.

In 1747, James Lind, a Scottish surgeon in the Royal Navy (Fig. 2), carried out the first controlled clinical study in recorded history. During an extended voyage on the 50-gun warship *HMS Salisbury*, Lind selected 12 sailors suffering from scurvy and separated them into groups of two. All 12 received the same diet, except that each group was given a different remedy for scurvy from among those recommended at the time. The sailors given lemons and oranges recovered and returned to duty. The sailors given boiled apple juice improved slightly. The remainder continued to deteriorate. Lind’s *Treatise on the Scurvy* was published in 1753, but inaction persisted in the Royal Navy for another 40 years.



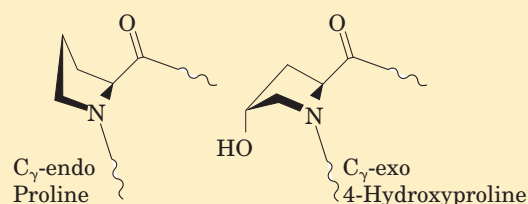
**FIGURE 2** James Lind, 1716–1794; naval surgeon, 1739–1748.



In 1795 the British admiralty finally mandated a ration of concentrated lime or lemon juice for all British sailors (hence the name “limeys”). Scurvy continued to be a problem in some other parts of the world until 1932, when Hungarian scientist Albert Szent-Györgyi, and W. A. Waugh and C. G. King at the University of Pittsburgh, isolated and synthesized ascorbic acid.

L-Ascorbic acid (vitamin C) is a white, odorless, crystalline powder. It is freely soluble in water and relatively insoluble in organic solvents. In a dry state, away from light, it is stable for a considerable length of time. The appropriate daily intake of this vitamin is still in dispute. The recommended daily allowance in the United States is 60 mg (Australia and the United Kingdom recommend 30 to 40 mg; Russia recommends 100 mg). Higher doses of vitamin C are sometimes recommended, although the benefit of such a regimen is disputed. Notably, animals that synthesize their own vitamin C maintain levels found in humans only if they consume hundreds of times the recommended daily allowance. Along with citrus fruits and almost all other fresh fruits, other good sources of vitamin C include peppers, tomatoes, potatoes, and broccoli. The vitamin C of fruits and vegetables is destroyed by overcooking or prolonged storage.

So why is ascorbate so necessary to good health? Of particular interest to us here is its role in the formation of collagen. The proline derivative 4(R)-L-hydroxyproline (4-Hyp) plays an essential role in the folding of collagen and in maintaining its structure. As noted in the text, collagen is constructed of the repeating tripeptide unit Gly–X–Y, where X and Y are generally Pro or 4-Hyp. A constructed peptide with 10 Gly–Pro–Pro repeats will fold to form a collagen triple helix, but the structure melts at 41 °C. If the 10 repeats are changed to Gly–Pro–4-Hyp, the melting temperature jumps to 69 °C. The stability of collagen arises from the detailed structure of the collagen helix, determined independently by Helen Berman and Adriana Zagari and their colleagues. The proline ring is normally found as a mixture of two puckered conformations, called C<sub>γ</sub>-endo and C<sub>γ</sub>-exo (Fig. 3). The collagen helix structure requires the Pro residue in the Y positions to be in the C<sub>γ</sub>-exo conformation, and it is this conformation that is enforced by the hydroxyl substitution at C-4 in 4-hydroxyproline. However, the collagen structure requires the Pro residue in the X positions to have the C<sub>γ</sub>-endo conformation, and introduction of 4-Hyp here can destabilize the helix. The inability to hydroxylate the Pro at the Y positions when vitamin C is absent leads to collagen instability and the connective tissue problems seen in scurvy.



**FIGURE 3** The C<sub>γ</sub>-endo conformation of proline and the C<sub>γ</sub>-exo conformation of 4-hydroxyproline.

The hydroxylation of specific Pro residues in procollagen, the precursor of collagen, requires the action of the enzyme prolyl 4-hydroxylase. This enzyme ( $M_r$  240,000) is an  $\alpha_2\beta_2$  tetramer in all vertebrate sources. The proline-hydroxylating activity is found in the  $\alpha$  subunits. (Researchers were surprised to find that the  $\beta$  subunits are identical to the enzyme protein disulfide isomerase (PDI; p. 152); these subunits do not participate in the prolyl hydroxylation activity.) Each  $\alpha$  subunit contains one atom of nonheme iron ( $\text{Fe}^{2+}$ ), and the enzyme is one of a class of hydroxylases that require  $\alpha$ -ketoglutarate in their reactions.

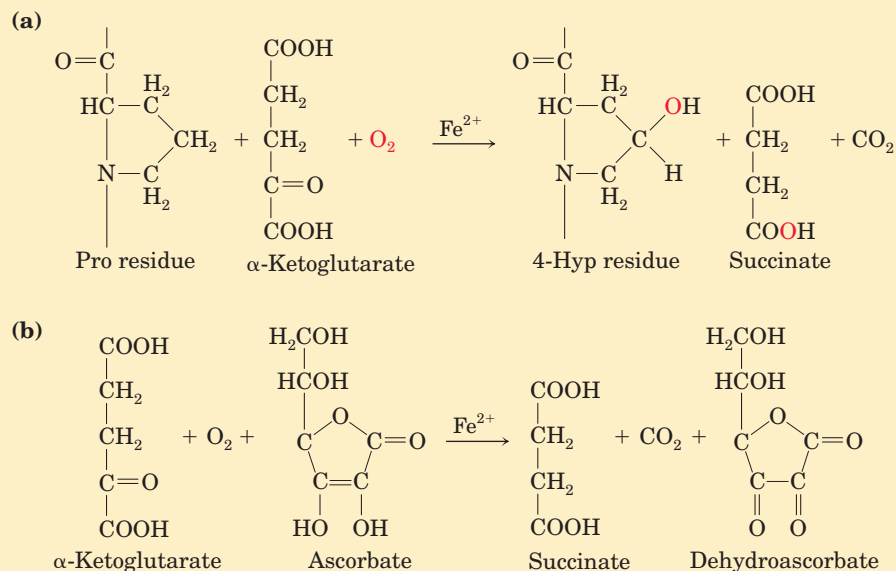
In the normal prolyl 4-hydroxylase reaction (Fig. 4a), one molecule of  $\alpha$ -ketoglutarate and one of  $\text{O}_2$  bind to the enzyme. The  $\alpha$ -ketoglutarate is oxidatively decarboxylated to form  $\text{CO}_2$  and succinate. The remaining oxygen atom is then used to hydroxylate an appropriate Pro residue in procollagen. No ascorbate is needed in this reaction. However, prolyl 4-hydroxylase also catalyzes an oxidative decarboxylation of  $\alpha$ -ketoglutarate that is not coupled to proline hydroxylation—and this is the reaction that requires ascorbate (Fig. 4b). During this reaction, the heme  $\text{Fe}^{2+}$  becomes oxidized, and the oxidized form of the enzyme is inactive—unable to hydroxylate proline. The ascorbate consumed in the reaction presumably functions to reduce the heme iron and restore enzyme activity.

But there is more to the vitamin C story than proline hydroxylation. Very similar hydroxylation reactions generate the less abundant 3-hydroxyproline and 5-hydroxylysine residues that also occur in collagen. The enzymes that catalyze these reactions are members of the same  $\alpha$ -ketoglutarate-dependent dioxygenase family, and for all these enzymes ascorbate plays the same role. These dioxygenases are just a few of the dozens of closely related enzymes that play a variety of metabolic roles in different classes of organisms. Ascorbate serves other roles too. It is an antioxidant, reacting enzymatically and nonenzymatically with reactive oxygen species, which in mammals play an important role in aging and cancer.

(continued on next page)



## BOX 4-3 BIOCHEMISTRY IN MEDICINE (continued from previous page)



**FIGURE 4** The reactions catalyzed by prolyl 4-hydroxylase. **(a)** The normal reaction, coupled to proline hydroxylation, which does not require ascorbate. The fate of the two oxygen atoms from O<sub>2</sub> is shown in red. **(b)** The uncoupled reaction, in which  $\alpha$ -ketoglutarate is oxidatively decarboxylated without hydroxylation of proline. Ascorbate is consumed stoichiometrically in this process as it is converted to dehydroascorbate.

In plants, ascorbate is required as a substrate for the enzyme ascorbate peroxidase, which converts H<sub>2</sub>O<sub>2</sub> to water. The peroxide is generated from the O<sub>2</sub> produced in photosynthesis, an unavoidable consequence of generating O<sub>2</sub> in a compartment laden with powerful oxidation-reduction systems (Chapter 19). Ascorbate is also a precursor of oxalate and tartrate in plants, and is involved in the hydroxylation of Pro residues in cell wall proteins called extensins. Ascorbate is found in all subcellular compartments of plants, at concentrations of 2 to 25 mM—which is why plants are such good sources of vitamin C.

Scurvy remains a problem today. The malady is still encountered not only in remote regions where nutritious food is scarce but, surprisingly, on U.S. college campuses. The only vegetables consumed by some students are those in tossed salads, and days go by without these young adults consuming fruit. A 1998 study of 230 students at Arizona State University revealed that 10% had serious vitamin C deficiencies, and 2 students had vitamin C levels so low that they probably had scurvy. Only half the students in the study consumed the recommended daily allowance of vitamin C.

Eat your fresh fruit and vegetables.

to function, and even the evolutionary paths by which proteins arrived at their present state, which can be glimpsed in the family resemblances that are revealed as protein databases are sifted and sorted. The sheer variety of structures can seem daunting. Yet as new protein structures become available it is becoming increasingly clear that they are manifestations of a finite set of recognizable, stable folding patterns.

Our discussion of globular protein structure begins with the principles gleaned from the earliest protein structures to be elucidated. This is followed by a detailed description of protein substructure and comparative categorization. Such discussions are possible only because of the vast amount of information available over the Internet from resources such as the Protein Data Bank (PDB; [www.rcsb.org/pdb](http://www.rcsb.org/pdb)), an archive of experimentally determined three-dimensional structures of biological macromolecules.

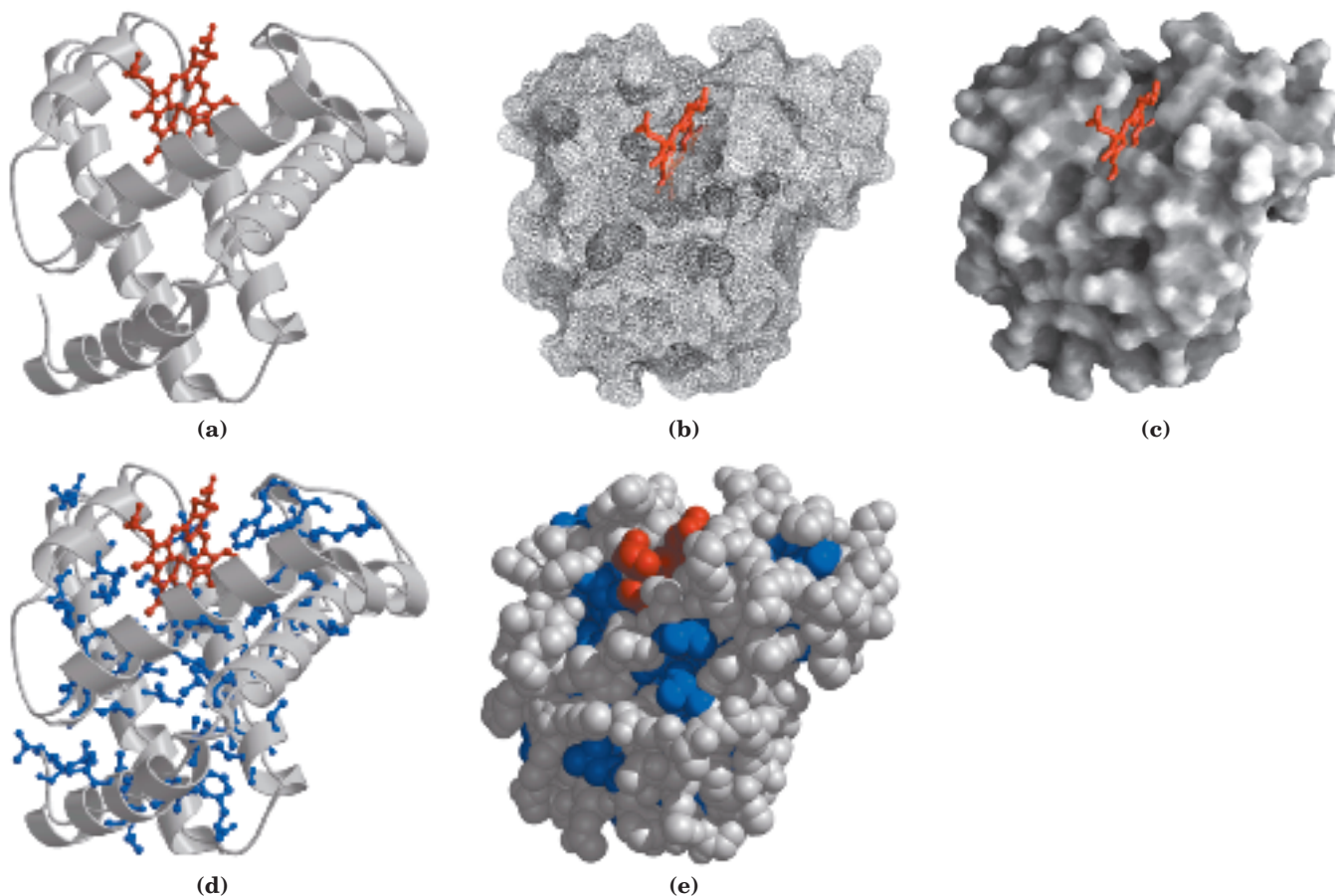
### Myoglobin Provided Early Clues about the Complexity of Globular Protein Structure

**Protein Architecture—Tertiary Structure of Small Globular Proteins, II. Myoglobin** The first breakthrough in understanding the three-dimensional structure of a globular protein came from x-ray diffraction studies of myoglobin carried out by John Kendrew and his colleagues in the 1950s. Myoglobin is a relatively small ( $M_r$  16,700), oxygen-binding protein of muscle cells. It functions both to store oxygen and to facilitate oxygen diffusion in rapidly contracting muscle tissue. Myoglobin contains a single polypeptide chain of 153 amino acid residues of known sequence and a single iron protoporphyrin, or heme, group. The same heme group is found in hemoglobin, the oxygen-binding protein of erythrocytes, and is responsible for the deep red-brown color of both myoglobin and hemoglobin. Myoglobin is particularly abun-

dant in the muscles of diving mammals such as the whale, seal, and porpoise, whose muscles are so rich in this protein that they are brown. Storage and distribution of oxygen by muscle myoglobin permit these animals to remain submerged for long periods of time. The activities of myoglobin and other globin molecules are investigated in greater detail in Chapter 5.

Figure 4–16 shows several structural representations of myoglobin, illustrating how the polypeptide chain is folded in three dimensions—its tertiary structure. The red group surrounded by protein is heme. The backbone of the myoglobin molecule is made up of eight relatively straight segments of  $\alpha$  helix interrupted by bends, some of which are  $\beta$  turns. The longest  $\alpha$  helix has 23 amino acid residues and the shortest only 7; all helices are right-handed. More than 70% of the residues in myoglobin are in these  $\alpha$ -helical regions. X-ray analysis has revealed the precise position of each of the R groups, which occupy nearly all the space within the folded chain.

Many important conclusions were drawn from the structure of myoglobin. The positioning of amino acid side chains reflects a structure that derives much of its stability from hydrophobic interactions. Most of the hydrophobic R groups are in the interior of the myoglobin molecule, hidden from exposure to water. All but two of the polar R groups are located on the outer surface of the molecule, and all are hydrated. The myoglobin molecule is so compact that its interior has room for only four molecules of water. This dense hydrophobic core is typical of globular proteins. The fraction of space occupied by atoms in an organic liquid is 0.4 to 0.6; in a typical crystal the fraction is 0.70 to 0.78, near the theoretical maximum. In a globular protein the fraction is about 0.75, comparable to that in a crystal. In this packed environment, weak interactions strengthen and reinforce each other. For example, the nonpolar side chains in the core are so close together that short-range van der Waals interactions make a significant contribution to stabilizing hydrophobic interactions.



