4

The Three-Dimensional Structure of Proteins

Proteins are big molecules. 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 in principle assume a virtually uncountable number of conformations. However, each protein has a specific chemical or structural function, suggesting that each has a unique three-dimensional structure **(Fig. 4–1)**. How stable is this structure, what factors guide its formation, and what holds it together? 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, generally, the ordered array of molecules in a crystal can form only if the molecular units are identical, the finding that many proteins could be crystallized was evidence that even very large proteins are discrete chemical entities with unique structures. This conclusion revolutionized thinking about proteins and their functions, but the insight it provided was incomplete. Protein structure is always malleable in sometimes surprising ways. Changes in structure can be as important to a protein's function as the structure itself.

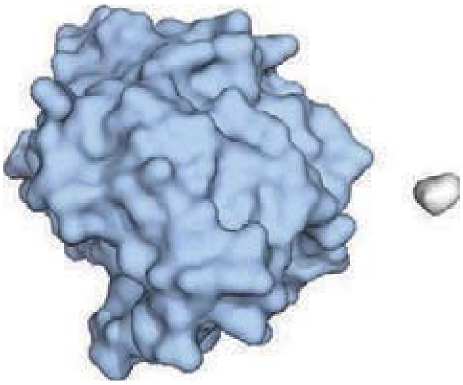


FIGURE 4–1 Structure of the enzyme chymotrypsin, a globular protein. A molecule of glycine (gray) is shown for size comparison. The known three-dimensional structures of proteins are archived in the Protein Data Bank, or PDB (see Box 4–4). The image shown here was made using data from the entry with PDB ID 6GCH.

In this chapter, we examine the structure of proteins. We emphasize six themes. First, the three-dimensional structure or structures taken up by a protein are determined by its amino acid sequence. Second, the function of a typical protein depends on its structure. Third, most isolated proteins exist 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. Fifth, amid the huge number of unique protein structures, we can recognize some common structural patterns that help to organize our understanding of protein architecture. Sixth, protein structures are not static. All proteins undergo changes in conformation ranging from subtle to quite dramatic. Parts of many proteins have no discernible structure. For some proteins, a lack of definable structure is critical to their function.

4.1 Overview of Protein Structure

The spatial arrangement of atoms in a protein or any part of a protein is called its **conformation**. The possible conformations of a protein or protein segment include any structural state it can achieve without breaking covalent bonds. A change in conformation could occur, for example, by rotation about single bonds. Of the many 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 take place in most proteins as they bind to other molecules or catalyze reactions. The conformations existing under a given set of conditions are usually the ones that are thermodynamically the most stable—that is, having the lowest Gibbs free energy (G). Proteins in any of their functional, folded conformations are called **native** proteins.

For the vast majority of proteins, a particular structure or small set of structures is critical to function. However, in many cases, parts of proteins lack discernible structure. These protein segments are intrinsically disordered. In a few cases, entire proteins are intrinsically disordered, yet fully functional.

What principles determine the most stable conformations of a typical 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 we must add the newer emphasis on common and classifiable folding patterns, variously called supersecondary structures, folds, or motifs, which provide 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 ΔG 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 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 the solvent (water), tend to maintain the unfolded state. The chemical interactions that counteract these effects and stabilize the native conformation include disulfide (covalent) bonds and the weak (noncovalent) interactions described in Chapter 2: hydrogen bonds and hydrophobic and ionic interactions.

Many proteins do not have disulfide bonds. The environment within most cells is highly reducing due to high concentrations of reductants such as glutathione, and most sulfhydryls will thus remain in the reduced state. Outside the cell, the environment is often more oxidizing, and disulfide formation is more likely to occur. In eukaryotes, disulfide bonds are found primarily in secreted, extracellular proteins (for example, the hormone insulin). Disulfide bonds are also uncommon in bacterial proteins. However, thermophilic bacteria, as well as the archaea, typically have many proteins with disulfide bonds, which stabilize proteins; this is presumably an adaptation to life at high temperatures.

For all proteins of all organisms, weak interactions are especially important in the folding of polypeptide chains into their secondary and tertiary structures. The association of multiple polypeptides to form quaternary structures also relies on these weak interactions.

About 200 to 460 kJ/mol are required to break a single covalent bond, whereas weak interactions can be disrupted by a mere 0.4 to 30 kJ/mol. Individual covalent bonds, 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. For every hydrogen bond formed in a protein during folding, a hydrogen bond (of similar strength) between the same group and water was broken. The net stability contributed by a given hydrogen bond, or the *difference* in free energies of the folded and unfolded states, may be close to zero. Ionic interactions may be either stabilizing or destabilizing. We must therefore look elsewhere to understand why a particular native conformation is favored.

On carefully examining the contribution of weak interactions to protein stability, we find that **hydrophobic interactions** generally predominate. Pure water contains a network of hydrogen-bonded H_2O molecules. No other molecule has the hydrogen-bonding potential of water, and the presence of other molecules in an aqueous solution disrupts 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 around the molecule (see Fig. 2–7). 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 cluster together, the extent of the solvation layer decreases, 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 increase in entropy is the major thermodynamic driving force for the association of hydrophobic groups in aqueous solution. Hydrophobic amino acid side chains therefore tend to cluster in a protein's interior, away from water (think of an oil droplet in water). The amino acid sequences of most proteins thus feature a significant content of hydrophobic amino acid side chains (especially Leu, Ile, Val, Phe, and Trp). These are positioned so that they are clustered when the protein is folded, forming a hydrophobic protein core.

Under physiological conditions, the formation of hydrogen bonds 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 layer also forms to some extent around polar molecules. Even though the energy of formation of an intramolecular hydrogen bond between two polar groups in a macromolecule is largely canceled by the elimination of such interactions between these polar groups and water, the release of structured water as intramolecular interactions form provides an entropic driving force for folding. Most of the net change in free energy as weak interactions form 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 its folded conformation.

Hydrophobic interactions are clearly important in stabilizing conformation; the interior of a structured 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 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 resulting from the combination of several such groups with partners in the surrounding solution can be greater than the free-energy difference between the folded and unfolded states. In addition, hydrogen bonds between groups in a protein form cooperatively (formation of one makes the next one more likely) in repeating secondary structures that optimize hydrogen bonding, as described below. In this way, hydrogen bonds often have an important role in guiding the protein-folding process.

The interaction of oppositely charged groups that form an ion pair, or salt bridge, can have either a stabilizing or destabilizing effect on protein structure. As in the case of hydrogen bonds, charged amino acid side chains interact with water and salts when the protein is unfolded, and the loss of those interactions must be considered when evaluating the effect of a salt bridge on the overall stability of a folded protein. However, the strength of a salt bridge increases as it moves to an environment of lower dielectric constant, ε (p. 50): from the polar aqueous solvent (ε near 80) to the nonpolar protein interior (ε near 4). Salt bridges, especially those that are partly or entirely buried, can thus provide significant stabilization to a protein structure. This trend explains the increased occurrence of buried salt bridges in the proteins of thermophilic organisms. Ionic interactions also limit structural flexibility and confer a uniqueness to a particular protein structure that nonspecific hydrophobic interactions cannot provide.

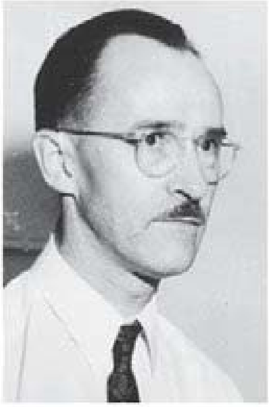
In the tightly packed atomic environment of a protein, one more type of weak interaction can have a significant effect—van der Waals interactions (p. 54). Van der Waals interactions are dipole-dipole interactions involving the permanent electric dipoles in groups such as carbonyls, transient dipoles derived from fluctuations of the electron cloud surrounding any atom, and dipoles induced by interaction of an atom with another that has a permanent or transient dipole. As atoms approach each other, these dipole-dipole interactions provide an attractive intermolecular force that operates only over a limited intermolecular distance (0.3 to 0.6 nm). Van der Waals interactions are weak and individually contribute little to overall protein stability. However, in a well-packed protein, or in an interaction between a protein and another protein or other molecule at a complementary surface, the number of such interactions can be substantial.

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 and ionic interactions within the protein is maximized, thus reducing the number of hydrogen-bonding and ionic groups that are not paired with a suitable partner. Proteins within membranes (which we examine in Chapter 11) and proteins that are intrinsically disordered or have intrinsically disordered segments follow different rules. This reflects their particular function or environment, but weak interactions are still critical structural elements. For example, soluble but intrinsically disordered protein segments are enriched in amino acid side chains that are charged (especially Arg, Lys, Glu) or small (Gly, Ala), providing little or no opportunity for the formation of a stable hydrophobic core.

The Peptide Bond Is Rigid and Planar

Protein Architecture—Primary Structure Covalent bonds, too, 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 current understanding of protein structure. They began with a careful analysis of the peptide bond.

Linus Pauling, 1901–1994



Robert Corey, 1897–1971

The α carbons of adjacent amino acid residues are separated by three covalent bonds, arranged as C_α—C—N—C_α. X-ray diffraction studies of crystals of amino acids and of simple dipeptides and tripeptides showed 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 hydrogen bonded to the nitrogen has a net 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 trans to the hydrogen atom of the amide nitrogen. From these findings Pauling and Corey concluded that the peptide C—N bonds, because of their partial double-bond character, cannot rotate freely. Rotation is permitted about the N—C_α and the C_α—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_α (Fig. 4–2b). The rigid peptide bonds limit the range of conformations possible for a polypeptide chain.