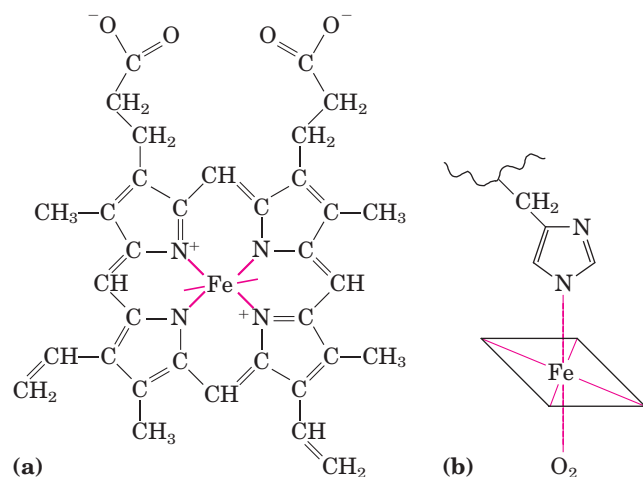
**FIGURE 4-16** Tertiary structure of sperm whale myoglobin. (PDB ID 1MBO) The orientation of the protein is similar in all panels; the heme group is shown in red. In addition to illustrating the myoglobin structure, this figure provides examples of several different ways to display protein structure. **(a)** The polypeptide backbone, shown in a ribbon representation of a type introduced by Jane Richardson, which highlights regions of secondary structure. The  $\alpha$ -helical regions are evident. **(b)** A “mesh” image emphasizes the protein surface. **(c)** A sur-

face contour image is useful for visualizing pockets in the protein where other molecules might bind. **(d)** A ribbon representation, including side chains (blue) for the hydrophobic residues Leu, Ile, Val, and Phe. **(e)** A space-filling model with all amino acid side chains. Each atom is represented by a sphere encompassing its van der Waals radius. The hydrophobic residues are again shown in blue; most are not visible, because they are buried in the interior of the protein.

Deduction of the structure of myoglobin confirmed some expectations and introduced some new elements of secondary structure. As predicted by Pauling and Corey, all the peptide bonds are in the planar trans configuration. The  $\alpha$  helices in myoglobin provided the first direct experimental evidence for the existence of this type of secondary structure. Three of the four Pro residues of myoglobin are found at bends (recall that proline, with its fixed  $\phi$  bond angle and lack of a peptide-bond N—H group for participation in hydrogen bonds, is largely incompatible with  $\alpha$ -helical structure). The fourth Pro residue occurs within an  $\alpha$  helix, where it creates a kink necessary for tight helix packing. Other bends contain Ser, Thr, and Asn residues, which are among the amino acids whose bulk and shape tend to make them incompatible with  $\alpha$ -helical structure if they are in close proximity in the amino acid sequence (p. 121).

The flat heme group rests in a crevice, or pocket, in the myoglobin molecule. The iron atom in the center of the heme group has two bonding (coordination) positions perpendicular to the plane of the heme (Fig. 4-17). One of these is bound to the R group of the His residue at position 93; the other is the site at which an  $O_2$  molecule binds. Within this pocket, the accessibility of the heme group to solvent is highly restricted. This is important for function, because free heme groups in an oxygenated solution are rapidly oxidized from the ferrous ( $Fe^{2+}$ ) form, which is active in the reversible binding of  $O_2$ , to the ferric ( $Fe^{3+}$ ) form, which does not bind  $O_2$ .

Knowledge of the structure of myoglobin allowed researchers for the first time to understand in detail the



**FIGURE 4-17 The heme group.** This group is present in myoglobin, hemoglobin, cytochromes, and many other heme proteins. (a) Heme consists of a complex organic ring structure, protoporphyrin, to which is bound an iron atom in its ferrous ( $Fe^{2+}$ ) state. The iron atom has six coordination bonds, four in the plane of, and bonded to, the flat porphyrin molecule and two perpendicular to it. (b) In myoglobin and hemoglobin, one of the perpendicular coordination bonds is bound to a nitrogen atom of a His residue. The other is “open” and serves as the binding site for an  $O_2$  molecule.

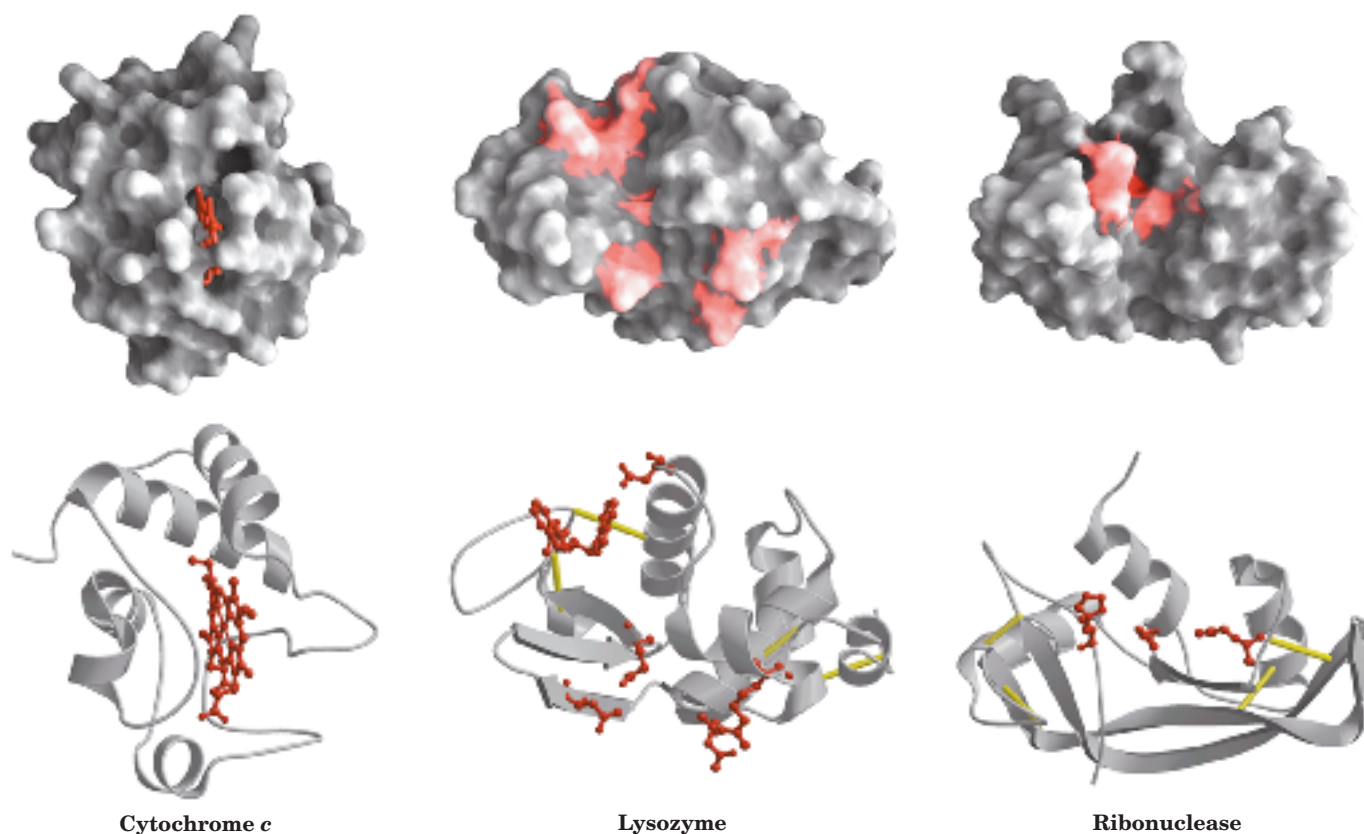
correlation between the structure and function of a protein. Many different myoglobin structures have been elucidated, allowing investigators to see how the structure changes when oxygen or other molecules bind to it. Hundreds of proteins have been subjected to similar analysis since then. Today, techniques such as NMR spectroscopy supplement x-ray diffraction data, providing more information on a protein’s structure (Box 4-4). The ongoing sequencing of genomic DNA from many organisms (Chapter 9) has identified thousands of genes that encode proteins of known sequence but unknown function. Our first insight into what these proteins do often comes from our still-limited understanding of how primary structure determines tertiary structure, and how tertiary structure determines function.

### Globular Proteins Have a Variety of Tertiary Structures

With elucidation of the tertiary structures of hundreds of other globular proteins by x-ray analysis, it became clear that myoglobin illustrates only one of many ways in which a polypeptide chain can be folded. In Figure 4-18 the structures of cytochrome *c*, lysozyme, and ribonuclease are compared. These proteins have different amino acid sequences and different tertiary structures, reflecting differences in function. All are relatively small and easy to work with, facilitating structural analysis. Cytochrome *c* is a component of the respiratory chain of mitochondria (Chapter 19). Like myoglobin, cytochrome *c* is a heme protein. It contains a single polypeptide chain of about 100 residues ( $M_r$  12,400) and a single heme group. In this case, the protoporphyrin of the heme group is covalently attached to the polypeptide. Only about 40% of the polypeptide is in  $\alpha$ -helical segments, compared with 70% of the myoglobin chain. The rest of the cytochrome *c* chain contains  $\beta$  turns and irregularly coiled and extended segments.

Lysozyme ( $M_r$  14,600) is an enzyme abundant in egg white and human tears that catalyzes the hydrolytic cleavage of polysaccharides in the protective cell walls of some families of bacteria. Lysozyme, because it can lyse, or degrade, bacterial cell walls, serves as a bactericidal agent. As in cytochrome *c*, about 40% of its 129 amino acid residues are in  $\alpha$ -helical segments, but the arrangement is different and some  $\beta$ -sheet structure is also present (Fig. 4-18). Four disulfide bonds contribute stability to this structure. The  $\alpha$  helices line a long crevice in the side of the molecule, called the active site, which is the site of substrate binding and catalysis. The bacterial polysaccharide that is the substrate for lysozyme fits into this crevice. **Protein Architecture—Tertiary Structure of Small Globular Proteins, III. Lysozyme**

Ribonuclease, another small globular protein ( $M_r$  13,700), is an enzyme secreted by the pancreas into the small intestine, where it catalyzes the hydrolysis of certain bonds in the ribonucleic acids present in ingested



**FIGURE 4-18** Three-dimensional structures of some small proteins.

Shown here are cytochrome *c* (PDB ID 1CCR), lysozyme (PDB ID 3LYM), and ribonuclease (PDB ID 3RN3). Each protein is shown in surface contour and in a ribbon representation, in the same orientation. In the ribbon depictions, regions in the  $\beta$  conformation are

food. Its tertiary structure, determined by x-ray analysis, shows that little of its 124 amino acid polypeptide chain is in an  $\alpha$ -helical conformation, but it contains many segments in the  $\beta$  conformation (Fig. 4-18). Like lysozyme, ribonuclease has four disulfide bonds between loops of the polypeptide chain.

In small proteins, hydrophobic residues are less likely to be sheltered in a hydrophobic interior—simple geometry dictates that the smaller the protein, the lower the ratio of volume to surface area. Small proteins also have fewer potential weak interactions available to stabilize them. This explains why many smaller proteins such as those in Figure 4-18 are stabilized by a number of covalent bonds. Lysozyme and ribonuclease, for example, have disulfide linkages, and the heme group in cytochrome *c* is covalently linked to the protein on two sides, providing significant stabilization of the entire protein structure.

Table 4-2 shows the proportions of  $\alpha$  helix and  $\beta$  conformation (expressed as percentage of residues in each secondary structure) in several small, single-chain, globular proteins. Each of these proteins has a distinct structure, adapted for its particular biological function, but together they share several important properties. Each is folded compactly, and in each case the hydro-

phobic amino acid side chains are oriented toward the interior (away from water) and the hydrophilic side chains are on the surface. The structures are also stabilized by a multitude of hydrogen bonds and some ionic interactions.

**TABLE 4-2** Approximate Amounts of  $\alpha$  Helix and  $\beta$  Conformation in Some Single-Chain Proteins

Protein (total residues)	Residues (%)*	
	$\alpha$ Helix	$\beta$ Conformation
Chymotrypsin (247)	14	45
Ribonuclease (124)	26	35
Carboxypeptidase (307)	38	17
Cytochrome <i>c</i> (104)	39	0
Lysozyme (129)	40	12
Myoglobin (153)	78	0

Source: Data from Cantor, C.R. & Schimmel, P.R. (1980) *Biophysical Chemistry, Part I: The Conformation of Biological Macromolecules*, p. 100, W. H. Freeman and Company, New York.

\*Portions of the polypeptide chains that are not accounted for by  $\alpha$  helix or  $\beta$  conformation consist of bends and irregularly coiled or extended stretches. Segments of  $\alpha$  helix and  $\beta$  conformation sometimes deviate slightly from their normal dimensions and geometry.



## BOX 4-4 WORKING IN BIOCHEMISTRY

**Methods for Determining the Three-Dimensional Structure of a Protein****X-Ray Diffraction**

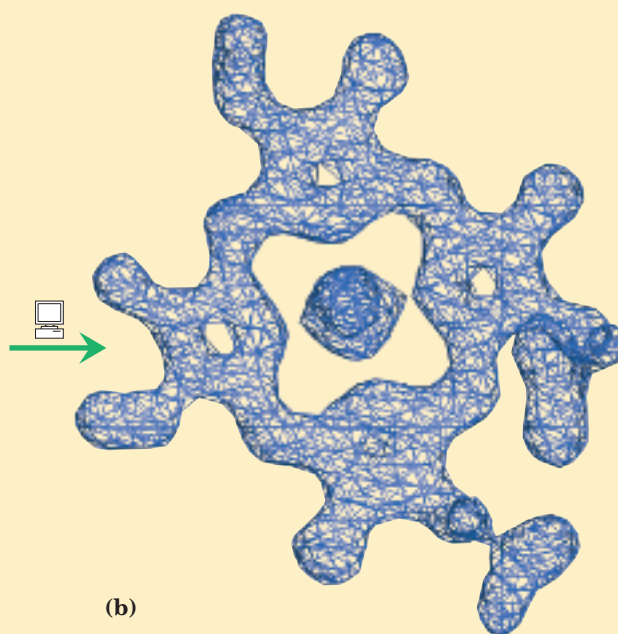
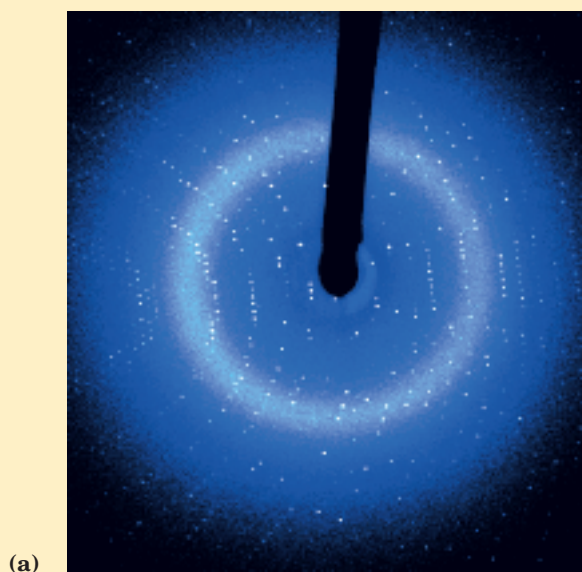
The spacing of atoms in a crystal lattice can be determined by measuring the locations and intensities of spots produced on photographic film by a beam of x rays of given wavelength, after the beam has been diffracted by the electrons of the atoms. For example, x-ray analysis of sodium chloride crystals shows that  $\text{Na}^+$  and  $\text{Cl}^-$  ions are arranged in a simple cubic lattice. The spacing of the different kinds of atoms in complex organic molecules, even very large ones such as proteins, can also be analyzed by x-ray diffraction methods. However, the technique for analyzing crystals of complex molecules is far more laborious than for simple salt crystals. When the repeating pattern of the crystal is a molecule as large as, say, a protein, the numerous atoms in the molecule yield thousands of diffraction spots that must be analyzed by computer.

The process may be understood at an elementary level by considering how images are generated in a light microscope. Light from a point source is focused on an object. The light waves are scattered by the object, and these scattered waves are recombined by a series of lenses to generate an enlarged image of the object. The smallest object whose structure can be determined by such a system—that is, the resolving power of the microscope—is determined by the wavelength of the light, in this case visible light, with

wavelengths in the range of 400 to 700 nm. Objects smaller than half the wavelength of the incident light cannot be resolved. To resolve objects as small as proteins we must use x rays, with wavelengths in the range of 0.7 to 1.5 Å (0.07 to 0.15 nm). However, there are no lenses that can recombine x rays to form an image; instead the pattern of diffracted x rays is collected directly and an image is reconstructed by mathematical techniques.

The amount of information obtained from x-ray crystallography depends on the degree of structural order in the sample. Some important structural parameters were obtained from early studies of the diffraction patterns of the fibrous proteins arranged in fairly regular arrays in hair and wool. However, the orderly bundles formed by fibrous proteins are not crystals—the molecules are aligned side by side, but not all are oriented in the same direction. More detailed three-dimensional structural information about proteins requires a highly ordered protein crystal. Protein crystallization is something of an empirical science, and the structures of many important proteins are not yet known, simply because they have proved difficult to crystallize. Practitioners have compared making protein crystals to holding together a stack of bowling balls with cellophane tape.

Operationally, there are several steps in x-ray structural analysis (Fig. 1). Once a crystal is obtained, it is placed in an x-ray beam between the x-ray source and a detector, and a regular array of spots called re-



flections is generated. The spots are created by the diffracted x-ray beam, and each atom in a molecule makes a contribution to each spot. An electron-density map of the protein is reconstructed from the overall diffraction pattern of spots by using a mathematical technique called a Fourier transform. In effect, the computer acts as a “computational lens.” A model for the structure is then built that is consistent with the electron-density map.

John Kendrew found that the x-ray diffraction pattern of crystalline myoglobin (isolated from muscles of the sperm whale) is very complex, with nearly 25,000 reflections. Computer analysis of these reflections took place in stages. The resolution improved at each stage, until in 1959 the positions of virtually all the non-hydrogen atoms in the protein had been determined. The amino acid sequence of the protein, obtained by chemical analysis, was consistent with the molecular model. The structures of thousands of proteins, many of them much more complex than myoglobin, have since been determined to a similar level of resolution.

The physical environment within a crystal, of course, is not identical to that in solution or in a living cell. A crystal imposes a space and time average on the structure deduced from its analysis, and x-ray diffraction studies provide little information about molecular motion within the protein. The conformation of proteins in a crystal could in principle also be affected by nonphysiological factors such as incidental

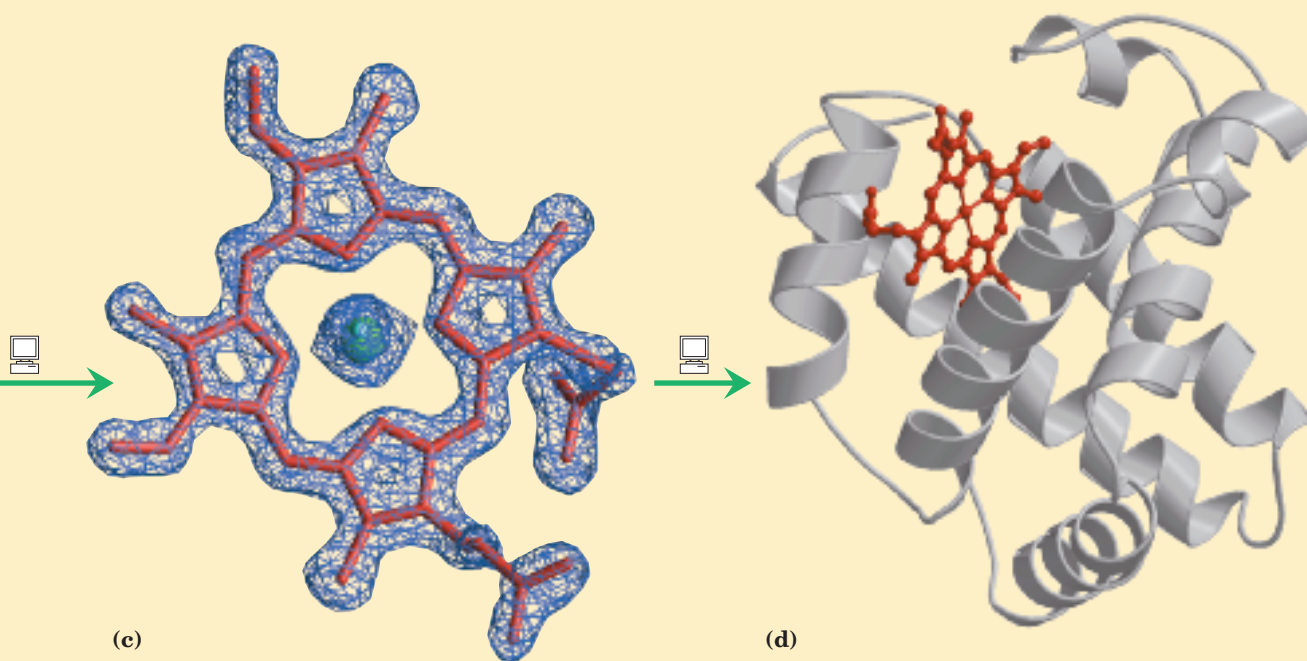
protein-protein contacts within the crystal. However, when structures derived from the analysis of crystals are compared with structural information obtained by other means (such as NMR, as described below), the crystal-derived structure almost always represents a functional conformation of the protein. X-ray crystallography can be applied successfully to proteins too large to be structurally analyzed by NMR.

### Nuclear Magnetic Resonance

An important complementary method for determining the three-dimensional structures of macromolecules is nuclear magnetic resonance (NMR). Modern NMR techniques are being used to determine the structures of ever-larger macromolecules, including carbohydrates, nucleic acids, and small to average-sized proteins. An advantage of NMR studies is that they are

*(continued on next page)*

**FIGURE 1** Steps in the determination of the structure of sperm whale myoglobin by x-ray crystallography. **(a)** X-ray diffraction patterns are generated from a crystal of the protein. **(b)** Data extracted from the diffraction patterns are used to calculate a three-dimensional electron-density map of the protein. The electron density of only part of the structure, the heme, is shown. **(c)** Regions of greatest electron density reveal the location of atomic nuclei, and this information is used to piece together the final structure. Here, the heme structure is modeled into its electron-density map. **(d)** The completed structure of sperm whale myoglobin, including the heme (PDB ID 2MBW).



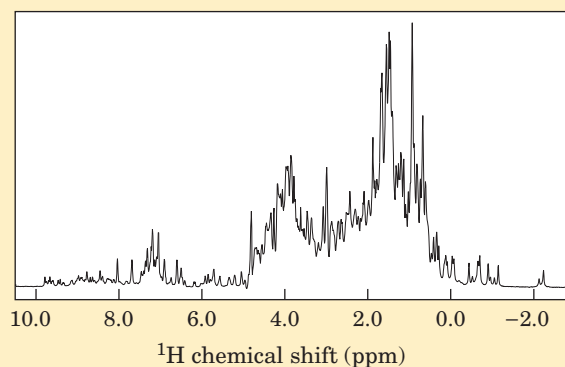
## BOX 4-4 WORKING IN BIOCHEMISTRY (continued from previous page)

carried out on macromolecules in solution, whereas x-ray crystallography is limited to molecules that can be crystallized. NMR can also illuminate the dynamic side of protein structure, including conformational changes, protein folding, and interactions with other molecules.

NMR is a manifestation of nuclear spin angular momentum, a quantum mechanical property of atomic nuclei. Only certain atoms, including  $^1\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$ ,  $^{19}\text{F}$ , and  $^{31}\text{P}$ , possess the kind of nuclear spin that gives rise to an NMR signal. Nuclear spin generates a magnetic dipole. When a strong, static magnetic field is applied to a solution containing a single type of macromolecule, the magnetic dipoles are aligned in the field in one of two orientations, parallel (low energy) or antiparallel (high energy). A short ( $\sim 10 \mu\text{s}$ ) pulse of electromagnetic energy of suitable frequency (the resonant frequency, which is in the radio frequency range) is applied at right angles to the nuclei aligned in the magnetic field. Some energy is absorbed as nuclei switch to the high-energy state, and the absorption spectrum that results contains information about the identity of the nuclei and their immediate chemical environment. The data from many such experiments performed on a sample are averaged, increasing the signal-to-noise ratio, and an NMR spectrum such as that in Figure 2 is generated.

$^1\text{H}$  is particularly important in NMR experiments because of its high sensitivity and natural abundance. For macromolecules,  $^1\text{H}$  NMR spectra can become quite complicated. Even a small protein has hundreds of  $^1\text{H}$  atoms, typically resulting in a one-dimensional NMR spectrum too complex for analysis. Structural analysis of proteins became possible with the advent of two-dimensional NMR techniques (Fig. 3). These methods allow measurement of distance-dependent coupling of nuclear spins in nearby atoms through space (the nuclear Overhauser effect (NOE), in a method dubbed NOESY) or the coupling of nuclear spins in atoms connected by covalent bonds (total correlation spectroscopy, or TOCSY).

Translating a two-dimensional NMR spectrum into a complete three-dimensional structure can be a laborious process. The NOE signals provide some information about the distances between individual atoms, but



**FIGURE 2** A one-dimensional NMR spectrum of a globin from a marine blood worm. This protein and sperm whale myoglobin are very close structural analogs, belonging to the same protein structural family and sharing an oxygen-transport function.

for these distance constraints to be useful, the atoms giving rise to each signal must be identified. Complementary TOCSY experiments can help identify which NOE signals reflect atoms that are linked by covalent bonds. Certain patterns of NOE signals have been associated with secondary structures such as  $\alpha$  helices. Modern genetic engineering (Chapter 9) can be used to prepare proteins that contain the rare isotopes  $^{13}\text{C}$  or  $^{15}\text{N}$ . The new NMR signals produced by these atoms, and the coupling with  $^1\text{H}$  signals resulting from these substitutions, help in the assignment of individual  $^1\text{H}$  NOE signals. The process is also aided by a knowledge of the amino acid sequence of the polypeptide.

To generate a three-dimensional structure, researchers feed the distance constraints into a computer along with known geometric constraints such as chirality, van der Waals radii, and bond lengths and angles. The computer generates a family of closely related structures that represent the range of conformations consistent with the NOE distance constraints (Fig. 3c). The uncertainty in structures generated by NMR is in part a reflection of the molecular vibrations (breathing) within a protein structure in solution, discussed in more detail in Chapter 5. Normal experimental uncertainty can also play a role.

When a protein structure has been determined by both x-ray crystallography and NMR, the structures

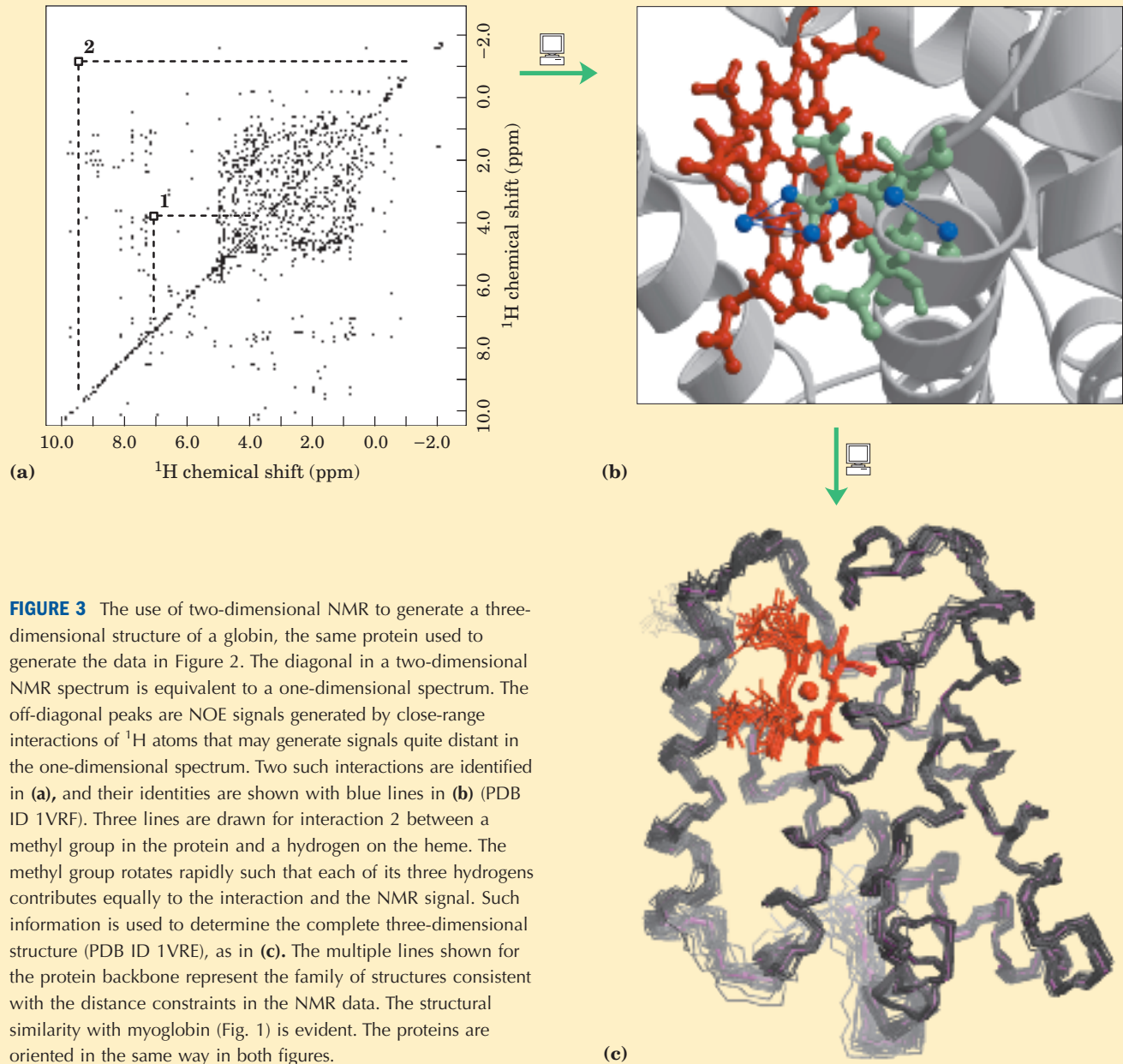
### Analysis of Many Globular Proteins Reveals Common Structural Patterns

**Protein Architecture—Tertiary Structure of Large Globular Proteins** For the beginning student, the very complex tertiary structures of globular proteins much larger than those shown in Figure 4-18 are best approached by fo-

cus on structural patterns that recur in different and often unrelated proteins. The three-dimensional structure of a typical globular protein can be considered an assemblage of polypeptide segments in the  $\alpha$ -helix and  $\beta$ -sheet conformations, linked by connecting segments. The structure can then be described to a first approximation by defining how these segments stack on one

generally agree well. In some cases, the precise locations of particular amino acid side chains on the protein exterior are different, often because of effects related to the packing of adjacent protein molecules in

a crystal. The two techniques together are at the heart of the rapid increase in the availability of structural information about the macromolecules of living cells.



**FIGURE 3** The use of two-dimensional NMR to generate a three-dimensional structure of a globin, the same protein used to generate the data in Figure 2. The diagonal in a two-dimensional NMR spectrum is equivalent to a one-dimensional spectrum. The off-diagonal peaks are NOE signals generated by close-range interactions of  $^1\text{H}$  atoms that may generate signals quite distant in the one-dimensional spectrum. Two such interactions are identified in **(a)**, and their identities are shown with blue lines in **(b)** (PDB ID 1VRF). Three lines are drawn for interaction 2 between a methyl group in the protein and a hydrogen on the heme. The methyl group rotates rapidly such that each of its three hydrogens contributes equally to the interaction and the NMR signal. Such information is used to determine the complete three-dimensional structure (PDB ID 1VRE), as in **(c)**. The multiple lines shown for the protein backbone represent the family of structures consistent with the distance constraints in the NMR data. The structural similarity with myoglobin (Fig. 1) is evident. The proteins are oriented in the same way in both figures.

another and how the segments that connect them are arranged. This formalism has led to the development of databases that allow informative comparisons of protein structures, complementing other databases that permit comparisons of protein sequences.

An understanding of a complete three-dimensional structure is built upon an analysis of its parts. We begin

by defining terms used to describe protein substructures, then turn to the folding rules elucidated from analysis of the structures of many proteins.

**Supersecondary structures**, also called **motifs** or simply **folds**, are particularly stable arrangements of several elements of secondary structure and the connections between them. There is no universal agreement

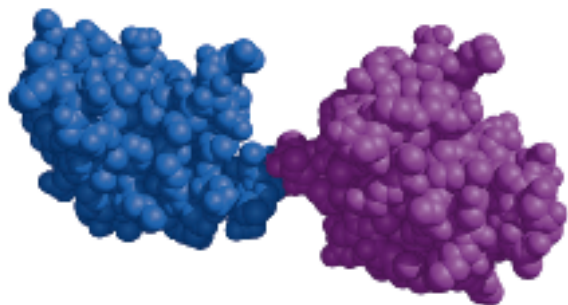


among biochemists on the application of the three terms, and they are often used interchangeably. The terms are also applied to a wide range of structures. Recognized motifs range from simple to complex, sometimes appearing in repeating units or combinations. A single large motif may comprise the entire protein. We have already encountered one well-studied motif, the coiled coil of  $\alpha$ -keratin, also found in a number of other proteins.