Peptide conformation is defined by three dihedral angles (also known as torsion angles) called ϕ (phi), ψ (psi), and ω (omega), reflecting rotation about each of the three repeating bonds in the peptide backbone. A dihedral angle is the angle at the intersection of two planes. In the case of peptides, the planes are defined by bond vectors in the peptide backbone. Two successive bond vectors describe a plane. Three successive bond vectors describe two planes (the central bond vector is common to both; Fig. 4–2c), and the angle between these two planes is what we measure to describe protein conformation.

KEY CONVENTION: The important dihedral angles in a peptide are defined by the three bond vectors connecting four consecutive main-chain (peptide backbone) atoms (Fig. 4–2c): ϕ involves the C—N—C_α—C bonds (with the rotation occurring about the N—C_α bond), and ψ involves the N—C_α—C—N bonds. Both ϕ and ψ are defined as ±180° when the polypeptide is fully extended and all peptide groups are in the same plane (Fig. 4–2d). As one looks down the central bond vector in the direction of the vector arrow (as depicted in Fig. 4–2c for ψ), the dihedral angles increase as the distal

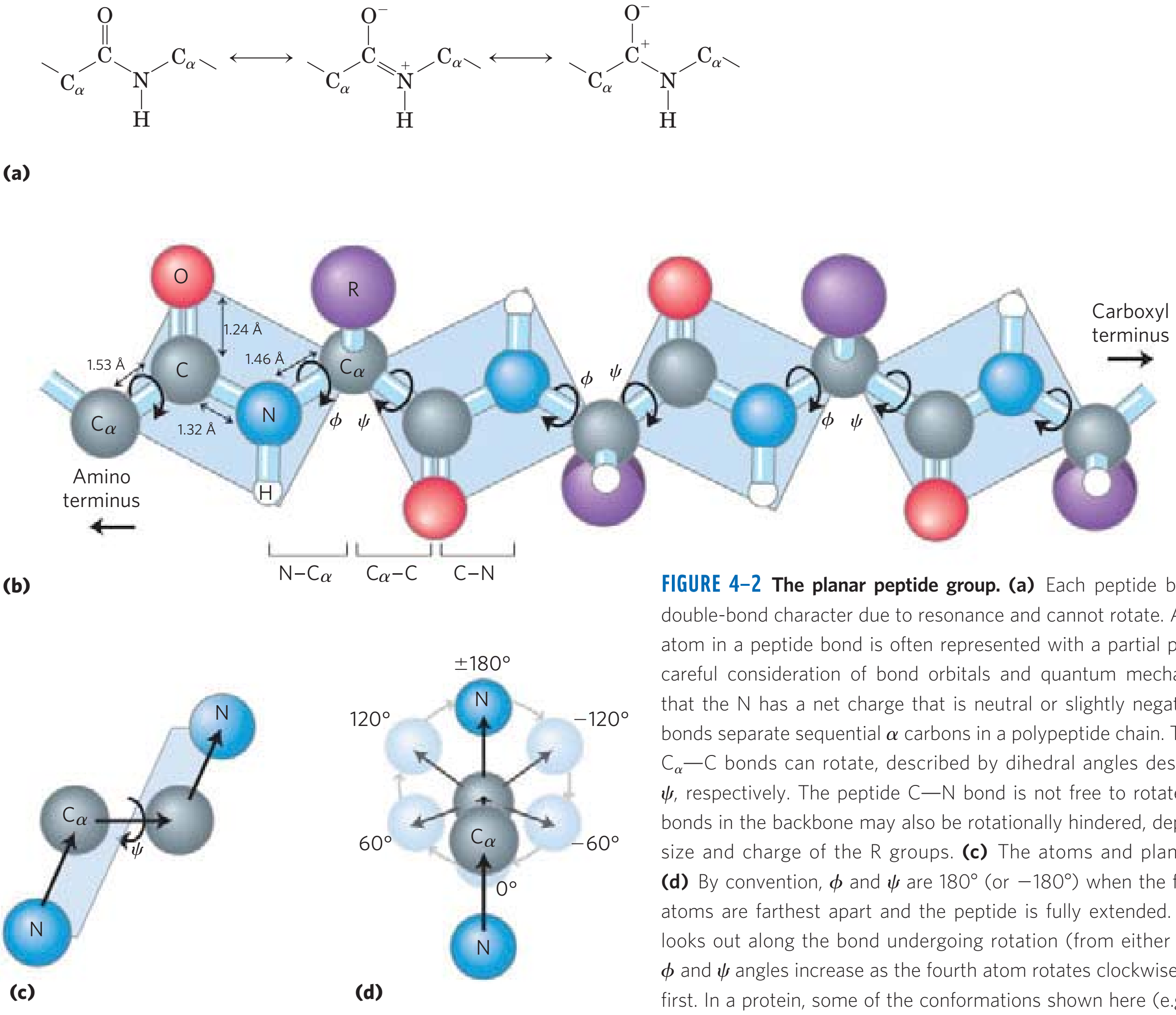


FIGURE 4–2 The planar peptide group. (a) Each peptide bond has some double-bond character due to resonance and cannot rotate. Although the N atom in a peptide bond is often represented with a partial positive charge, careful consideration of bond orbitals and quantum mechanics indicates that the N has a net charge that is neutral or slightly negative. **(b)** Three bonds separate sequential α carbons in a polypeptide chain. The N—C_α and C_α—C bonds can rotate, described by dihedral angles designated ϕ and ψ, 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. **(c)** The atoms and planes defining ψ. **(d)** By convention, ϕ and ψ are 180° (or −180°) when the first and fourth atoms are farthest apart and the peptide is fully extended. As the viewer looks out along the bond undergoing rotation (from either direction), the ϕ and ψ angles increase as the fourth atom rotates clockwise relative to the first. In a protein, some of the conformations shown here (e.g., 0°) are prohibited by steric overlap of atoms. In (b) through (d), the balls representing atoms are smaller than the van der Waals radii for this scale.

(fourth) atom is rotated clockwise (Fig. 4–2d). From the ±180° position, the dihedral angle increases from –180° to 0°, at which point the first and fourth atoms are eclipsed. The rotation can be continued from 0° to +180° (same position as –180°) to bring the structure back to the starting point. The third dihedral angle, ω, is not often considered. It involves the C_α—C—N—C_α bonds. The central bond in this case is the peptide bond, where rotation is constrained. The peptide bond is normally (99.6% of the time) in the trans configuration, constraining ω to a value of ±180°. For a rare cis peptide bond, $\omega = 0°$. ■

In principle, ϕ and ψ can have any value between –180° and +180°, but many values are prohibited by steric interference between atoms in the polypeptide backbone and amino acid side chains. The conformation in which both ϕ and ψ are 0° (Fig. 4–2d) is prohibited for this reason; this conformation is merely a reference point for describing the dihedral angles. Allowed values for ϕ and ψ become evident when ψ is plotted versus ϕ in a **Ramachandran plot (Fig. 4–3)**, introduced by G. N. Ramachandran. We will see that Ramachandran plots are very useful tools that are often used to test the quality of three-dimensional protein structures that are deposited in international databases.

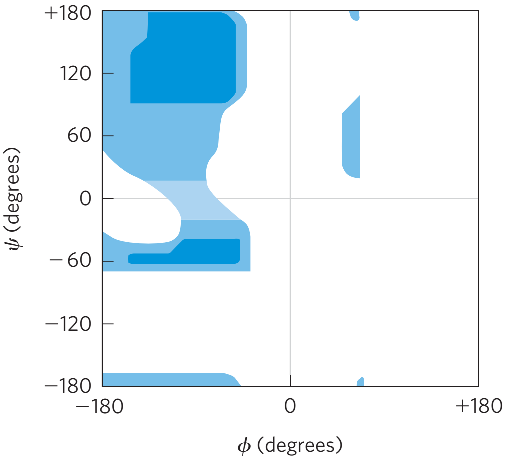


FIGURE 4–3 Ramachandran plot for L-Ala residues. Peptide conformations are defined by the values of ϕ and ψ. Conformations deemed possible are those that involve little or no steric interference, based on calculations using known van der Waals radii and dihedral angles. The areas shaded dark blue represent conformations that involve no steric overlap if the van der Waals radii of each atom are modeled as a hard sphere and thus are fully allowed; medium blue indicates conformations permitted if atoms are allowed to approach each other by an additional 0.1 nm, a slight clash; the lightest blue indicates conformations that are permissible if a very modest flexibility (a few degrees) is allowed in the ω dihedral angle that describes the peptide bond itself (generally constrained to 180°). The white regions are conformations that are not allowed. The asymmetry of the plot results from the L stereochemistry of the amino acid residues. The plots for other L residues with unbranched side chains are nearly identical. Allowed ranges for branched residues such as Val, Ile, and Thr are somewhat smaller than for Ala. The Gly residue, which is less sterically hindered, has a much broader range of allowed conformations. The range for Pro residues is greatly restricted because ϕ is limited by the cyclic side chain to the range of −35° to −85°.