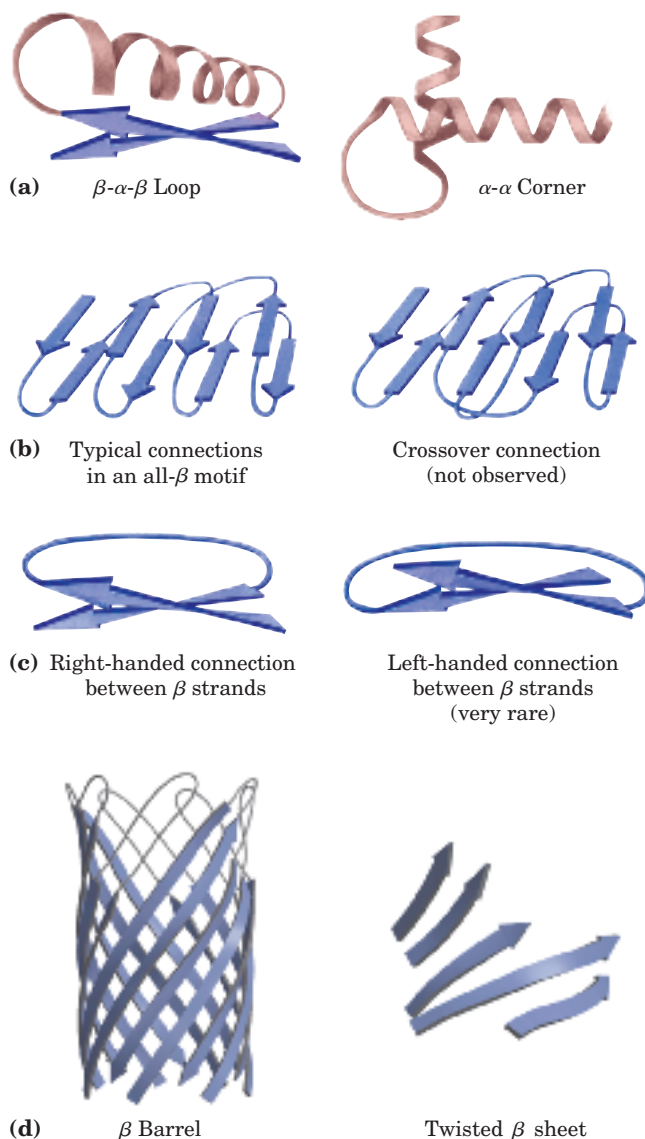
Polypeptides with more than a few hundred amino acid residues often fold into two or more stable, globular units called **domains**. In many cases, a domain from a large protein will retain its correct three-dimensional structure even when it is separated (for example, by proteolytic cleavage) from the remainder of the polypeptide chain. A protein with multiple domains may appear to have a distinct globular lobe for each domain (Fig. 4–19), but, more commonly, extensive contacts between domains make individual domains hard to discern. Different domains often have distinct functions, such as the binding of small molecules or interaction with other proteins. Small proteins usually have only one domain (the domain *is* the protein).

Folding of polypeptides is subject to an array of physical and chemical constraints. A sampling of the prominent folding rules that have emerged provides an opportunity to introduce some simple motifs.

1. Hydrophobic interactions make a large contribution to the stability of protein structures. Burial of hydrophobic amino acid R groups so as to exclude water requires at least two layers of secondary structure. Two simple motifs, the  **$\beta$ - $\alpha$ - $\beta$  loop** and the  **$\alpha$ - $\alpha$  corner** (Fig. 4–20a), create two layers.
2. Where they occur together in proteins,  $\alpha$  helices and  $\beta$  sheets generally are found in different structural layers. This is because the backbone of a polypeptide segment in the  $\beta$  conformation (Fig. 4–7) cannot readily hydrogen-bond to an  $\alpha$  helix aligned with it.



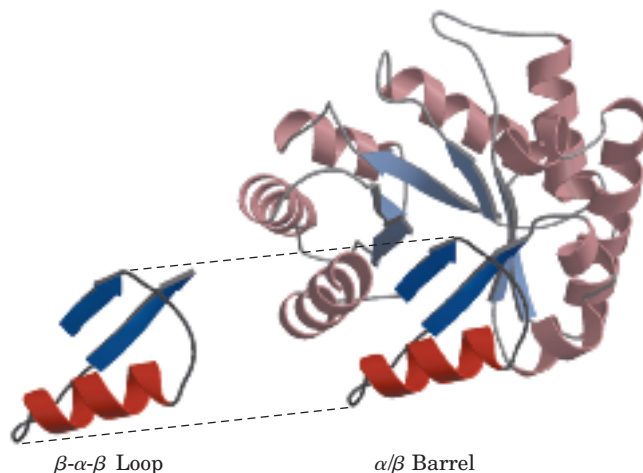
**FIGURE 4-19** Structural domains in the polypeptide troponin C. (PDB ID 4TNC) This calcium-binding protein associated with muscle has separate calcium-binding domains, indicated in blue and purple.



**FIGURE 4-20** Stable folding patterns in proteins. (a) Two simple and common motifs that provide two layers of secondary structure. Amino acid side chains at the interface between elements of secondary structure are shielded from water. Note that the  $\beta$  strands in the  $\beta$ - $\alpha$ - $\beta$  loop tend to twist in a right-handed fashion. (b) Connections between  $\beta$  strands in layered  $\beta$  sheets. The strands are shown from one end, with no twisting included in the schematic. Thick lines represent connections at the ends nearest the viewer; thin lines are connections at the far ends of the  $\beta$  strands. The connections on a given end (e.g., near the viewer) do not cross each other. (c) Because of the twist in  $\beta$  strands, connections between strands are generally right-handed. Left-handed connections must traverse sharper angles and are harder to form. (d) Two arrangements of  $\beta$  strands stabilized by the tendency of the strands to twist. This  $\beta$  barrel is a single domain of  $\alpha$ -hemolysin (a pore-forming toxin that kills a cell by creating a hole in its membrane) from the bacterium *Staphylococcus aureus* (derived from PDB ID 7AHL). The twisted  $\beta$  sheet is from a domain of photolyase (a protein that repairs certain types of DNA damage) from *E. coli* (derived from PDB ID 1DNP).

- Polypeptide segments adjacent to each other in the primary sequence are usually stacked adjacent to each other in the folded structure. Although distant segments of a polypeptide may come together in the tertiary structure, this is not the norm.
- Connections between elements of secondary structure cannot cross or form knots (Fig. 4–20b).
- The  $\beta$  conformation is most stable when the individual segments are twisted slightly in a right-handed sense. This influences both the arrangement of  $\beta$  sheets relative to one another and the path of the polypeptide connection between them. Two parallel  $\beta$  strands, for example, must be connected by a crossover strand (Fig. 4–20c). In principle, this crossover could have a right- or left-handed conformation, but in proteins it is almost always right-handed. Right-handed connections tend to be shorter than left-handed connections and tend to bend through smaller angles, making them easier to form. The twisting of  $\beta$  sheets also leads to a characteristic twisting of the structure formed when many segments are put together. Two examples of resulting structures are the  $\beta$  barrel and twisted  $\beta$  sheet (Fig. 4–20d), which form the core of many larger structures.

Following these rules, complex motifs can be built up from simple ones. For example, a series of  $\beta$ - $\alpha$ - $\beta$  loops, arranged so that the  $\beta$  strands form a barrel, creates a particularly stable and common motif called the  **$\alpha/\beta$  barrel** (Fig. 4–21). In this structure, each parallel  $\beta$  segment is attached to its neighbor by an  $\alpha$ -helical segment. All connections are right-handed. The  $\alpha/\beta$  barrel is found in many enzymes, often with a binding site for a



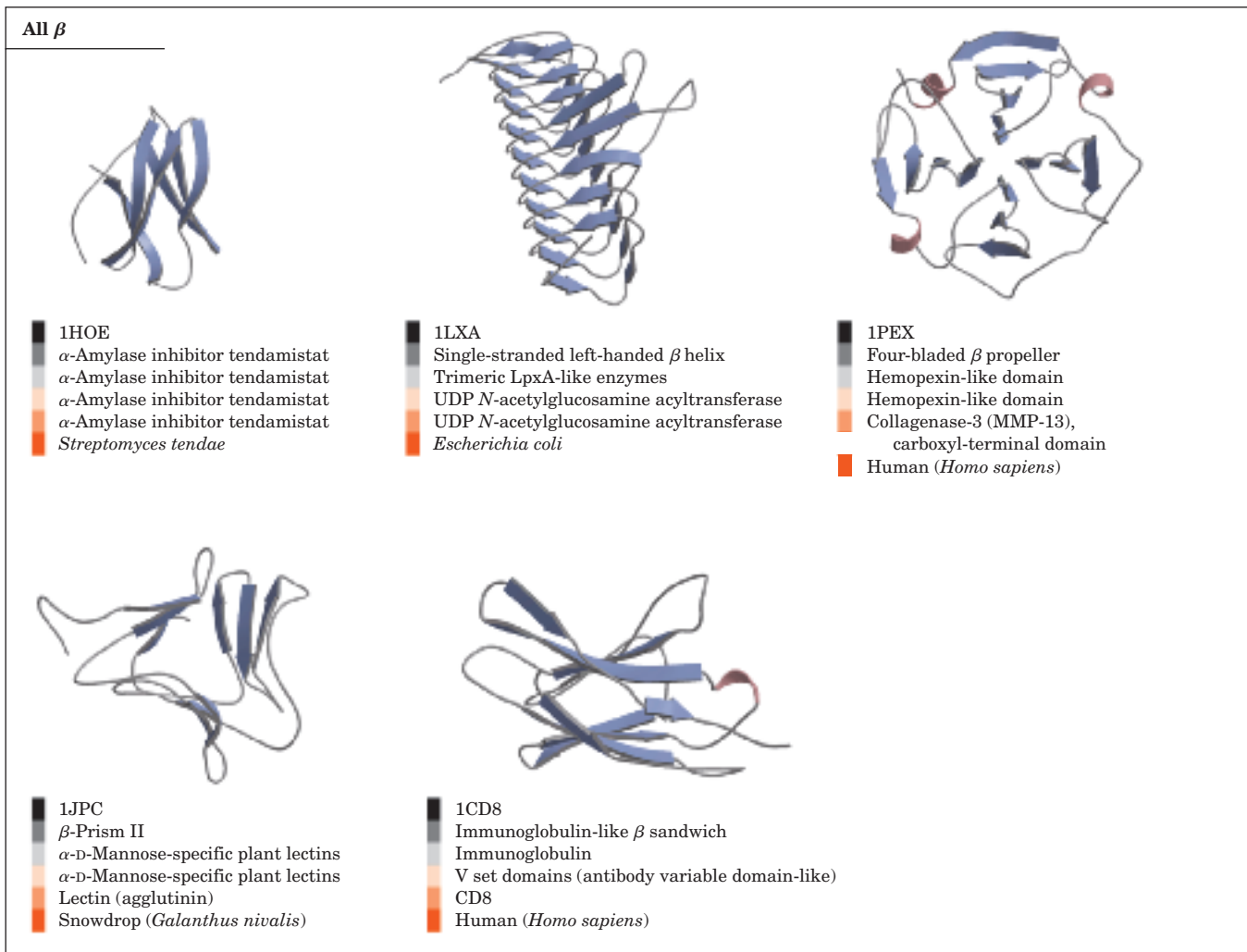
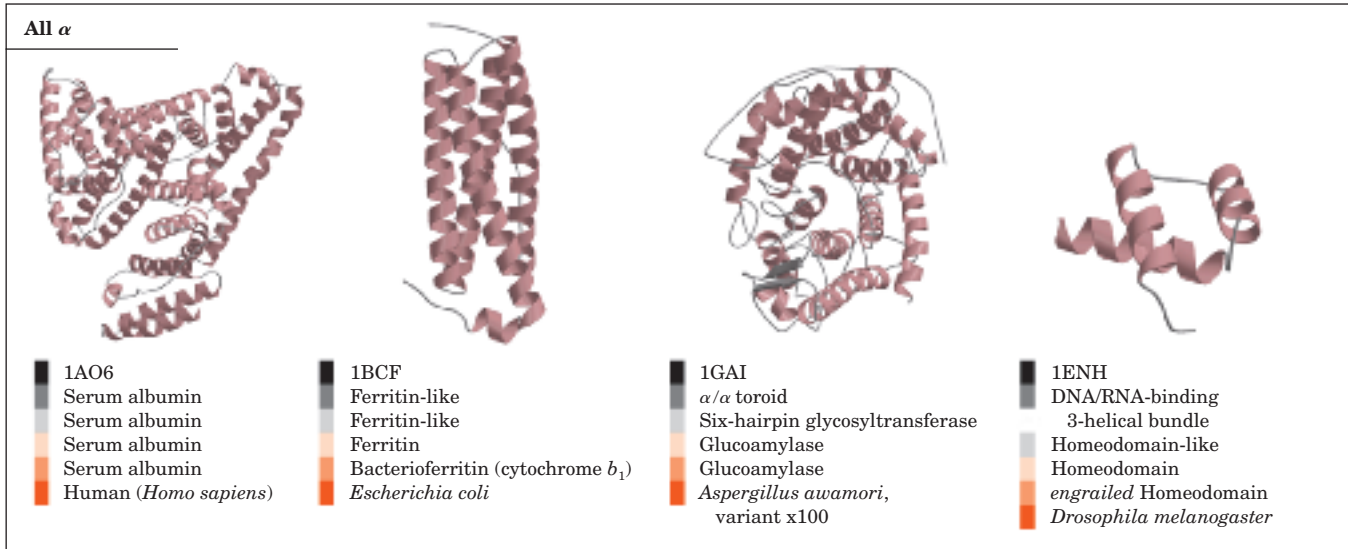
**FIGURE 4-21** Constructing large motifs from smaller ones. The  $\alpha/\beta$  barrel is a common motif constructed from repetitions of the simpler  $\beta$ - $\alpha$ - $\beta$  loop motif. This  $\alpha/\beta$  barrel is a domain of the pyruvate kinase (a glycolytic enzyme) from rabbit (derived from PDB ID 1PKN).

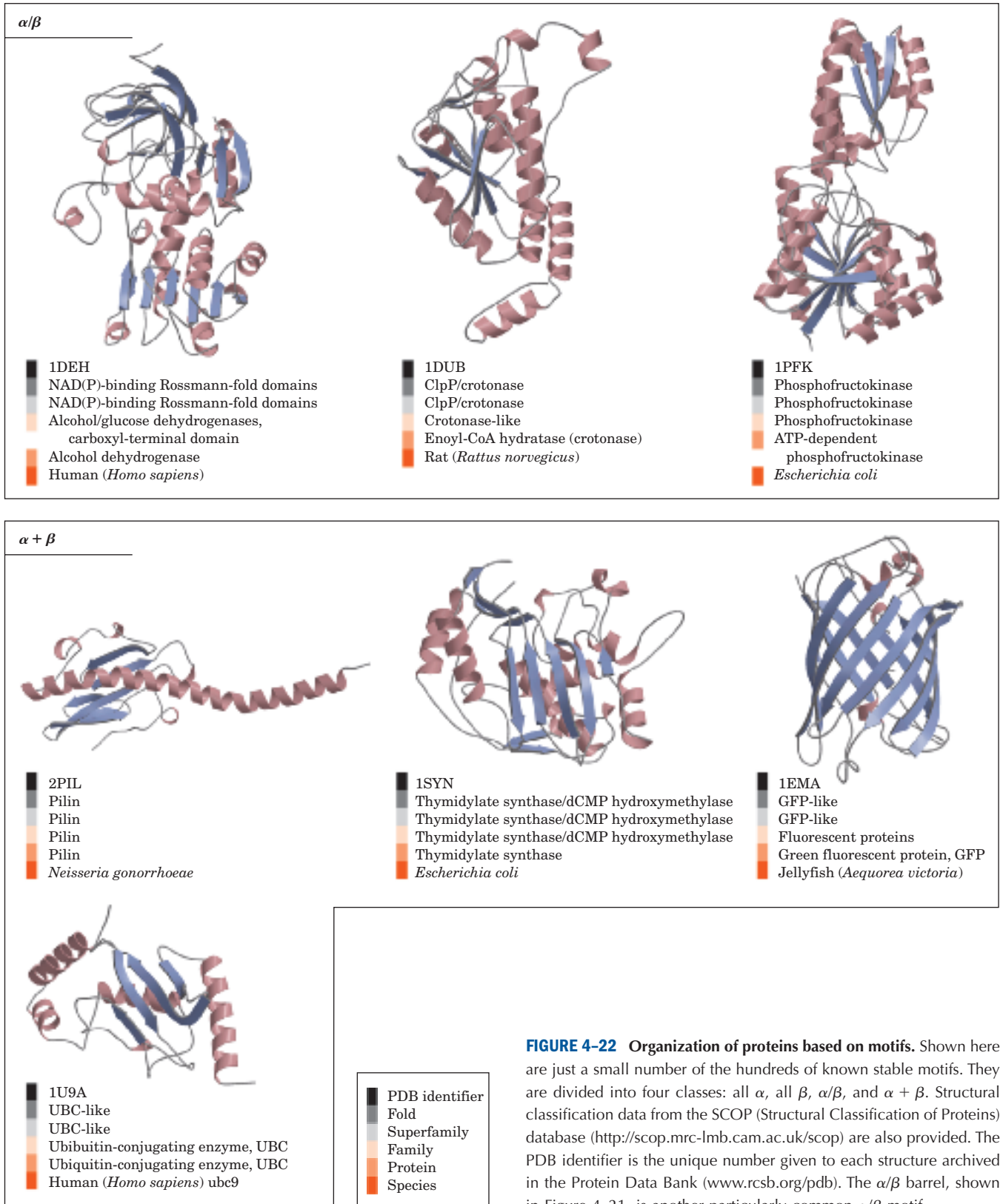
cofactor or substrate in the form of a pocket near one end of the barrel. Note that domains exhibiting similar folding patterns are said to have the same motif even though their constituent  $\alpha$  helices and  $\beta$  sheets may differ in length.

### Protein Motifs Are the Basis for Protein Structural Classification

**Protein Architecture—Tertiary Structure of Large Globular Proteins, IV. Structural Classification of Proteins** As we have seen, the complexities of tertiary structure are decreased by considering substructures. Taking this idea further, researchers have organized the complete contents of databases according to hierarchical levels of structure. The Structural Classification of Proteins (SCOP) database offers a good example of this very important trend in biochemistry. At the highest level of classification, the SCOP database (<http://scop.mrc-lmb.cam.ac.uk/scop>) borrows a scheme already in common use, in which protein structures are divided into four classes: all  $\alpha$ , all  $\beta$ ,  $\alpha/\beta$  (in which the  $\alpha$  and  $\beta$  segments are interspersed or alternate), and  $\alpha + \beta$  (in which the  $\alpha$  and  $\beta$  regions are somewhat segregated) (Fig. 4–22). Within each class are tens to hundreds of different folding arrangements, built up from increasingly identifiable substructures. Some of the substructure arrangements are very common, others have been found in just one protein. Figure 4–22 displays a variety of motifs arrayed among the four classes of protein structure. Those illustrated are just a minute sample of the hundreds of known motifs. The number of folding patterns is not infinite, however. As the rate at which new protein structures are elucidated has increased, the fraction of those structures containing a new motif has steadily declined. Fewer than 1,000 different folds or motifs may exist in all proteins. Figure 4–22 also shows how proteins can be organized based on the presence of the various motifs. The top two levels of organization, **class** and **fold**, are purely structural. Below the fold level, categorization is based on evolutionary relationships.

Many examples of recurring domain or motif structures are available, and these reveal that protein tertiary structure is more reliably conserved than primary sequence. The comparison of protein structures can thus provide much information about evolution. Proteins with significant primary sequence similarity, and/or with demonstrably similar structure and function, are said to be in the same **protein family**. A strong evolutionary relationship is usually evident within a protein family. For example, the globin family has many different proteins with both structural and sequence similarity to myoglobin (as seen in the proteins used as examples in Box 4–4 and again in the next chapter). Two or more families with little primary sequence similarity sometimes make use of the same major structural





**FIGURE 4-22 Organization of proteins based on motifs.** Shown here are just a small number of the hundreds of known stable motifs. They are divided into four classes: all  $\alpha$ , all  $\beta$ ,  $\alpha/\beta$ , and  $\alpha + \beta$ . Structural classification data from the SCOP (Structural Classification of Proteins) database (<http://scop.mrc-lmb.cam.ac.uk/scop>) are also provided. The PDB identifier is the unique number given to each structure archived in the Protein Data Bank ([www.rcsb.org/pdb](http://www.rcsb.org/pdb)). The  $\alpha/\beta$  barrel, shown in Figure 4-21, is another particularly common  $\alpha/\beta$  motif.




motif and have functional similarities; these families are grouped as **superfamilies**. An evolutionary relationship between the families in a superfamily is considered probable, even though time and functional distinctions—hence different adaptive pressures—may have erased many of the telltale sequence relationships. A protein family may be widespread in all three domains of cellular life, the Bacteria, Archaea, and Eukarya, suggesting a very ancient origin. Other families may be present in only a small group of organisms, indicating that the structure arose more recently. Tracing the natural history of structural motifs, using structural classifications in databases such as SCOP, provides a powerful complement to sequence analyses in tracing many evolutionary relationships.

The SCOP database is curated manually, with the objective of placing proteins in the correct evolutionary framework based on conserved structural features. Two similar enterprises, the CATH (*class, architecture, topology, and homologous superfamily*) and FSSP (*fold classification based on structure-structure alignment of proteins*) databases, make use of more automated methods and can provide additional information.

Structural motifs become especially important in defining protein families and superfamilies. Improved classification and comparison systems for proteins lead inevitably to the elucidation of new functional relationships. Given the central role of proteins in living systems, these structural comparisons can help illuminate every aspect of biochemistry, from the evolution of individual proteins to the evolutionary history of complete metabolic pathways.

### Protein Quaternary Structures Range from Simple Dimers to Large Complexes

 **Protein Architecture—Quaternary Structure** Many proteins have multiple polypeptide subunits. The association of polypeptide chains can serve a variety of functions. Many multisubunit proteins have regulatory roles; the binding of small molecules may affect the interaction between subunits, causing large changes in the protein's activity in response to small changes in the concentration of substrate or regulatory molecules (Chapter 6). In other cases, separate subunits can take on separate but related functions, such as catalysis and regulation. Some associations, such as the fibrous proteins considered earlier in this chapter and the coat proteins of viruses, serve primarily structural roles. Some very large protein assemblies are the site of complex, multistep reactions. One example is the ribosome, site of protein synthesis, which incorporates dozens of protein subunits along with a number of RNA molecules.

A multisubunit protein is also referred to as a **multimer**. Multimeric proteins can have from two to hundreds of subunits. A multimer with just a few subunits

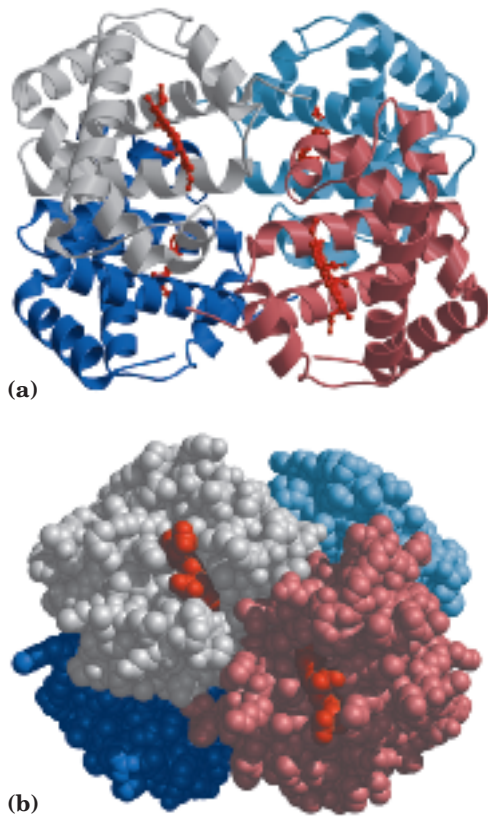
is often called an **oligomer**. If a multimer is composed of a number of nonidentical subunits, the overall structure of the protein can be asymmetric and quite complicated. However, most multimers have identical subunits or repeating groups of nonidentical subunits, usually in symmetric arrangements. As noted in Chapter 3, the repeating structural unit in such a multimeric protein, whether it is a single subunit or a group of subunits, is called a **protomer**.

The first oligomeric protein for which the three-dimensional structure was determined was hemoglobin ( $M_r$  64,500), which contains four polypeptide chains and four heme prosthetic groups, in which the iron atoms are in the ferrous ( $\text{Fe}^{2+}$ ) state (Fig. 4-17). The protein portion, called globin, consists of two  $\alpha$  chains (141 residues each) and two  $\beta$  chains (146 residues each). Note that in this case  $\alpha$  and  $\beta$  do not refer to secondary structures. Because hemoglobin is four times as large as myoglobin, much more time and effort were required to solve its three-dimensional structure by x-ray analysis, finally achieved by Max Perutz, John Kendrew, and their colleagues in 1959. The subunits of hemoglobin are arranged in symmetric pairs (Fig. 4-23), each pair having one  $\alpha$  and one  $\beta$  subunit. Hemoglobin can therefore be described either as a tetramer or as a dimer of  $\alpha\beta$  protomers.

Identical subunits of multimeric proteins are generally arranged in one or a limited set of symmetric patterns. A description of the structure of these proteins requires an understanding of conventions used to define symmetries. Oligomers can have either **rotational symmetry** or **helical symmetry**; that is, individual subunits can be superimposed on others (brought to coincidence) by rotation about one or more rotational axes, or by a helical rotation. In proteins with rotational symmetry, the subunits pack about the rotational axes to form closed structures. Proteins with helical symme-



Max Perutz, 1914–2002 (left)  
John Kendrew, 1917–1997 (right)

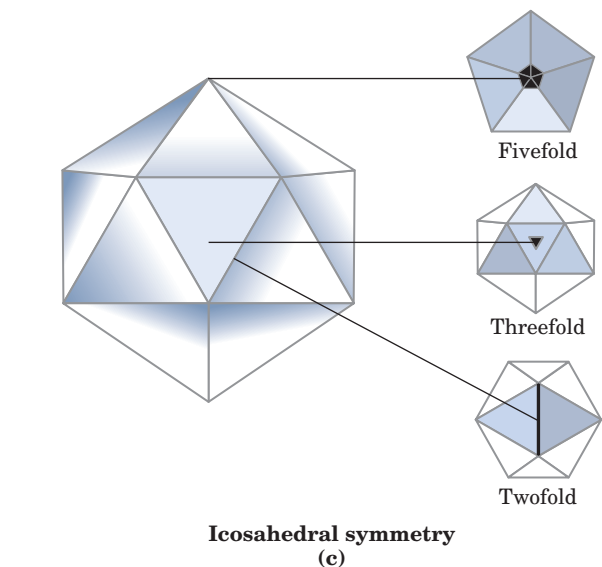
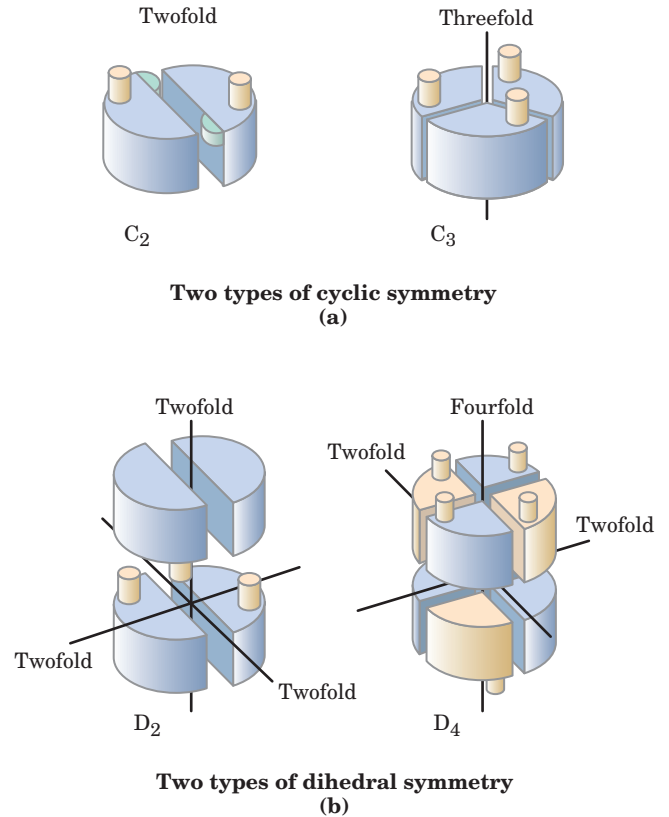


**FIGURE 4-23** Quaternary structure of deoxyhemoglobin. (PDB ID 2HHB) X-ray diffraction analysis of deoxyhemoglobin (hemoglobin without oxygen molecules bound to the heme groups) shows how the four polypeptide subunits are packed together. **(a)** A ribbon representation. **(b)** A space-filling model. The  $\alpha$  subunits are shown in gray and light blue; the  $\beta$  subunits in pink and dark blue. Note that the heme groups (red) are relatively far apart.

try tend to form structures that are more open-ended, with subunits added in a spiraling array.

There are several forms of rotational symmetry. The simplest is **cyclic symmetry**, involving rotation about a single axis (Fig. 4-24a). If subunits can be superimposed by rotation about a single axis, the protein has a symmetry defined by convention as  $C_n$  ( $C$  for cyclic,  $n$  for the number of subunits related by the axis). The axis itself is described as an  $n$ -fold rotational axis. The  $\alpha\beta$  protomers of hemoglobin (Fig. 4-23) are related by  $C_2$  symmetry. A somewhat more complicated rotational symmetry is **dihedral symmetry**, in which a twofold rotational axis intersects an  $n$ -fold axis at right angles. The symmetry is defined as  $D_n$  (Fig. 4-24b). A protein with dihedral symmetry has  $2n$  protomers.

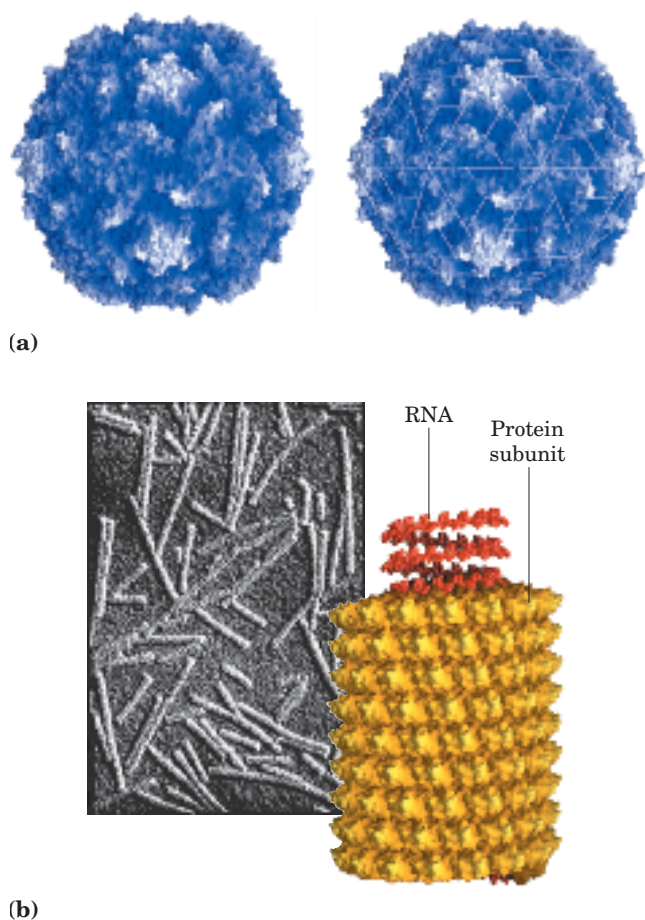
Proteins with cyclic or dihedral symmetry are particularly common. More complex rotational symmetries are possible, but only a few are regularly encountered. One example is **icosahedral symmetry**. An icosahedron is a regular 12-cornered polyhedron having 20 equilateral triangular faces (Fig. 4-24c). Each face can



**FIGURE 4-24** Rotational symmetry in proteins. **(a)** In cyclic symmetry, subunits are related by rotation about a single  $n$ -fold axis, where  $n$  is the number of subunits so related. The axes are shown as black lines; the numbers are values of  $n$ . Only two of many possible  $C_n$  arrangements are shown. **(b)** In dihedral symmetry, all subunits can be related by rotation about one or both of two axes, one of which is twofold.  $D_2$  symmetry is most common. **(c)** Icosahedral symmetry. Relating all 20 triangular faces of an icosahedron requires rotation about one or more of three separate rotational axes: twofold, threefold, and fivefold. An end-on view of each of these axes is shown at the right.

be brought to coincidence with another by rotation about one or more of three rotational axes. This is a common structure in virus coats, or capsids. The human poliovirus has an icosahedral capsid (Fig. 4–25a). Each triangular face is made up of three protomers, each protomer containing single copies of four different polypeptide chains, three of which are accessible at the outer surface. Sixty protomers form the 20 faces of the icosahedral shell enclosing the genetic material (RNA).

The other major type of symmetry found in oligomers, helical symmetry, also occurs in capsids. Tobacco mosaic virus is a right-handed helical filament made up of 2,130 identical subunits (Fig. 4–25b). This cylindrical structure encloses the viral RNA. Proteins with subunits arranged in helical filaments can also form long, fibrous structures such as the actin filaments of muscle (see Fig. 5–30).



**FIGURE 4–25 Viral capsids.** (a) Poliovirus (derived from PDB ID 2PLV). The coat proteins of poliovirus assemble into an icosahedron 300 Å in diameter. Icosahedral symmetry is a type of rotational symmetry (see Fig. 4–24c). On the left is a surface contour image of the poliovirus capsid. In the image on the right, lines have been superimposed to show the axes of symmetry. (b) Tobacco mosaic virus (derived from PDB ID 1VTM). This rod-shaped virus (as shown in the electron micrograph) is 3,000 Å long and 180 Å in diameter; it has helical symmetry.

### There Are Limits to the Size of Proteins

The relatively large size of proteins reflects their functions. The function of an enzyme, for example, requires a stable structure containing a pocket large enough to bind its substrate and catalyze a reaction. Protein size has limits, however, imposed by two factors: the genetic coding capacity of nucleic acids and the accuracy of the protein biosynthetic process. The use of many copies of one or a few proteins to make a large enclosing structure (capsid) is important for viruses because this strategy conserves genetic material. Remember that there is a linear correspondence between the sequence of a gene in the nucleic acid and the amino acid sequence of the protein for which it codes (see Fig. 1–31). The nucleic acids of viruses are much too small to encode the information required for a protein shell made of a single polypeptide. By using many copies of much smaller polypeptides, a much shorter nucleic acid is needed for coding the capsid subunits, and this nucleic acid can be efficiently used over and over again. Cells also use large complexes of polypeptides in muscle, cilia, the cytoskeleton, and other structures. It is simply more efficient to make many copies of a small polypeptide than one copy of a very large protein. In fact, most proteins with a molecular weight greater than 100,000 have multiple subunits, identical or different.