SUMMARY 4.1 Overview of Protein Structure

- A typical protein usually has one or more stable three-dimensional structures, or conformations, that reflect its function. Some proteins have segments that are intrinsically disordered.
- Protein structure is stabilized largely by multiple weak interactions. Hydrophobic interactions, derived from the increase in entropy of the surrounding water when nonpolar molecules or groups are clustered together, are the major contributors to stabilizing the globular form of most soluble proteins. Van der Waals interactions also contribute. Hydrogen bonds and ionic interactions are optimized in the thermodynamically most stable structures.
- Nonpeptide covalent bonds, particularly disulfide bonds, play a role in the stabilization of structure in some proteins.
- 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_α and C_α—C bonds can rotate to define the dihedral angles ϕ and ψ, respectively.
- The Ramachandran plot is a visual description of the combinations of ϕ and ψ dihedral angles that are permitted in a peptide backbone or that are not permitted due to steric constraints.

4.2 Protein Secondary Structure

The term **secondary structure** refers to any chosen segment of a polypeptide chain and describes the local spatial arrangement of its main-chain atoms, without regard to the positioning of its side chains or its relationship to other segments. A *regular* secondary structure occurs when each dihedral angle, ϕ and ψ, remains the same or nearly the same throughout the segment. There are a few types of secondary structure that are particularly stable and occur widely in proteins. The most prominent are the α helix and β conformations; another common type is the β turn. Where a regular pattern is not found, the secondary structure is sometimes referred to as undefined or as a random coil. This last designation, however, does not properly describe the structure of these segments. The path of most of the polypeptide backbone in a typical protein is not random; rather, it is unchanging and highly specific to the structure and function of that particular protein. Our discussion here focuses on the regular structures that are most common.

The α Helix Is a Common Protein Secondary Structure

Protein Architecture—α Helix Pauling and Corey were aware of the importance of hydrogen bonds in orienting 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 α-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—it is approximately the length of a typical C—H bond.) 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 first breakthrough came in 1948. Pauling was a visiting lecturer at Oxford University, became ill, and retired to his apartment for several days of rest. Bored with the reading available, Pauling grabbed some paper and pencils to work out a plausible stable structure that could be taken up by a polypeptide chain. The model he developed, and later confirmed in work with Corey and coworker Herman Branson, was the simplest arrangement the polypeptide chain can assume that maximizes the use of internal hydrogen bonding. It is a helical structure, and Pauling and Corey called it the **α 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 backbone atoms of the amino acid residues in the prototypical α helix have a characteristic set of dihedral angles that define the α-helix conformation (Table 4–1), and each helical turn includes 3.6 amino acid residues. The α-helical segments in proteins often deviate slightly from these dihedral angles, and even vary somewhat within a single contiguous segment to produce subtle bends or kinks in the helical axis. Pauling and Corey considered both right- and left-handed variants of the α helix. The subsequent elucidation of the three-dimensional structure of myoglobin and other proteins showed that the right-handed α helix is the common form (Box 4–1). Extended left-handed α helices are theoretically less stable and have not been observed in proteins. The α helix proved to be the predominant structure in

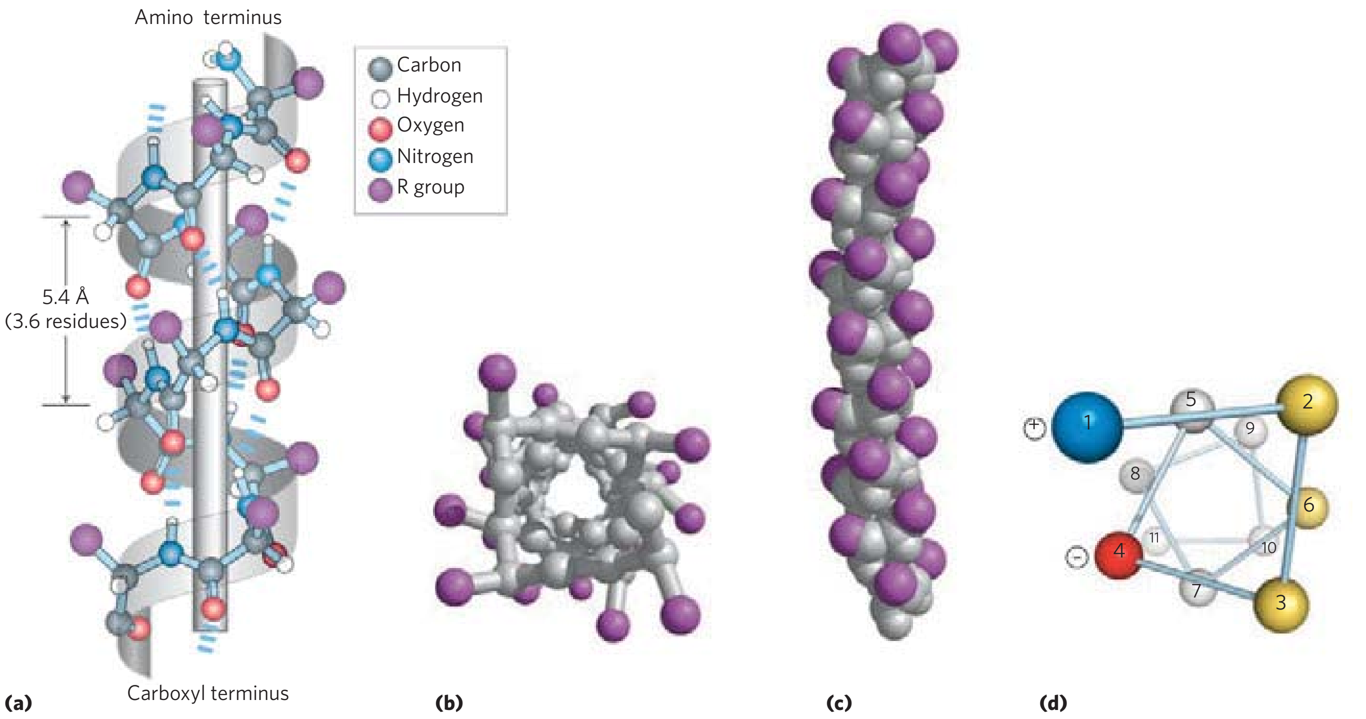


FIGURE 4–4 Models of the α helix, showing different aspects of its structure. (a) Ball-and-stick model showing the intrachain hydrogen bonds. The repeat unit is a single turn of the helix, 3.6 residues. **(b)** The α helix viewed from one end, looking down the longitudinal axis (derived from PDB ID 4TNC). Note the positions of the R groups, represented by purple spheres. This ball-and-stick model, which emphasizes 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. **(c)** As this space-filling model shows, the atoms in the center of the α helix are in very close contact. **(d)** Helical wheel projection of an α helix. This representation can be colored to identify surfaces with particular properties. The yellow residues, for example, could be hydrophobic and conform to an interface between the helix shown here and another part of the same or another polypeptide. The red (negative) and blue (positive) residues illustrate the potential for interaction of oppositely charged side chains separated by two residues in the helix.

TABLE 4–1 Idealized ϕ and ψ Angles for Common Secondary Structures in Proteins

Structure	ϕ	ψ
α Helix	−57°	−47°
β Conformation		
Antiparallel	−139°	+135°
Parallel	−119°	+113°
Collagen triple helix	−51°	+153°
β Turn type I		
i + 1*	−60°	−30°
i + 2*	−90°	0°
β Turn type II		
i + 1	−60°	+120°
i + 2	+80°	0°

Note: In real proteins, the dihedral angles often vary somewhat from these idealized values.

*The i+1 and i+2 angles are those for the second and third amino acid residues in the β turn, respectively.

α-keratins. More generally, about one-fourth of all amino acid residues in proteins are found in α helices, the exact fraction varying greatly from one protein to another.

Why does the α helix form more readily than many other possible conformations? The answer lies in part in its 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. 4–4a). Within the α helix, every peptide bond (except those close to each end of the helix) participates in such hydrogen bonding. Each successive turn of the α helix is held to adjacent turns by three to four hydrogen bonds, conferring significant stability on the overall structure. At the ends of an α-helical segment, there are always three or four amide carbonyl or amino groups that cannot participate in this helical pattern of hydrogen bonding. These may be exposed to the surrounding solvent, where they hydrogen-bond with water, or other parts of the protein may cap the helix to provide the needed hydrogen-bonding partners.

Further experiments have shown that an α 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. The most stable form of an α helix consisting of D-amino acids is left-handed.

WORKED EXAMPLE 4–1 Secondary Structure and Protein Dimensions

What is the length of a polypeptide with 80 amino acid residues in a single contiguous α helix?

Solution: An idealized α helix has 3.6 residues per turn and the rise along the helical axis is 5.4 Å. Thus, the rise along the axis for each amino acid residue is 1.5 Å. The length of the polypeptide is therefore 80 residues × 1.5 Å/residue = 120 Å.

Amino Acid Sequence Affects Stability of the α Helix

Not all polypeptides can form a stable α helix. Each amino acid residue in a polypeptide has an intrinsic propensity to form an α helix (Table 4–2), reflecting the properties of the R group and how they affect the capacity of the adjoining main-chain atoms to take up the characteristic ϕ and ψ angles. Alanine shows the greatest tendency to form α helices in most experimental model systems.

The position of an amino acid residue relative to its neighbors is also important. Interactions between amino

BOX 4–1 METHODS 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 away from you. 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 (clockwise). The resulting helix is right-handed. Your left hand will demonstrate a left-handed helix, which rotates in the counterclockwise direction as it spirals up your thumb.

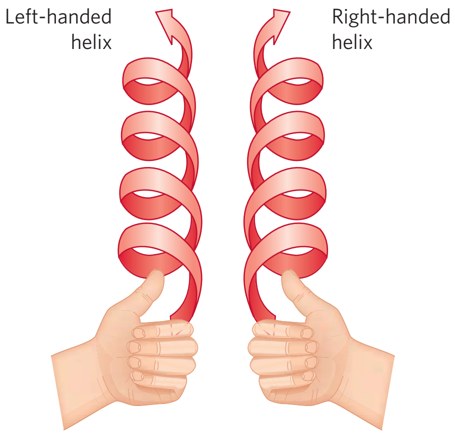


TABLE 4–2 Propensity of Amino Acid Residues to Take Up an α-Helical Conformation

Amino acid	$\Delta\Delta G°$ (kJ/mol)*	Amino acid	$\Delta\Delta G°$ (kJ/mol)*
Ala	0	Leu	0.79
Arg	0.3	Lys	0.63
Asn	3	Met	0.88
Asp	2.5	Phe	2.0
Cys	3	Pro	>4
Gln	1.3	Ser	2.2
Glu	1.4	Thr	2.4
Gly	4.6	Tyr	2.0
His	2.6	Trp	2.0
Ile	1.4	Val	2.1

Sources: Data (except proline) from Bryson, J.W., Betz, S.F., Lu, H.S., Suich, D.J., Zhou, H.X., O'Neil, K.T., & DeGrado, W.F. (1995) Protein design: a hierarchic approach. *Science* 270, 935. Proline data from Myers, J.K., Pace, C.N., & Scholtz, J.M. (1997) Helix propensities are identical in proteins and peptides. *Biochemistry* 36, 10,926.

*$\Delta\Delta G°$ is the difference in free-energy change, relative to that for alanine, required for the amino acid residue to take up the α-helical conformation. Larger numbers reflect greater difficulty taking up the α-helical structure. Data are a composite derived from multiple experiments and experimental systems.

acid side chains can stabilize or destabilize the α-helical structure. For example, if a polypeptide chain has a long block of Glu residues, this segment of the chain will not form an α 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 α helix. For the same reason, if there are many adjacent Lys and/or Arg residues, with positively charged R groups at pH 7.0, they also repel each other and prevent formation of the α helix. The bulk and shape of Asn, Ser, Thr, and Cys residues can also destabilize an α helix if they are close together in the chain.

The twist of an α 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. This is clear when the α helix is depicted as a helical wheel (Fig. 4–4d). 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.

A constraint on the formation of the α helix is the presence of Pro or Gly residues, which have the least proclivity to form α helices. In proline, the nitrogen atom is part of a rigid ring (see Fig. 4–8), and rotation about the N—C$_\alpha$ bond is not possible. Thus, a Pro residue introduces a destabilizing kink in an α helix. In addition, the nitrogen atom of a Pro residue in a peptide linkage has no substituent hydrogen to participate in hydrogen bonds with other residues. For these reasons, proline is only rarely found in an α helix. Glycine occurs infrequently in α 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 α helix.

A final factor affecting the stability of an α helix is the identity of the amino acid residues near the ends of the α-helical segment of the polypeptide. A small electric dipole exists in each peptide bond (Fig. 4–2a). These dipoles are aligned through the hydrogen bonds of the helix, resulting in a net dipole along the helical axis that increases with helix length **(Fig. 4–5)**. The partial positive and negative charges of the helix dipole reside on the peptide amino and carbonyl groups near the amino-terminal and carboxyl-terminal ends, 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.

In summary, five types of constraints affect the stability of an α helix: (1) the intrinsic propensity of an amino acid residue to form an α helix; (2) the interactions between R groups, particularly those spaced three (or four) residues apart; (3) the bulkiness of adjacent R groups; (4) the occurrence of Pro and Gly residues; and (5) interactions between amino acid residues at the ends of the helical segment and the electric dipole inherent to the α helix. The tendency of a given segment of a polypeptide chain to form an α helix therefore depends on the identity and sequence of amino acid residues within the segment.

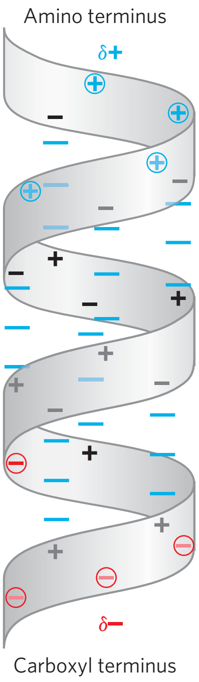


FIGURE 4–5 Helix dipole. The electric dipole of a peptide bond (see Fig. 4-2a) is transmitted along an α-helical segment through the intrachain hydrogen bonds, resulting in an overall helix dipole. In this illustration, the amino and carbonyl constituents of each peptide bond are indicated by + and − symbols, respectively. Non-hydrogen-bonded amino and carbonyl constituents of the peptide bonds near each end of the α-helical region are circled and shown in color.

The *β* Conformation Organizes Polypeptide Chains into Sheets

Protein Architecture—*β* Sheet In 1951, Pauling and Corey predicted a second type of repetitive structure, the ***β* conformation**. This is a more extended conformation of polypeptide chains, and its structure is again defined by backbone atoms arranged according to a characteristic set of dihedral angles (Table 4–1). In the *β* conformation, the backbone of the polypeptide chain is extended into a zigzag rather than helical structure **(Fig. 4–6)**. The arrangement of several segments side by side, all of which are in the *β* conformation, is called a ***β* sheet**. The zigzag structure of the individual polypeptide segments gives rise to a pleated appearance of the overall sheet. Hydrogen bonds form between adjacent segments of polypeptide chain within the sheet. The individual segments that form a *β* 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 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 view in Figure 4–6.

The adjacent polypeptide chains in a *β* 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, vs. 7 Å for antiparallel) and the hydrogen-bonding patterns are different. The interstrand hydrogen bonds are essentially in-line (see Fig. 2–5) in the antiparallel *β* sheet, whereas they are distorted or not in-line for the parallel variant. The idealized structures exhibit the bond angles given in Table 4–1; these values vary somewhat in real proteins, resulting in structural variation, as seen above for *α* helices.

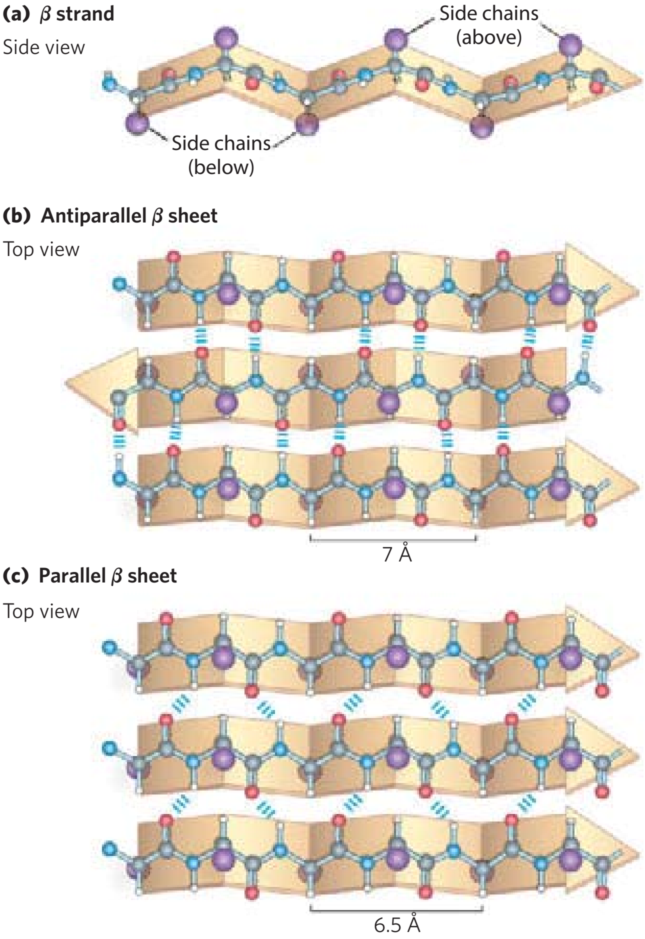


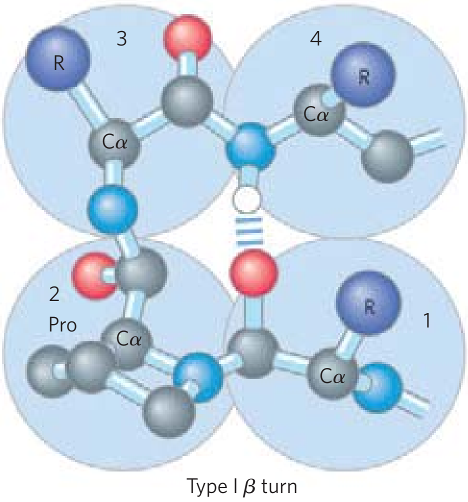
FIGURE 4–6 The *β* conformation of polypeptide chains. These **(a)** side and **(b, c)** top views reveal the R groups extending out from the *β* sheet and emphasize the pleated shape formed by the planes of the peptide bonds. (An alternative name for this structure is *β*-pleated sheet.) Hydrogen-bond cross-links between adjacent chains are also shown. The amino-terminal to carboxyl-terminal orientations of adjacent chains (arrows) can be the same or opposite, forming (b) an antiparallel *β* sheet or (c) a parallel *β* sheet.