The second factor limiting the size of proteins is the error frequency during protein biosynthesis. The error frequency is low (about 1 mistake per 10,000 amino acid residues added), but even this low rate results in a high probability of a damaged protein if the protein is very large. Simply put, the potential for incorporating a “wrong” amino acid in a protein is greater for a large protein than for a small one.

### SUMMARY 4.3 Protein Tertiary and Quaternary Structures

- Tertiary structure is the complete three-dimensional structure of a polypeptide chain. There are two general classes of proteins based on tertiary structure: fibrous and globular.
- Fibrous proteins, which serve mainly structural roles, have simple repeating elements of secondary structure.
- Globular proteins have more complicated tertiary structures, often containing several types of secondary structure in the same polypeptide chain. The first globular protein structure to be determined, using x-ray diffraction methods, was that of myoglobin.
- The complex structures of globular proteins can be analyzed by examining stable substructures called supersecondary structures,



motifs, or folds. The thousands of known protein structures are generally assembled from a repertoire of only a few hundred motifs. Regions of a polypeptide chain that can fold stably and independently are called domains.

- Quaternary structure results from interactions between the subunits of multisubunit (multimeric) proteins or large protein assemblies. Some multimeric proteins have a repeated unit consisting of a single subunit or a group of subunits referred to as a protomer. Protomers are usually related by rotational or helical symmetry.

## 4.4 Protein Denaturation and Folding

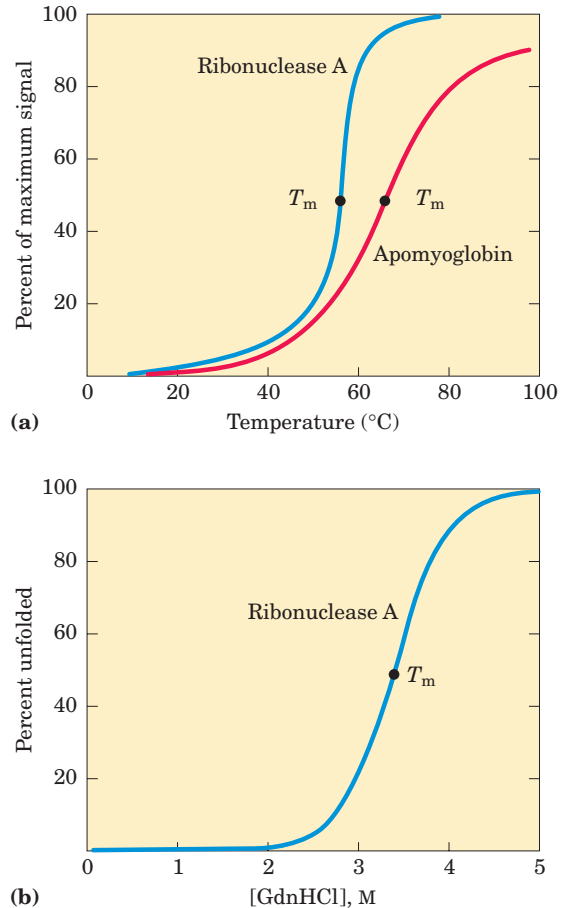
All proteins begin their existence on a ribosome as a linear sequence of amino acid residues (Chapter 27). This polypeptide must fold during and following synthesis to take up its native conformation. We have seen that a native protein conformation is only marginally stable. Modest changes in the protein's environment can bring about structural changes that can affect function. We now explore the transition that occurs between the folded and unfolded states.

### Loss of Protein Structure Results in Loss of Function

Protein structures have evolved to function in particular cellular environments. Conditions different from those in the cell can result in protein structural changes, large and small. A loss of three-dimensional structure sufficient to cause loss of function is called **denaturation**. The denatured state does not necessarily equate with complete unfolding of the protein and randomization of conformation. Under most conditions, denatured proteins exist in a set of partially folded states that are poorly understood.

Most proteins can be denatured by heat, which affects the weak interactions in a protein (primarily hydrogen bonds) in a complex manner. If the temperature is increased slowly, a protein's conformation generally remains intact until an abrupt loss of structure (and function) occurs over a narrow temperature range (Fig. 4–26). The abruptness of the change suggests that unfolding is a cooperative process: loss of structure in one part of the protein destabilizes other parts. The effects of heat on proteins are not readily predictable. The very heat-stable proteins of thermophilic bacteria have evolved to function at the temperature of hot springs (~100 °C). Yet the structures of these proteins often differ only slightly from those of homologous proteins derived from bacteria such as *Escherichia coli*. How these small differences promote structural stability at high temperatures is not yet understood.

Proteins can be denatured not only by heat but by extremes of pH, by certain miscible organic solvents



**FIGURE 4–26 Protein denaturation.** Results are shown for proteins denatured by two different environmental changes. In each case, the transition from the folded to unfolded state is fairly abrupt, suggesting cooperativity in the unfolding process. **(a)** Thermal denaturation of horse apomyoglobin (myoglobin without the heme prosthetic group) and ribonuclease A (with its disulfide bonds intact; see Fig. 4–27). The midpoint of the temperature range over which denaturation occurs is called the melting temperature, or  $T_m$ . The denaturation of apomyoglobin was monitored by circular dichroism, a technique that measures the amount of helical structure in a macromolecule. Denaturation of ribonuclease A was tracked by monitoring changes in the intrinsic fluorescence of the protein, which is affected by changes in the environment of Trp residues. **(b)** Denaturation of disulfide-intact ribonuclease A by guanidine hydrochloride (GdnHCl), monitored by circular dichroism.

such as alcohol or acetone, by certain solutes such as urea and guanidine hydrochloride, or by detergents. Each of these denaturing agents represents a relatively mild treatment in the sense that no covalent bonds in the polypeptide chain are broken. Organic solvents, urea, and detergents act primarily by disrupting the hydrophobic interactions that make up the stable core of globular proteins; extremes of pH alter the net charge on the protein, causing electrostatic repulsion and the disruption of some hydrogen bonding. The denatured states obtained with these various treatments need not be equivalent.



### Amino Acid Sequence Determines Tertiary Structure

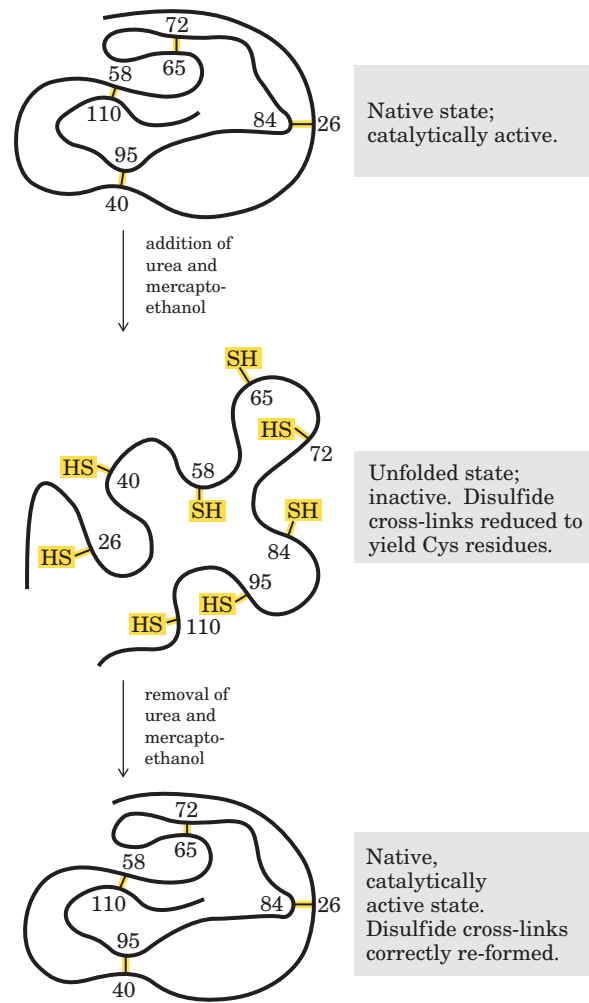
The tertiary structure of a globular protein is determined by its amino acid sequence. The most important proof of this came from experiments showing that denaturation of some proteins is reversible. Certain globular proteins denatured by heat, extremes of pH, or denaturing reagents will regain their native structure and their biological activity if returned to conditions in which the native conformation is stable. This process is called **renaturation**.

A classic example is the denaturation and renaturation of ribonuclease. Purified ribonuclease can be completely denatured by exposure to a concentrated urea solution in the presence of a reducing agent. The reducing agent cleaves the four disulfide bonds to yield eight Cys residues, and the urea disrupts the stabilizing hydrophobic interactions, thus freeing the entire polypeptide from its folded conformation. Denaturation of ribonuclease is accompanied by a complete loss of catalytic activity. When the urea and the reducing agent are removed, the randomly coiled, denatured ribonuclease spontaneously refolds into its correct tertiary structure, with full restoration of its catalytic activity (Fig. 4–27). The refolding of ribonuclease is so accurate that the four intrachain disulfide bonds are re-formed in the same positions in the renatured molecule as in the native ribonuclease. As calculated mathematically, the eight Cys residues could recombine at random to form up to four disulfide bonds in 105 different ways. In fact, an essentially random distribution of disulfide bonds is obtained when the disulfides are allowed to reform in the presence of denaturant, indicating that weak bonding interactions are required for correct positioning of disulfide bonds and assumption of the native conformation.

This classic experiment, carried out by Christian Anfinsen in the 1950s, provided the first evidence that the amino acid sequence of a polypeptide chain contains all the information required to fold the chain into its native, three-dimensional structure. Later, similar results were obtained using chemically synthesized, catalytically active ribonuclease. This eliminated the possibility that some minor contaminant in Anfinsen's purified ribonuclease preparation might have contributed to the renaturation of the enzyme, thus dispelling any remaining doubt that this enzyme folds spontaneously.

### Polypeptides Fold Rapidly by a Stepwise Process

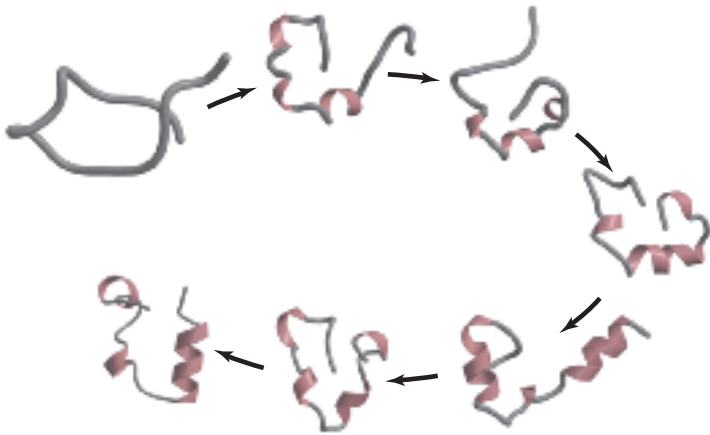
In living cells, proteins are assembled from amino acids at a very high rate. For example, *E. coli* cells can make a complete, biologically active protein molecule containing 100 amino acid residues in about 5 seconds at 37 °C. How does such a polypeptide chain arrive at its native conformation? Let's assume conservatively that each of the amino acid residues could take up 10 dif-



**FIGURE 4–27 Renaturation of unfolded, denatured ribonuclease.** Urea is used to denature ribonuclease, and mercaptoethanol ( $\text{HOCH}_2\text{CH}_2\text{SH}$ ) to reduce and thus cleave the disulfide bonds to yield eight Cys residues. Renaturation involves reestablishment of the correct disulfide cross-links.

ferent conformations on average, giving  $10^{100}$  different conformations for the polypeptide. Let's also assume that the protein folds itself spontaneously by a random process in which it tries out all possible conformations around every single bond in its backbone until it finds its native, biologically active form. If each conformation were sampled in the shortest possible time ( $\sim 10^{-13}$  second, or the time required for a single molecular vibration), it would take about  $10^{77}$  years to sample all possible conformations. Thus protein folding cannot be a completely random, trial-and-error process. There must be shortcuts. This problem was first pointed out by Cyrus Levinthal in 1968 and is sometimes called Levinthal's paradox.

The folding pathway of a large polypeptide chain is unquestionably complicated, and not all the principles that guide the process have been worked out. However, extensive study has led to the development of several



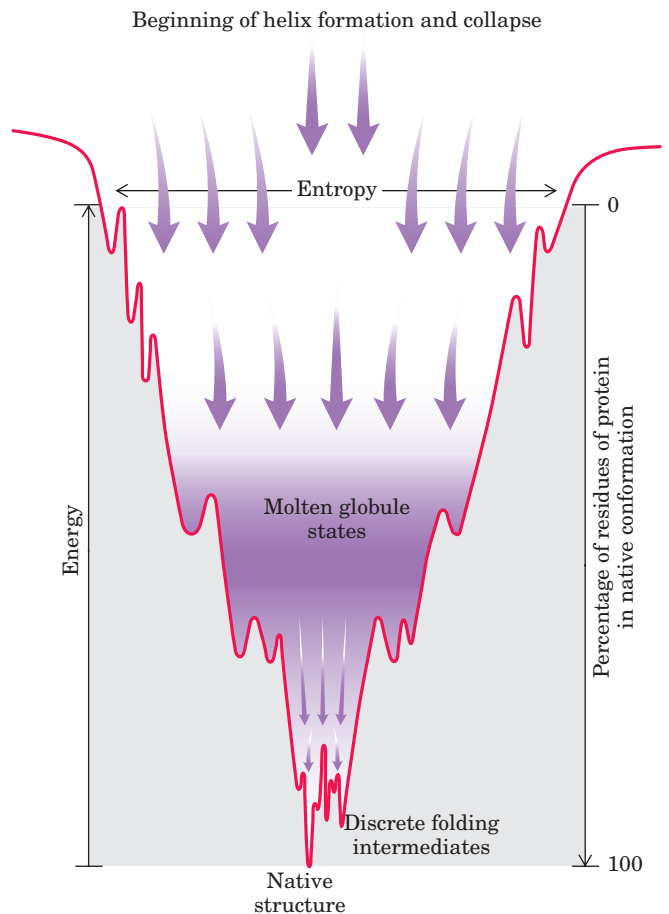
**FIGURE 4-28** A simulated folding pathway. The folding pathway of a 36-residue segment of the protein villin (an actin-binding protein found principally in the microvilli lining the intestine) was simulated by computer. The process started with the randomly coiled peptide and 3,000 surrounding water molecules in a virtual “water box.” The molecular motions of the peptide and the effects of the water molecules were taken into account in mapping the most likely paths to the final structure among the countless alternatives. The simulated folding took place in a theoretical time span of 1 ms; however, the calculation required half a billion integration steps on two Cray supercomputers, each running for two months.

plausible models. In one, the folding process is envisioned as hierarchical. Local secondary structures form first. Certain amino acid sequences fold readily into  $\alpha$  helices or  $\beta$  sheets, guided by constraints we have reviewed in our discussion of secondary structure. This is followed by longer-range interactions between, say, two  $\alpha$  helices that come together to form stable supersecondary structures. The process continues until complete domains form and the entire polypeptide is folded (Fig. 4-28). In an alternative model, folding is initiated by a spontaneous collapse of the polypeptide into a compact state, mediated by hydrophobic interactions among nonpolar residues. The state resulting from this “hydrophobic collapse” may have a high content of secondary structure, but many amino acid side chains are not entirely fixed. The collapsed state is often referred to as a **molten globule**. Most proteins probably fold by a process that incorporates features of both models. Instead of following a single pathway, a population of peptide molecules may take a variety of routes to the same end point, with the number of different partly folded conformational species decreasing as folding nears completion.

Thermodynamically, the folding process can be viewed as a kind of free-energy funnel (Fig. 4-29). The unfolded states are characterized by a high degree of conformational entropy and relatively high free energy. As folding proceeds, the narrowing of the funnel repre-

sents a decrease in the number of conformational species present. Small depressions along the sides of the free-energy funnel represent semistable intermediates that can briefly slow the folding process. At the bottom of the funnel, an ensemble of folding intermediates has been reduced to a single native conformation (or one of a small set of native conformations).