β Turns Are Common in Proteins

Protein Architecture—*β* Turn In globular proteins, which have a compact folded structure, some amino acid residues are in turns or loops where the polypeptide chain reverses direction **(Fig. 4–7)**. These are the connecting elements that link successive runs of *α* helix or *β* conformation. Particularly common are ***β* turns** that connect the ends of two adjacent segments of an antiparallel *β* 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 inter-residue hydrogen bonding. Several types of *β* turns have been described, each defined by the ϕ and ψ angles of the bonds that link the four amino acid residues that make up the particular turn (Table 4–1). Gly and Pro residues often occur in *β* turns, the former because it is small and flexible, the latter because peptide bonds involving the imino nitrogen of proline readily assume the cis configuration **(Fig. 4–8)**, a form that is particularly amenable to a tight turn. The two types of *β* turns shown in Figure 4–7 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 γ turn, a three-residue turn with a hydrogen bond between the first and third residues.

Common Secondary Structures Have Characteristic Dihedral Angles

The *α* helix and the *β* conformation are the major repetitive secondary structures in a wide variety of proteins, although other repetitive structures exist in some specialized proteins (an example is collagen; see Fig. 4–13). Every type of secondary structure can be completely described by the dihedral angles ϕ and ψ associated with each residue. As shown by a Ramachandran plot, the dihedral angles that define the *α* helix and *β* conformation fall within a relatively restricted range of



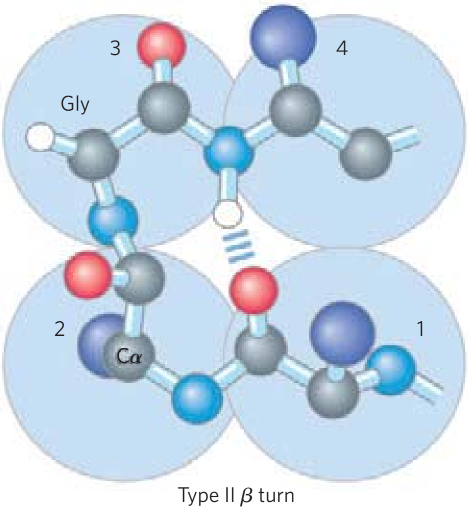


FIGURE 4–7 Structures of β turns. Type I and type II β turns are most common, distinguished by the ϕ and ψ angles taken up by the peptide backbone in the turn (see Table 4-1). Type I turns occur more than twice as frequently as type II. Type II β turns usually have Gly as the third residue. Note the 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. Not all H atoms are shown in these depictions.)

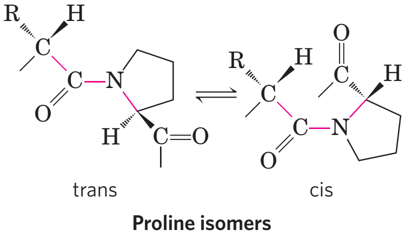


FIGURE 4–8 Trans and cis isomers of a peptide bond involving the imino nitrogen of proline. Of the peptide bonds between amino acid residues other than Pro, more than 99.95% are in the trans configuration. For peptide bonds involving the imino nitrogen of proline, however, about 6% are in the cis configuration; many of these occur at β turns.

sterically allowed structures **(Fig. 4–9a)**. Most values of ϕ and ψ taken from known protein structures fall into the expected regions, with high concentrations near the α helix and β 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 is small, a Gly residue can take part in many conformations that are sterically forbidden for other amino acids.

Common Secondary Structures Can Be Assessed by Circular Dichroism

Any form of structural asymmetry in a molecule gives rise to differences in absorption of left-handed versus right-handed circularly polarized light. Measurement of this difference is called **circular dichroism (CD)**

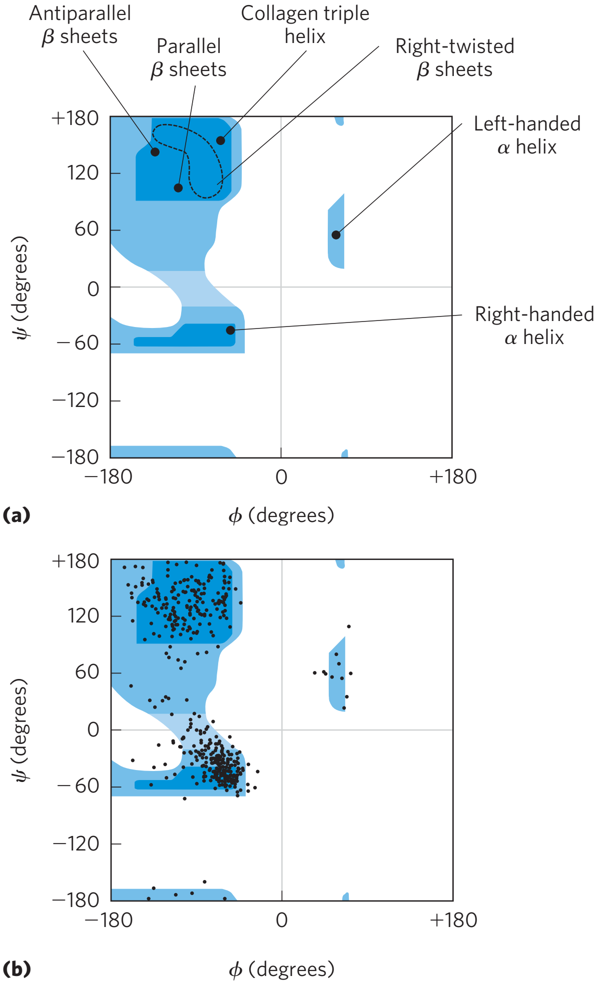


FIGURE 4–9 Ramachandran plots showing a variety of structures. **(a)** The values of ϕ and ψ for various allowed secondary structures are overlaid on the plot from Figure 4-3. Although left-handed α helices extending over several amino acid residues are theoretically possible, they have not been observed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

spectroscopy. An ordered structure, such as a folded protein, gives rise to an absorption spectrum that can have peaks or regions with both positive and negative values. For proteins, spectra are obtained in the far UV region (190 to 250 nm). The light-absorbing entity, or chromophore, in this region is the peptide bond; a signal is obtained when the peptide bond is in a folded environment. The difference in molar extinction coefficients (see Box 3–1) for left- and right-handed, circularly polarized light ($\Delta\varepsilon$) is plotted as a function of wavelength. The α helix and β conformations have characteristic CD spectra **(Fig. 4–10)**. Using CD spectra, biochemists can determine whether proteins are properly folded, estimate the fraction of the protein that is folded in either of the common secondary structures, and monitor transitions between the folded and unfolded states.

SUMMARY 4.2 Protein Secondary Structure

- Secondary structure is the local spatial arrangement of the main-chain atoms in a selected segment of a polypeptide chain.
- The most common regular secondary structures are the α helix, the β conformation, and β turns.
- The secondary structure of a polypeptide segment can be completely defined if the ϕ and ψ angles are known for all amino acid residues in that segment.
- Circular dichroism spectroscopy is a method for assessing common secondary structure and monitoring folding in proteins.

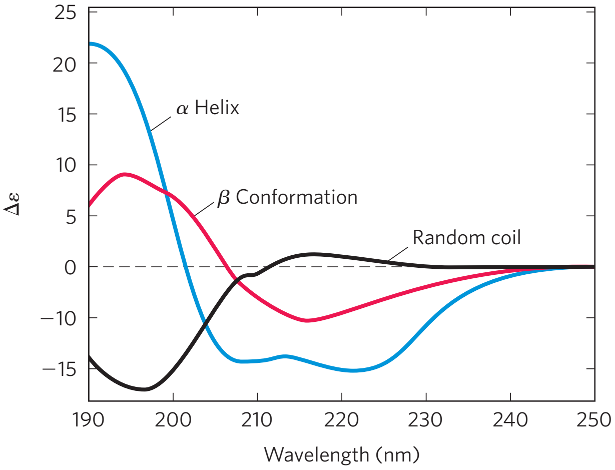


FIGURE 4–10 Circular dichroism spectroscopy. These spectra show polylysine entirely as α helix, as β conformation, or as a denatured, random coil. The *y* axis unit is a simplified version of the units most commonly used in CD experiments. Since the curves are different for α helix, β conformation, and random coil, the CD spectrum for a given protein can provide a rough estimate for the fraction of the protein made up of the two most common secondary structures. The CD spectrum of the native protein can serve as a benchmark for the folded state, useful for monitoring denaturation or conformational changes brought about by changes in solution conditions.

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 a segment of a polypeptide, tertiary structure includes *longer-range* aspects of amino acid sequence. Amino acids that are far apart in the polypeptide sequence and are in different types of secondary structure may interact within the completely folded structure of a protein. The location of bends (including β 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 several kinds of weak 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 designate two major groups into which many proteins can be classified: **fibrous proteins**, with polypeptide chains arranged in long strands or sheets, and **globular proteins**, with 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, and their tertiary structure is relatively simple. Globular proteins often contain several types of secondary structure. The two groups also differ functionally: 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.