Defects in protein folding may be the molecular basis for a wide range of human genetic disorders. For example, cystic fibrosis is caused by defects in a membrane-bound protein called *cystic fibrosis transmembrane conductance regulator* (CFTR), which acts as a channel for chloride ions. The most common cystic



**FIGURE 4-29** The thermodynamics of protein folding depicted as a free-energy funnel. At the top, the number of conformations, and hence the conformational entropy, is large. Only a small fraction of the intramolecular interactions that will exist in the native conformation are present. As folding progresses, the thermodynamic path down the funnel reduces the number of states present (decreases entropy), increases the amount of protein in the native conformation, and decreases the free energy. Depressions on the sides of the funnel represent semistable folding intermediates, which may, in some cases, slow the folding process.

fibrosis-causing mutation is the deletion of a Phe residue at position 508 in CFTR, which causes improper protein folding (see Box 11-3). Many of the disease-related mutations in collagen (p. 129) also cause defective folding. An improved understanding of protein

folding may lead to new therapies for these and many other diseases (Box 4-5). ■

Thermodynamic stability is not evenly distributed over the structure of a protein—the molecule has regions of high and low stability. For example, a protein

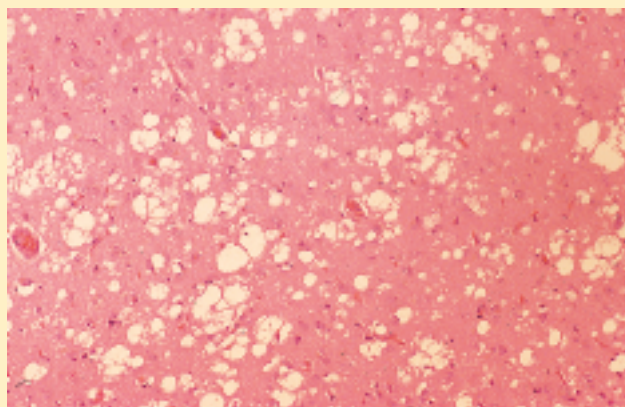


## BOX 4-5 BIOCHEMISTRY IN MEDICINE

### Death by Misfolding: The Prion Diseases

A misfolded protein appears to be the causative agent of a number of rare degenerative brain diseases in mammals. Perhaps the best known of these is mad cow disease (bovine spongiform encephalopathy, BSE), an outbreak of which made international headlines in the spring of 1996. Related diseases include kuru and Creutzfeldt-Jakob disease in humans, scrapie in sheep, and chronic wasting disease in deer and elk. These diseases are also referred to as spongiform encephalopathies, because the diseased brain frequently becomes riddled with holes (Fig. 1). Typical symptoms include dementia and loss of coordination. The diseases are fatal.

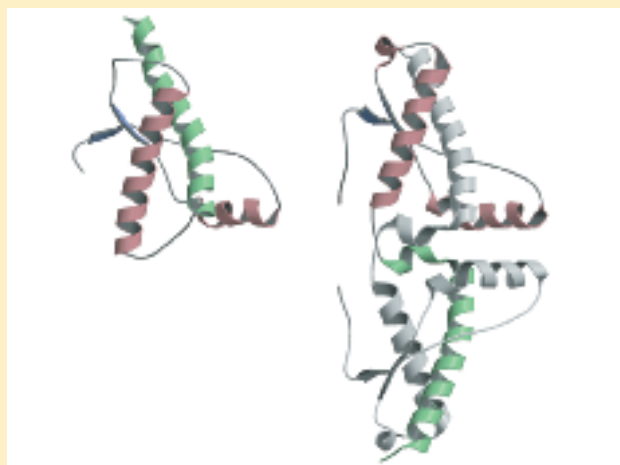
In the 1960s, investigators found that preparations of the disease-causing agents appeared to lack nucleic acids. At this time, Tikvah Alper suggested that the agent was a protein. Initially, the idea seemed heretical. All disease-causing agents known up to that time—viruses, bacteria, fungi, and so on—contained nucleic acids, and their virulence was related to genetic reproduction and propagation. However, four decades of investigations, pursued most notably by Stanley Prusiner, have provided evidence that spongiform encephalopathies are different.



**FIGURE 1** A stained section of the cerebral cortex from a patient with Creutzfeldt-Jakob disease shows spongiform (vacuolar) degeneration, the most characteristic neurohistological feature. The yellowish vacuoles are intracellular and occur mostly in pre- and post-synaptic processes of neurons. The vacuoles in this section vary in diameter from 20 to 100  $\mu\text{m}$ .

The infectious agent has been traced to a single protein ( $M_r$  28,000), which Prusiner dubbed **prion** (from *proteinaceous infectious only*) protein (PrP). Prion protein is a normal constituent of brain tissue in all mammals. Its role in the mammalian brain is not known in detail, but it appears to have a molecular signaling function. Strains of mice lacking the gene for PrP (and thus the protein itself) suffer no obvious ill effects. Illness occurs only when the normal cellular PrP, or PrP<sup>C</sup>, occurs in an altered conformation called PrP<sup>Sc</sup> (Sc denotes scrapie). The interaction of PrP<sup>Sc</sup> with PrP<sup>C</sup> converts the latter to PrP<sup>Sc</sup>, initiating a domino effect in which more and more of the brain protein converts to the disease-causing form. The mechanism by which the presence of PrP<sup>Sc</sup> leads to spongiform encephalopathy is not understood.

In inherited forms of prion diseases, a mutation in the gene encoding PrP produces a change in one amino acid residue that is believed to make the conversion of PrP<sup>C</sup> to PrP<sup>Sc</sup> more likely. A complete understanding of prion diseases awaits new information about how prion protein affects brain function. Structural information about PrP is beginning to provide insights into the molecular process that allows the prion proteins to interact so as to alter their conformation (Fig. 2).



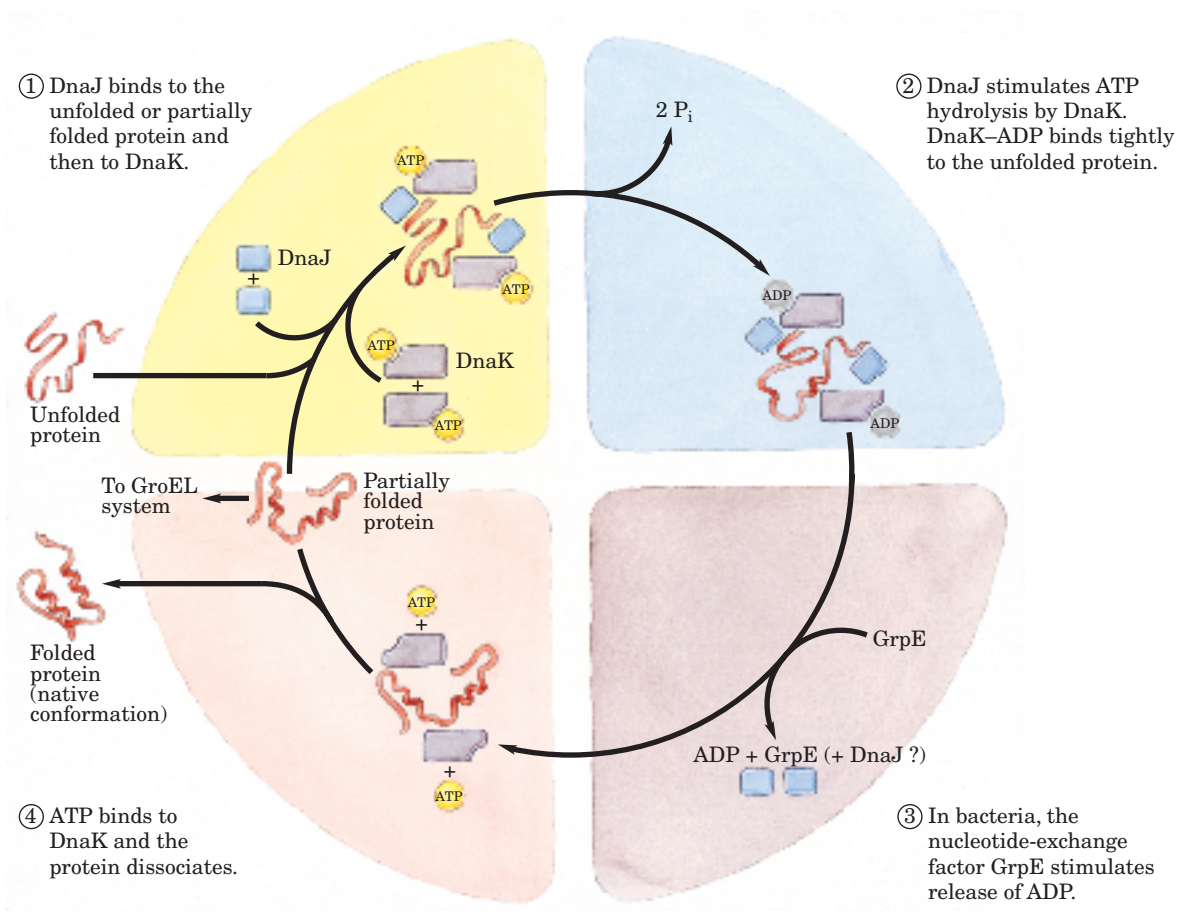
**FIGURE 2** The structure of the globular domain of human PrP in monomeric (left) and dimeric (right) forms. The second subunit is gray to highlight the dramatic conformational change in the green  $\alpha$  helix when the dimer is formed.

may have two stable domains joined by a segment with lower structural stability, or one small part of a domain may have a lower stability than the remainder. The regions of low stability allow a protein to alter its conformation between two or more states. As we shall see in the next two chapters, variations in the stability of regions within a given protein are often essential to protein function.

### Some Proteins Undergo Assisted Folding

Not all proteins fold spontaneously as they are synthesized in the cell. Folding for many proteins is facilitated by the action of specialized proteins. **Molecular chaperones** are proteins that interact with partially folded or improperly folded polypeptides, facilitating correct folding pathways or providing microenvironments in which folding can occur. Two classes of molecular chaperones have been well studied. Both are found in organisms ranging from bacteria to humans. The first class, a family of proteins called **Hsp70**, generally have

a molecular weight near 70,000 and are more abundant in cells stressed by elevated temperatures (hence, *heat shock proteins* of  $M_r$  70,000, or Hsp70). Hsp70 proteins bind to regions of unfolded polypeptides that are rich in hydrophobic residues, preventing inappropriate aggregation. These chaperones thus “protect” proteins that have been denatured by heat and peptides that are being synthesized (and are not yet folded). Hsp70 proteins also block the folding of certain proteins that must remain unfolded until they have been translocated across membranes (as described in Chapter 27). Some chaperones also facilitate the quaternary assembly of oligomeric proteins. The Hsp70 proteins bind to and release polypeptides in a cycle that also involves several other proteins (including a class called Hsp40) and ATP hydrolysis. Figure 4–30 illustrates chaperone-assisted folding as elucidated for the chaperones DnaK and DnaJ in *E. coli*, homologs of the eukaryotic Hsp70 and Hsp40. DnaK and DnaJ were first identified as proteins required for *in vitro* replication of certain viral DNA molecules (hence the “Dna” designation).



**FIGURE 4–30 Chaperones in protein folding.** The cyclic pathway by which chaperones bind and release polypeptides is illustrated for the *E. coli* chaperone proteins DnaK and DnaJ, homologs of the eukaryotic chaperones Hsp70 and Hsp40. The chaperones do not actively promote the folding of the substrate protein, but instead prevent aggregation of unfolded peptides. For a population of polypeptides, some

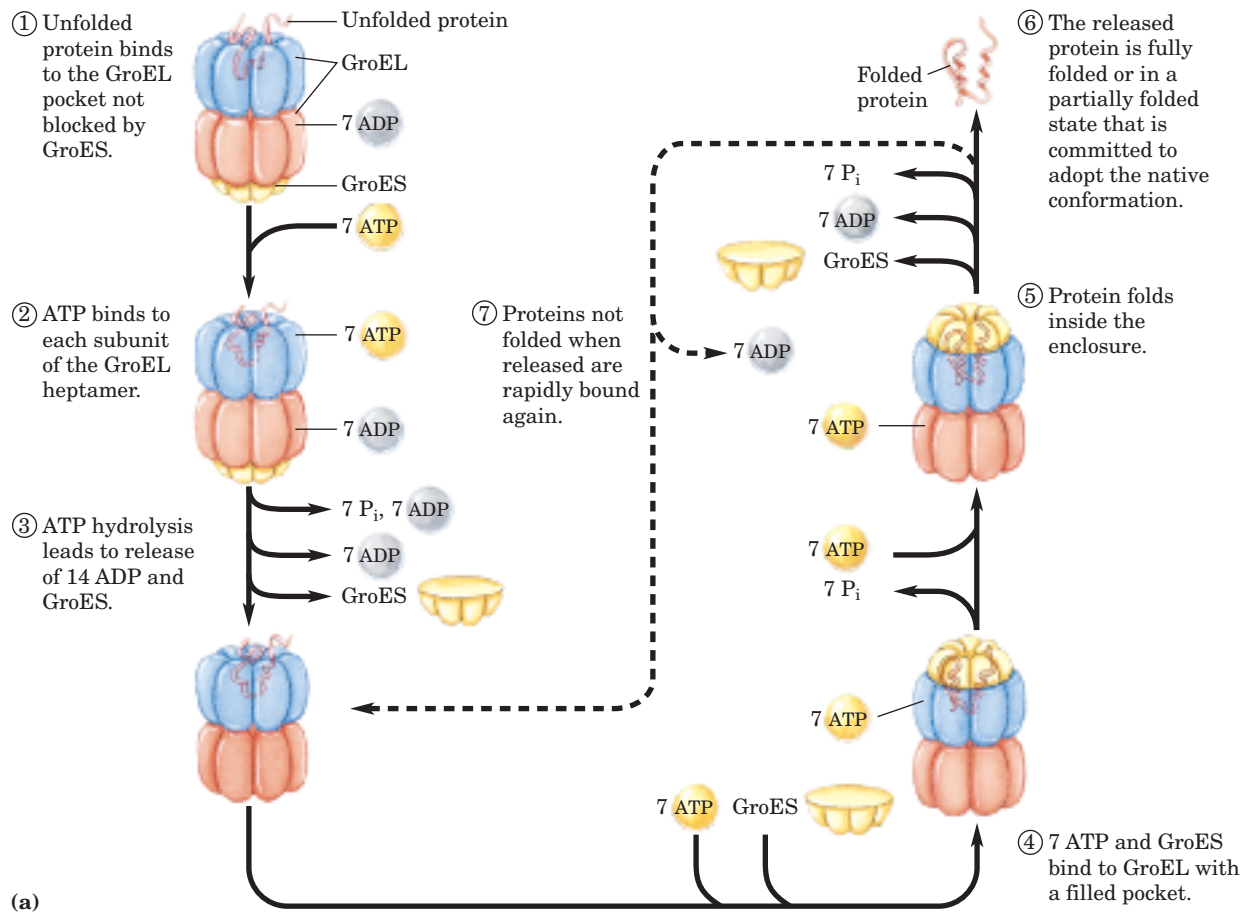
fraction of the polypeptides released at the end of the cycle are in the native conformation. The remainder are rebound by DnaK or are diverted to the chaperonin system (GroEL; see Fig. 4–31). In bacteria, a protein called GrpE interacts transiently with DnaK late in the cycle (step ③), promoting dissociation of ADP and possibly DnaJ. No eukaryotic analog of GrpE is known.



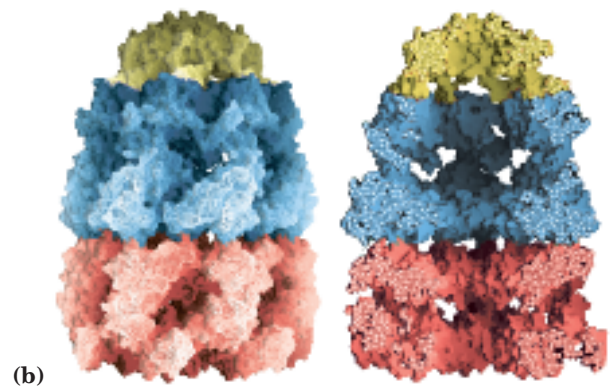
The second class of chaperones is called **chaperonins**. These are elaborate protein complexes required for the folding of a number of cellular proteins that do not fold spontaneously. In *E. coli* an estimated 10% to 15% of cellular proteins require the resident chaperonin system, called GroEL/GroES, for folding under normal conditions (up to 30% require this assistance when the cells are heat stressed). These proteins first became known when they were found to be necessary for the growth of certain bacterial viruses (hence the designation “Gro”). Unfolded proteins are bound within pockets in the GroEL complex, and the pockets are capped transiently by the GroES “lid” (Fig. 4–31). GroEL un-

dergoes substantial conformational changes, coupled to ATP hydrolysis and the binding and release of GroES, which promote folding of the bound polypeptide. Although the structure of the GroEL/GroES chaperonin is known, many details of its mechanism of action remain unresolved.

Finally, the folding pathways of a number of proteins require two enzymes that catalyze isomerization reactions. **Protein disulfide isomerase (PDI)** is a widely distributed enzyme that catalyzes the interchange or shuffling of disulfide bonds until the bonds of the native conformation are formed. Among its functions, PDI catalyzes the elimination of folding interme-



**FIGURE 4–31 Chaperonins in protein folding.** (a) A proposed pathway for the action of the *E. coli* chaperonins GroEL (a member of the Hsp60 protein family) and GroES. Each GroEL complex consists of two large pockets formed by two heptameric rings (each subunit  $M_r$  57,000). GroES, also a heptamer (subunits  $M_r$  10,000), blocks one of the GroEL pockets. (b) Surface and cut-away images of the GroEL/GroES complex (PDB ID 1AON). The cut-away (right) illustrates the large interior space within which other proteins are bound.



diates with inappropriate disulfide cross-links. **Peptide prolyl cis-trans isomerase (PPI)** catalyzes the interconversion of the cis and trans isomers of Pro peptide bonds (Fig. 4–8b), which can be a slow step in the folding of proteins that contain some Pro residue peptide bonds in the cis conformation.