Fibrous Proteins Are Adapted for a Structural Function

Protein Architecture—Tertiary Structure of Fibrous Proteins α-Keratin, collagen, and silk fibroin nicely illustrate the relationship between protein structure and biological function (Table 4–3). 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 as many similar polypeptide chains are packed 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.

α-Keratin The α-keratins have evolved for strength. Found only 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 α-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 the structural features exemplified by the α-keratins.

The α-keratin helix is a right-handed α helix, the same helix found in many other proteins. Francis Crick and Linus Pauling in the early 1950s independently suggested that the α helices of keratin were arranged as a coiled coil. Two strands of α-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 α helix to form a coiled coil explains the discrepancy between the 5.4 Å per turn predicted for an α 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 α helix. The surfaces where the two α 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, α-keratin is rich in the hydrophobic residues Ala, Val, Leu, Ile, Met, and Phe.

An individual polypeptide in the α-keratin coiled coil has a relatively simple tertiary structure, dominated by an α-helical secondary structure with its helical axis twisted in a left-handed superhelix. The intertwining of the two α-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. 5–27). The quaternary structure of α-keratin can be quite complex. Many coiled coils can be assembled into large supramolecular complexes, such as the arrangement of α-keratin to form the intermediate filament of hair (Fig. 4–11b).

The strength of fibrous proteins is enhanced by covalent cross-links between polypeptide chains in the multihelical "ropes" and between adjacent chains in a supramolecular assembly. In α-keratins, the cross-links stabilizing quaternary structure are disulfide bonds (Box 4–2). In the hardest and toughest α-keratins, such

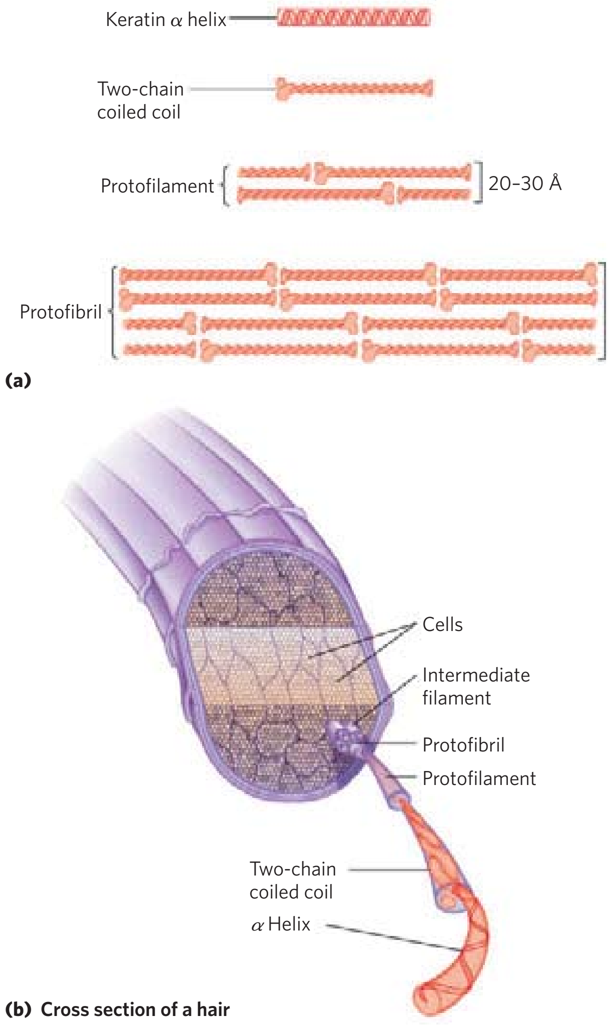


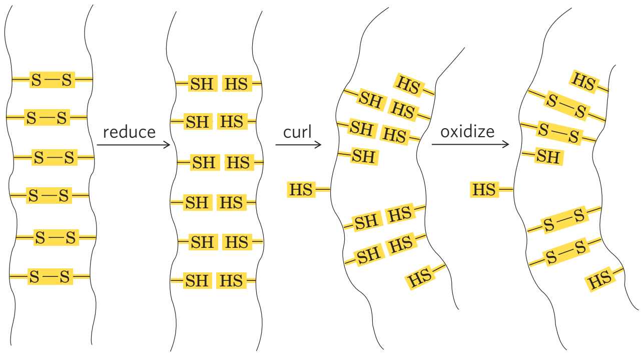
FIGURE 4–11 Structure of hair. (a) Hair α-keratin is an elongated α helix with somewhat thicker elements near the amino and carboxyl termini. Pairs of these helices are interwound in a left-handed sense to form two-chain coiled coils. These then combine in higher-order structures called protofilaments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

TABLE 4–3 Secondary Structures and Properties of Some Fibrous Proteins

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

BOX 4–2 Permanent Waving Is Biochemical Engineering

When hair is exposed to moist heat, it can be stretched. At the molecular level, the α helices in the α-keratin of hair are stretched out until they arrive at the fully extended β conformation. On cooling they spontaneously revert to the α-helical conformation. The characteristic "stretchability" of α-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 (—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 α-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 α-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 α-helical coils in the hair fibers. The same process can be used to straighten hair that is naturally curly. A permanent wave (or hair straightening) is not truly permanent, because the hair grows; in the new hair replacing the old, the α-keratin has the natural pattern of disulfide bonds.



as those of rhinoceros horn, up to 18% of the residues are cysteines involved in disulfide bonds.

Collagen Like the α-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 α helix. It is left-handed and has three amino acid residues per turn (**Fig. 4–12** and Table 4–1). Collagen is also a coiled coil, but one with distinct tertiary and quaternary structures: three separate polypeptides, called α chains (not to be confused with α 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 α 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 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 α 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).

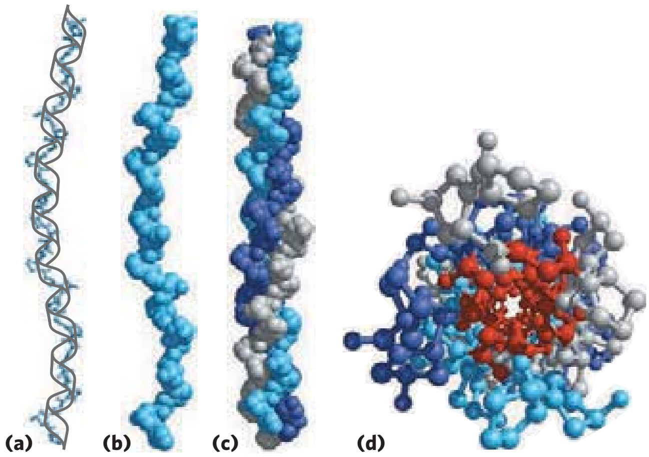


FIGURE 4–12 Structure of collagen. (Derived from PDB ID 1CGD) **(a)** The α chain of collagen has a repeating secondary structure unique to this protein. The repeating tripeptide sequence Gly–X–Pro or Gly–X–4-Hyp adopts a left-handed helical structure with three residues per turn. The repeating sequence used to generate this model is Gly–Pro–4-Hyp. **(b)** Space-filling model of the same α 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 very tightly packed.

BOX 4–3 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 . . .

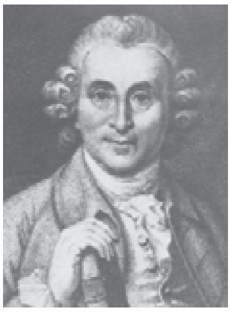
—The Memoirs of the Lord of Joinville, *ca. 1300*

This excerpt describes the plight of Louis IX's army toward the end of the Seventh Crusade (1248–1254), when 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. Milder cases of vitamin 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. Yet it did not come to wide public notice until the European voyages of discovery from 1500 to 1800. The first circumnavigation of the globe (1519–1522), led by Ferdinand Magellan, was accomplished only with the loss of more than 80% of his crew to scurvy. During Jacques Cartier's second voyage to explore the St. Lawrence River (1535–1536), his band 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). 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, 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. 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.


James Lind, 1716–1794; naval surgeon, 1739–1748

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 value in the United States is 90 mg (Australia and the United Kingdom recommend 60 mg; Russia recommends 125 mg). Along with citrus fruits and almost all other fresh fruits, 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. 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—the proline derivative (4*R*)-L-hydroxyproline, which plays an essential role in the folding of collagen and in maintaining its structure. The proline ring is normally found as a mixture of two puckered conformations, called C$_\gamma$-endo and C$_\gamma$-exo (Fig. 1). The collagen helix structure requires the Pro residue in the Y positions to

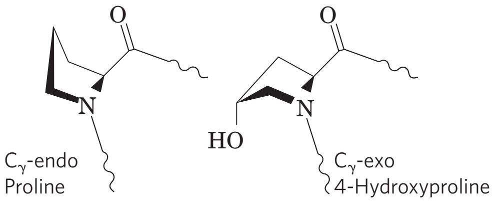


FIGURE 1 The C_γ-endo conformation of proline and the C_γ-exo conformation of 4-hydroxyproline.

be in the C_γ-exo conformation, and it is this conformation that is enforced by the hydroxyl substitution at C-4 in 4-Hyp. The collagen structure also requires that the Pro residue in the X positions have the C_γ-endo conformation, and introduction of 4-Hyp here can destabilize the helix. In the absence of vitamin C, cells cannot hydroxylate the Pro at the Y positions. This leads to collagen instability and the connective tissue problems seen in scurvy.

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 vertebrates. The proline-hydroxylating activity is found in the α subunits. Each α subunit contains one atom of nonheme iron (Fe^{2+}), and the enzyme is one of a class of hydroxylases that require α-ketoglutarate in their reactions.

In the normal prolyl 4-hydroxylase reaction (Fig. 2a), one molecule of α-ketoglutarate and one of O_2 bind to the enzyme. The α-ketoglutarate is oxidatively decarboxylated to form 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 α-ketoglutarate that is not coupled to proline hydroxylation (Fig. 2b). During this reaction the heme Fe^{2+} becomes oxidized, inactivating the enzyme and preventing the proline hydroxylation. The ascorbate consumed in the reaction is needed to restore enzyme activity—by reducing the heme iron.

Scurvy remains a problem today, 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 fruits and vegetables.

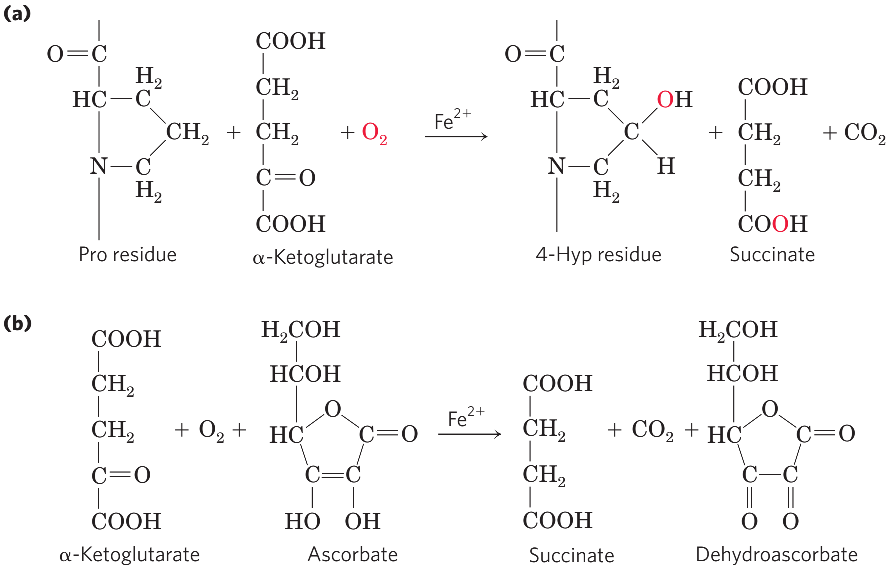
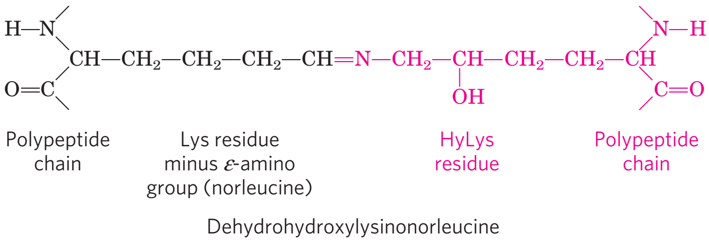


FIGURE 2 Reactions catalyzed by prolyl 4-hydroxylase. **(a)** The normal reaction, coupled to proline hydroxylation, which does not require ascorbate. The fate of the two oxygen atoms from O_2 is shown in red. **(b)** The uncoupled reaction, in which α-ketoglutarate is oxidatively decarboxylated without hydroxylation of proline. Ascorbate is consumed stoichiometrically in this process as it is converted to dehydroascorbate, preventing Fe^{2+} oxidation.

The tight wrapping of the α chains in the collagen triple helix provides tensile strength greater than that 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 α 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. 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.


H—N
CH—CH₂—CH₂—CH₂—CH=N—CH₂—CH—CH₂—CH₂—CH
O=C
OH
N—H
C=O
Polypeptide chain
Lys residue minus ε-amino group (norleucine)
HyLys residue
Polypeptide chain
Dehydrohydroxylysinonorleucine

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; at least eight variants of this condition, with different degrees of severity, occur in the human population. Ehlers-Danlos syndrome is characterized by loose joints, and at least six variants occur in humans. The composer Niccolò Paganini (1782–1840) was famed for his seemingly impossible dexterity in playing the violin. He suffered from a variant of Ehlers-Danlos syndrome that rendered him effectively double-jointed. In both disorders, some variants can be lethal, whereas others cause lifelong problems.

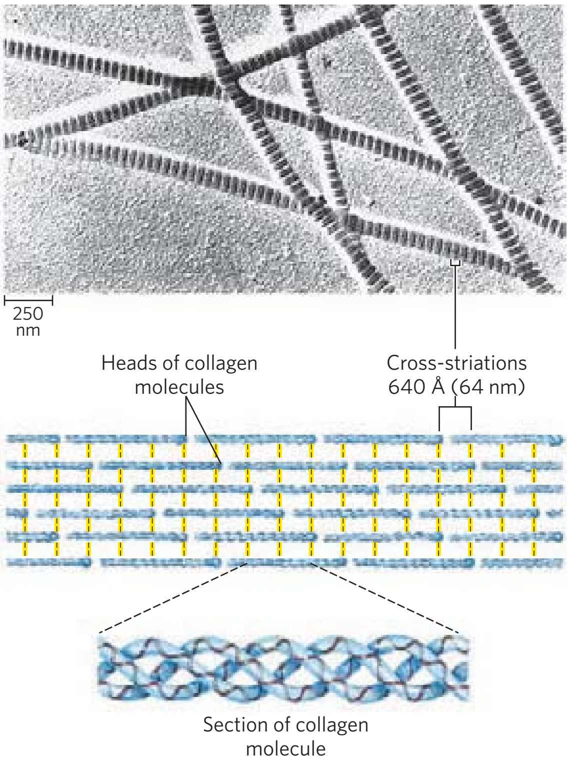

250 nm
Heads of collagen molecules
Cross-striations 640 Å (64 nm)
Section of collagen molecule

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 α chains may have different sequences; each chain 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 (64 nm) apart.

All of the variants of both conditions 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 an α chain in one or another collagen protein (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 β conformation. Fibroin is rich in Ala and Gly residues, permitting a close packing of β 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 β sheet and by the optimization of van der Waals interactions between sheets. Silk does not stretch, because the β conformation is already highly extended (Fig. 4–6). 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 α-keratins.

Structural Diversity Reflects Functional Diversity in Globular Proteins

In a globular protein, different segments of the polypeptide chain (or multiple polypeptide chains) fold back on each other, generating a more compact shape than is seen in the fibrous proteins **(Fig. 4–15)**. The folding also provides the structural diversity necessary for proteins to carry out a wide array of biological functions.