Protein folding is likely to be a more complex process in the densely packed cellular environment than in the test tube. More classes of proteins that facilitate protein folding may be discovered as the biochemical dissection of the folding process continues.

#### SUMMARY 4.4 Protein Denaturation and Folding

- The three-dimensional structure and the function of proteins can be destroyed by denaturation, demonstrating a relationship

between structure and function. Some denatured proteins can renature spontaneously to form biologically active protein, showing that protein tertiary structure is determined by amino acid sequence.

- Protein folding in cells probably involves multiple pathways. Initially, regions of secondary structure may form, followed by folding into supersecondary structures. Large ensembles of folding intermediates are rapidly brought to a single native conformation.
- For many proteins, folding is facilitated by Hsp70 chaperones and by chaperonins. Disulfide bond formation and the cis-trans isomerization of Pro peptide bonds are catalyzed by specific enzymes.

### Key Terms

*Terms in bold are defined in the glossary.*

<b>conformation</b> 116	<b>β conformation</b> 123	collagen 127	<b>protomer</b> 144
<b>native conformation</b> 117	β sheet 123	silk fibroin 129	symmetry 144
solvation layer 117	<b>β turn</b> 123	<b>supersecondary structures</b> 139	<b>denaturation</b> 147
peptide group 118	<b>tertiary structure</b> 125	<b>motif</b> 139	molten globule 149
Ramachandran plot 118	<b>quaternary structure</b> 125	<b>fold</b> 139	prion 150
<b>secondary structure</b> 120	<b>fibrous proteins</b> 125	<b>domain</b> 140	molecular chaperone 151
<b>α helix</b> 120	<b>globular proteins</b> 125	protein family 141	Hsp70 151
	α-keratin 126	multimer 144	chaperonin 152
		<b>oligomer</b> 144	

### Further Reading

#### General

**Anfinsen, C.B.** (1973) Principles that govern the folding of protein chains. *Science* **181**, 223–230.

The author reviews his classic work on ribonuclease.

**Branden, C. & Tooze, J.** (1991) *Introduction to Protein Structure*, Garland Publishing, Inc., New York.

**Creighton, T.E.** (1993) *Proteins: Structures and Molecular Properties*, 2nd edn, W. H. Freeman and Company, New York. A comprehensive and authoritative source.

Evolution of Catalytic Function. (1987) *Cold Spring Harb. Symp. Quant. Biol.* **52**.

A collection of excellent articles on many topics, including protein structure, folding, and function.

**Kendrew, J.C.** (1961) The three-dimensional structure of a protein molecule. *Sci. Am.* **205** (December), 96–111.

Describes how the structure of myoglobin was determined and what was learned from it.

**Richardson, J.S.** (1981) The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.* **34**, 167–339.

An outstanding summary of protein structural patterns and principles; the author originated the very useful “ribbon” representations of protein structure.

#### Secondary, Tertiary, and Quaternary Structures

**Berman, H.M.** (1999) The past and future of structure databases. *Curr. Opin. Biotechnol.* **10**, 76–80.

A broad summary of the different approaches being used to catalog protein structures.

**Brenner, S.E., Chothia, C., & Hubbard, T.J.P.** (1997) Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* **7**, 369–376.

**Fuchs, E. & Cleveland, D.W.** (1998) A structural scaffolding of intermediate filaments in health and disease. *Science* **279**, 514–519.

**McPherson, A.** (1989) Macromolecular crystals. *Sci. Am.* **260** (March), 62–69.

A description of how macromolecules such as proteins are crystallized.

**Ponting, C.P. & Russell, R.R.** (2002) The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* **31**, 45–71.

An explanation of how structural databases can be used to explore evolution.

**Prockop, D.J. & Kivirikko, K.I.** (1995) Collagens, molecular biology, diseases, and potentials for therapy. *Annu. Rev. Biochem.* **64**, 403–434.

### Protein Denaturation and Folding

**Baldwin, R.L.** (1994) Matching speed and stability. *Nature* **369**, 183–184.

**Bukau, B., Deuerling, E., Pfund, C., & Craig, E.A.** (2000) Getting newly synthesized proteins into shape. *Cell* **101**, 119–122.

A good summary of chaperone mechanisms.

**Collinge, J.** (2001) Prion diseases of humans and animals: their causes and molecular basis. *Annu. Rev. Neurosci.* **24**, 519–550.

**Creighton, T.E., Darby, N.J., & Kemmink, J.** (1996) The roles of partly folded intermediates in protein folding. *FASEB J.* **10**, 110–118.

**Daggett, V., & Fersht, A.R.** (2003) Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.* **28**, 18–25.

**Dill, K.A. & Chan, H.S.** (1997) From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**, 10–19.

**Luque, I., Leavitt, S.A., & Freire, E.** (2002) The linkage between protein folding and functional cooperativity: two sides of the same coin? *Annu. Rev. Biophys. Biomol. Struct.* **31**, 235–256.

A review of how variations in structural stability within one protein contribute to function.

**Nicotera, P.** (2001) A route for prion neuroinvasion. *Neuron* **31**, 345–348.

**Prusiner, S.B.** (1995) The prion diseases. *Sci. Am.* **272** (January), 48–57.

A good summary of the evidence leading to the prion hypothesis.

**Richardson, A., Landry, S.J., & Georgopolous, C.** (1998) The ins and outs of a molecular chaperone machine. *Trends Biochem. Sci.* **23**, 138–143.

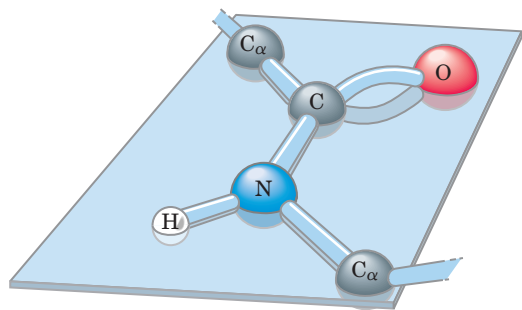
**Thomas, P.J., Qu, B.-H., & Pederson, P.L.** (1995) Defective protein folding as a basis of human disease. *Trends Biochem. Sci.* **20**, 456–459.

**Westaway, D. & Carlson, G.A.** (2002) Mammalian prion proteins: enigma, variation and vaccination. *Trends Biochem. Sci.* **27**, 301–307.

A good update.

## Problems

**1. Properties of the Peptide Bond** In x-ray studies of crystalline peptides, Linus Pauling and Robert Corey found that the C—N bond in the peptide link is intermediate in length (1.32 Å) between a typical C—N single bond (1.49 Å) and a C=N double bond (1.27 Å). They also found that the peptide bond is planar (all four atoms attached to the C—N group are located in the same plane) and that the two  $\alpha$ -carbon atoms attached to the C—N are always trans to each other (on opposite sides of the peptide bond):



(a) What does the length of the C—N bond in the peptide linkage indicate about its strength and its bond order (i.e., whether it is single, double, or triple)?

(b) What do the observations of Pauling and Corey tell us about the ease of rotation about the C—N peptide bond?

**2. Structural and Functional Relationships in Fibrous Proteins** William Astbury discovered that the x-ray pattern of wool shows a repeating structural unit spaced about 5.2 Å along the length of the wool fiber. When he steamed and

stretched the wool, the x-ray pattern showed a new repeating structural unit at a spacing of 7.0 Å. Steaming and stretching the wool and then letting it shrink gave an x-ray pattern consistent with the original spacing of about 5.2 Å. Although these observations provided important clues to the molecular structure of wool, Astbury was unable to interpret them at the time.

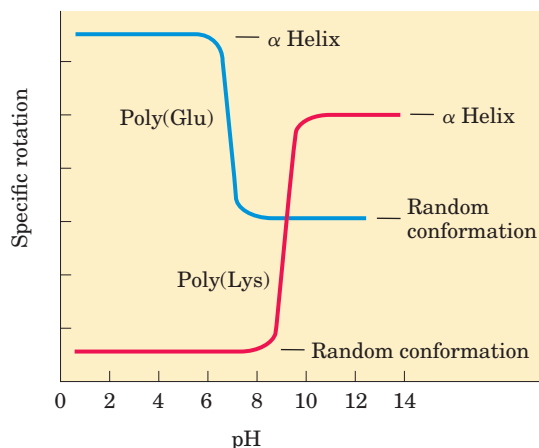
(a) Given our current understanding of the structure of wool, interpret Astbury's observations.

(b) When wool sweaters or socks are washed in hot water or heated in a dryer, they shrink. Silk, on the other hand, does not shrink under the same conditions. Explain.

**3. Rate of Synthesis of Hair  $\alpha$ -Keratin** Hair grows at a rate of 15 to 20 cm/yr. All this growth is concentrated at the base of the hair fiber, where  $\alpha$ -keratin filaments are synthesized inside living epidermal cells and assembled into ropelike structures (see Fig. 4–11). The fundamental structural element of  $\alpha$ -keratin is the  $\alpha$  helix, which has 3.6 amino acid residues per turn and a rise of 5.4 Å per turn (see Fig. 4–4b). Assuming that the biosynthesis of  $\alpha$ -helical keratin chains is the rate-limiting factor in the growth of hair, calculate the rate at which peptide bonds of  $\alpha$ -keratin chains must be synthesized (peptide bonds per second) to account for the observed yearly growth of hair.

**4. Effect of pH on the Conformation of  $\alpha$ -Helical Secondary Structures** The unfolding of the  $\alpha$  helix of a polypeptide to a randomly coiled conformation is accompanied by a large decrease in a property called its specific rotation, a measure of a solution's capacity to rotate plane-polarized light. Polyglutamate, a polypeptide made up of only L-Glu residues,

has the  $\alpha$ -helical conformation at pH 3. When the pH is raised to 7, there is a large decrease in the specific rotation of the solution. Similarly, polylysine (L-Lys residues) is an  $\alpha$  helix at pH 10, but when the pH is lowered to 7 the specific rotation also decreases, as shown by the following graph.



What is the explanation for the effect of the pH changes on the conformations of poly(Glu) and poly(Lys)? Why does the transition occur over such a narrow range of pH?

**5. Disulfide Bonds Determine the Properties of Many Proteins** A number of natural proteins are very rich in disulfide bonds, and their mechanical properties (tensile strength, viscosity, hardness, etc.) are correlated with the degree of disulfide bonding. For example, glutenin, a wheat protein rich in disulfide bonds, is responsible for the cohesive and elastic character of dough made from wheat flour. Similarly, the hard, tough nature of tortoise shell is due to the extensive disulfide bonding in its  $\alpha$ -keratin.

(a) What is the molecular basis for the correlation between disulfide-bond content and mechanical properties of the protein?

(b) Most globular proteins are denatured and lose their activity when briefly heated to 65 °C. However, globular proteins that contain multiple disulfide bonds often must be heated longer at higher temperatures to denature them. One such protein is bovine pancreatic trypsin inhibitor (BPTI), which has 58 amino acid residues in a single chain and contains three disulfide bonds. On cooling a solution of denatured BPTI, the activity of the protein is restored. What is the molecular basis for this property?

**6. Amino Acid Sequence and Protein Structure** Our growing understanding of how proteins fold allows researchers to make predictions about protein structure based on primary amino acid sequence data.

1	2	3	4	5	6	7	8	9	10
Ile	-Ala	-His	-Thr	-Tyr	-Gly	-Pro	-Phe	-Glu	-Ala
11	12	13	14	15	16	17	18	19	20
Ala	-Met	-Cys	-Lys	-Trp	-Glu	-Ala	-Gln	-Pro	-Asp
21	22	23	24	25	26	27	28		
Gly	-Met	-Glu	-Cys	-Ala	-Phe	-His	-Arg		

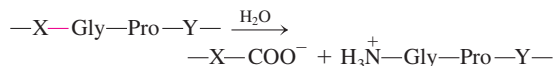
(a) In the amino acid sequence above, where would you predict that bends or  $\beta$  turns would occur?

(b) Where might intrachain disulfide cross-linkages be formed?

(c) Assuming that this sequence is part of a larger globular protein, indicate the probable location (the external surface or interior of the protein) of the following amino acid residues: Asp, Ile, Thr, Ala, Gln, Lys. Explain your reasoning. (Hint: See the hydropathy index in Table 3-1.)

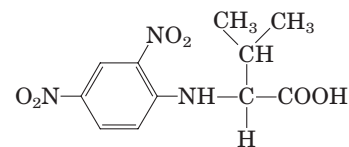
**7. Bacteriorhodopsin in Purple Membrane Proteins** Under the proper environmental conditions, the salt-loving bacterium *Halobacterium halobium* synthesizes a membrane protein ( $M_r$  26,000) known as bacteriorhodopsin, which is purple because it contains retinal (see Fig. 10-21). Molecules of this protein aggregate into "purple patches" in the cell membrane. Bacteriorhodopsin acts as a light-activated proton pump that provides energy for cell functions. X-ray analysis of this protein reveals that it consists of seven parallel  $\alpha$ -helical segments, each of which traverses the bacterial cell membrane (thickness 45 Å). Calculate the minimum number of amino acid residues necessary for one segment of  $\alpha$  helix to traverse the membrane completely. Estimate the fraction of the bacteriorhodopsin protein that is involved in membrane-spanning helices. (Use an average amino acid residue weight of 110.)

**8. Pathogenic Action of Bacteria That Cause Gas Gangrene** The highly pathogenic anaerobic bacterium *Clostridium perfringens* is responsible for gas gangrene, a condition in which animal tissue structure is destroyed. This bacterium secretes an enzyme that efficiently catalyzes the hydrolysis of the peptide bond indicated in red:



where X and Y are any of the 20 common amino acids. How does the secretion of this enzyme contribute to the invasiveness of this bacterium in human tissues? Why does this enzyme not affect the bacterium itself?

**9. Number of Polypeptide Chains in a Multisubunit Protein** A sample (660 mg) of an oligomeric protein of  $M_r$  132,000 was treated with an excess of 1-fluoro-2,4-dinitrobenzene (Sanger's reagent) under slightly alkaline conditions until the chemical reaction was complete. The peptide bonds of the protein were then completely hydrolyzed by heating it with concentrated HCl. The hydrolysate was found to contain 5.5 mg of the following compound:



2,4-Dinitrophenyl derivatives of the  $\alpha$ -amino groups of other amino acids could not be found.

(a) Explain how this information can be used to determine the number of polypeptide chains in an oligomeric protein.

(b) Calculate the number of polypeptide chains in this protein.

(c) What other protein analysis technique could you employ to determine whether the polypeptide chains in this protein are similar or different?



**Biochemistry on the Internet**

**10. Protein Modeling on the Internet** A group of patients suffering from Crohn's disease (an inflammatory bowel disease) underwent biopsies of their intestinal mucosa in an attempt to identify the causative agent. A protein was identified that was expressed at higher levels in patients with Crohn's disease than in patients with an unrelated inflammatory bowel disease or in unaffected controls. The protein was isolated and the following *partial* amino acid sequence was obtained (reads left to right):

EAELCPDRCI	HSFQNLGIQC	VKKRDLEQAI
SQRIQTNNNP	FQVPIEEQRG	DYDLNAVRLC
FQVTVRDPSG	RPLRLPPVLP	HPIFDNRAPN
TAEKICRVN	RNSGSLGGD	EIFLLCDKVQ
KEDIEVYFTG	PGWEARGSFS	QADVHRQVAI
VFRTPPYADP	SLQAPVRVSM	QLRRPSDREL
SEPMEFQYLP	DTDDRHRIEE	KRKRTYETFK
SIMKKSPPFSG	PTDPRPPRR	IAPSRSSAS
VPKPAPQPYP		

(a) You can identify this protein using a protein database on the Internet. Some good places to start include Protein Information Resource (PIR; [pir.georgetown.edu/pirwww](http://pir.georgetown.edu/pirwww)), Structural Classification of Proteins (SCOP; <http://scop.berkeley.edu>), and Prosite (<http://us.expasy.org/prosite>).

At your selected database site, follow links to locate the sequence comparison engine. Enter about 30 residues from the sequence of the protein in the appropriate search field and submit it for analysis. What does this analysis tell you about the identity of the protein?

(b) Try using different portions of the protein amino acid sequence. Do you always get the same result?

(c) A variety of websites provide information about the three-dimensional structure of proteins. Find information about the protein's secondary, tertiary, and quaternary structure using database sites such as the Protein Data Bank (PDB; [www.rcsb.org/pdb](http://www.rcsb.org/pdb)) or SCOP.

(d) In the course of your Web searches try to find information about the cellular function of the protein.