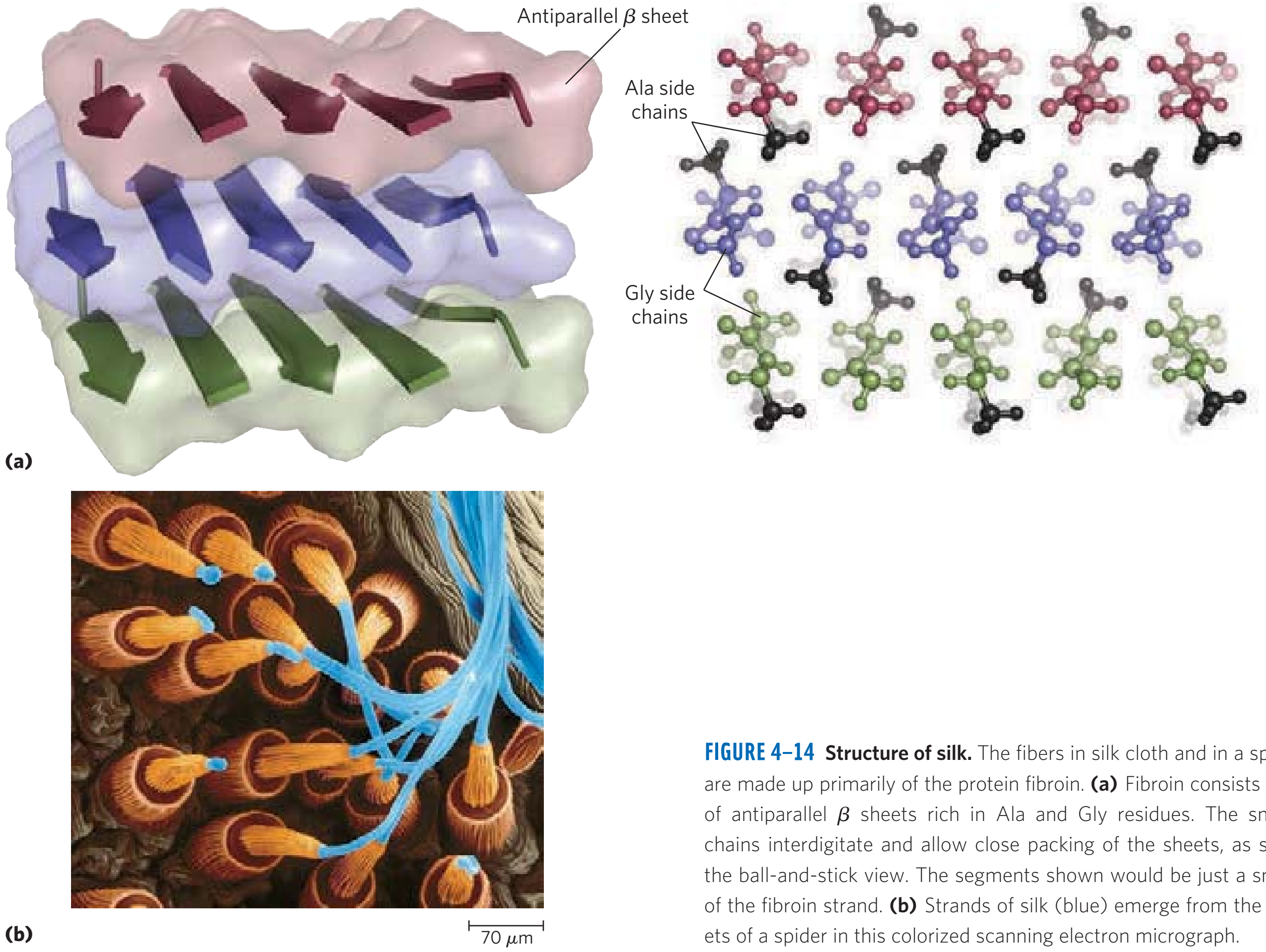


FIGURE 4–14 Structure of silk. The fibers in silk cloth and in a spiderweb are made up primarily of the protein fibroin. **(a)** Fibroin consists of layers of antiparallel β sheets rich in Ala and Gly residues. The small side chains interdigitate and allow close packing of the sheets, as shown in the ball-and-stick view. The segments shown would be just a small part of the fibroin strand. **(b)** Strands of silk (blue) emerge from the spinnerets of a spider in this colorized scanning electron micrograph.

Globular proteins include enzymes, transport proteins, motor proteins, regulatory proteins, immunoglobulins, and proteins with many other functions.

Our discussion of globular proteins begins with the principles gleaned from the first 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 publicly accessible databases, particularly the Protein Data Bank (Box 4–4).

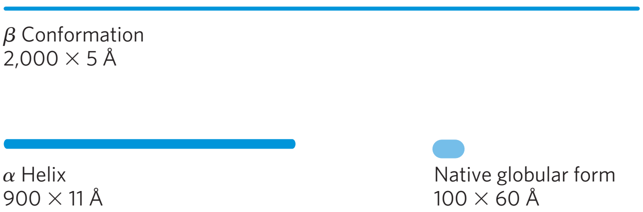


FIGURE 4–15 Globular protein structures are compact and varied. Human serum albumin (M_r 64,500) has 585 residues in a single chain. Given here are the approximate dimensions its single polypeptide chain would have if it occurred entirely in extended β conformation or as an α helix. Also shown is the 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.

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 that is found in myoglobin 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 abundant in the muscles of diving mammals such as the whale, seal, and porpoise—so abundant that the muscles of these animals are brown. Storage and distribution of oxygen by muscle myoglobin permits diving mammals to remain submerged for long periods. The activities of

BOX 4–4 The Protein Data Bank

The number of known three-dimensional protein structures is now in the tens of thousands and more than doubles every couple of years. This wealth of information is revolutionizing our understanding of protein structure, the relation of structure to function, and the evolutionary paths by which proteins arrived at their present state, which can be seen in the family resemblances that come to light as protein databases are sifted and sorted. One of the most important resources available to biochemists is the **Protein Data Bank** (**PDB**; www.pdb.org).

The PDB is an archive of experimentally determined three-dimensional structures of biological macromolecules, containing virtually all of the macromolecular structures (proteins, RNAs, DNAs, etc.) elucidated to date. Each structure is assigned an identifying label (a four-character identifier called the PDB ID). Such labels are provided in the figure legends for every PDB-derived structure illustrated in this text so that students and instructors can explore the same structures on their own. The data files in the PDB describe the spatial coordinates of each atom whose position has been determined (many of the cataloged structures are not complete). Additional data files provide information on how the structure was determined and its accuracy. The atomic coordinates can be converted into an image of the macromolecule by using structure visualization software. Students are encouraged to access the PDB and explore structures, using visualization software linked to the database. Macromolecular structure files can also be downloaded and explored on the desktop, using free software such as Jmol.

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 consists of eight relatively straight segments of α helix interrupted by bends, some of which are β turns. The longest α 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 α-helical regions. X-ray analysis has revealed the precise position of each of the R groups, which fill up nearly all the space within the folded chain that is not occupied by backbone atoms.

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 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 globular protein the fraction is about 0.75, comparable to that in a crystal (in a typical crystal the fraction is 0.70 to 0.78, near the theoretical maximum). 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.

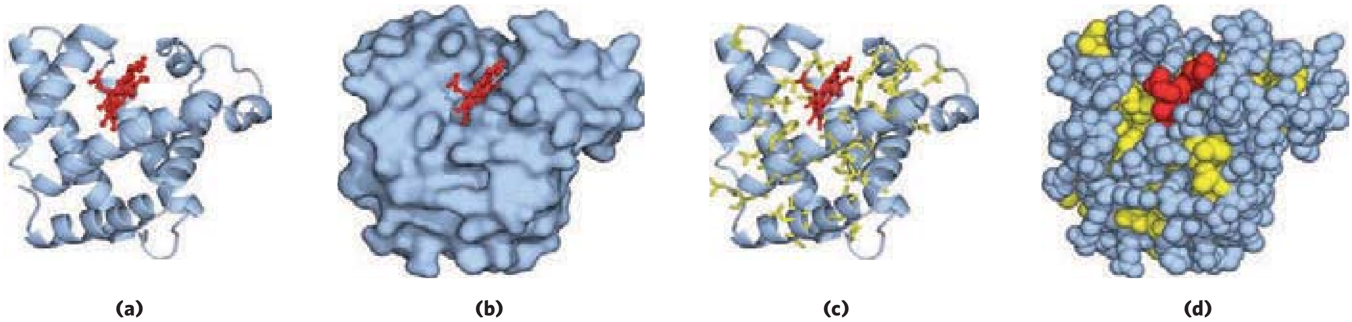


FIGURE 4–16 Tertiary structure of sperm whale myoglobin. (PDB ID 1MBO) Orientation of the protein is similar in (a) through (d); the heme group is shown in red. In addition to illustrating the myoglobin structure, this figure provides examples of several different ways to display protein structure. **(a)** The polypeptide backbone in a ribbon representation of a type introduced by Jane Richardson, which highlights regions of secondary structure. The α-helical regions are evident. **(b)** Surface contour image; this is useful for visualizing pockets in the protein where other molecules might bind. **(c)** Ribbon representation including side chains (yellow) for the hydrophobic residues Leu, Ile, Val, and Phe. **(d)** 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 yellow; most are buried in the interior of the protein and thus not visible.

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 α helices in myoglobin provided the first direct experimental evidence for the existence of this type of secondary structure. Three of the four Pro residues are found at bends. The fourth Pro residue occurs within an α helix, where it creates a kink necessary for tight helix packing.

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.

As many different myoglobin structures were resolved, investigators were able to observe the structural changes that accompany the binding of oxygen or other molecules and thus, for the first time, to understand the correlation between protein structure and function. Hundreds of proteins have now been subjected to similar analysis. Today, nuclear magnetic resonance (NMR) spectroscopy and other techniques supplement x-ray diffraction data, providing more information on a protein's structure (Box 4–5). In addition, the sequencing of the genomic DNA of many organisms (Chapter 9) has identified thousands of genes that encode proteins of known sequence but, as yet, unknown function; this work continues apace.

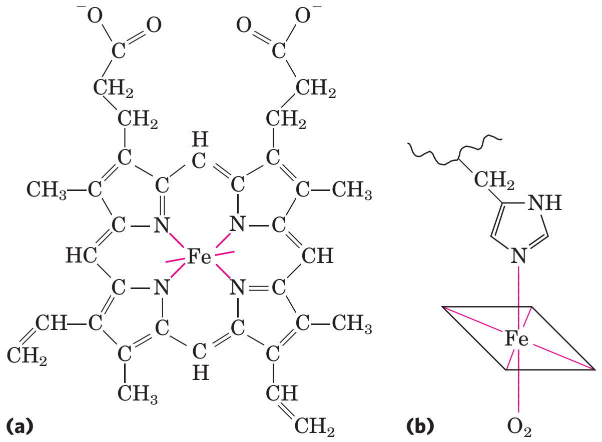


FIGURE 4–17 The heme group. This group is present in myoglobin, hemoglobin, cytochromes, and many other proteins (the heme proteins). **(a)** Heme consists of a complex organic ring structure, protoporphyrin, which binds an iron atom in its ferrous (Fe^{2+}) state. The iron atom has six coordination 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.

Globular Proteins Have a Variety of Tertiary Structures

From what we now know about the tertiary structures of hundreds of globular proteins, it is clear that myoglobin illustrates just one of many ways in which a polypeptide chain can fold. Table 4–4 shows the proportions of α helix and β conformations (expressed as percentage of residues in each type) 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 with myoglobin. Each is folded compactly, and in each case the hydrophobic 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.

For the beginning student, the very complex tertiary structures of globular proteins—some much larger than myoglobin—are best approached by focusing on common structural patterns, recurring 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 α-helical and β conformations, linked by connecting segments. The structure can then be defined by how these segments stack on one another and how the segments that connect them are arranged.

To understand a complete three-dimensional structure, we need to analyze its folding patterns. We begin by defining two important terms that describe protein structural patterns or elements in a polypeptide chain and then turn to the folding rules.

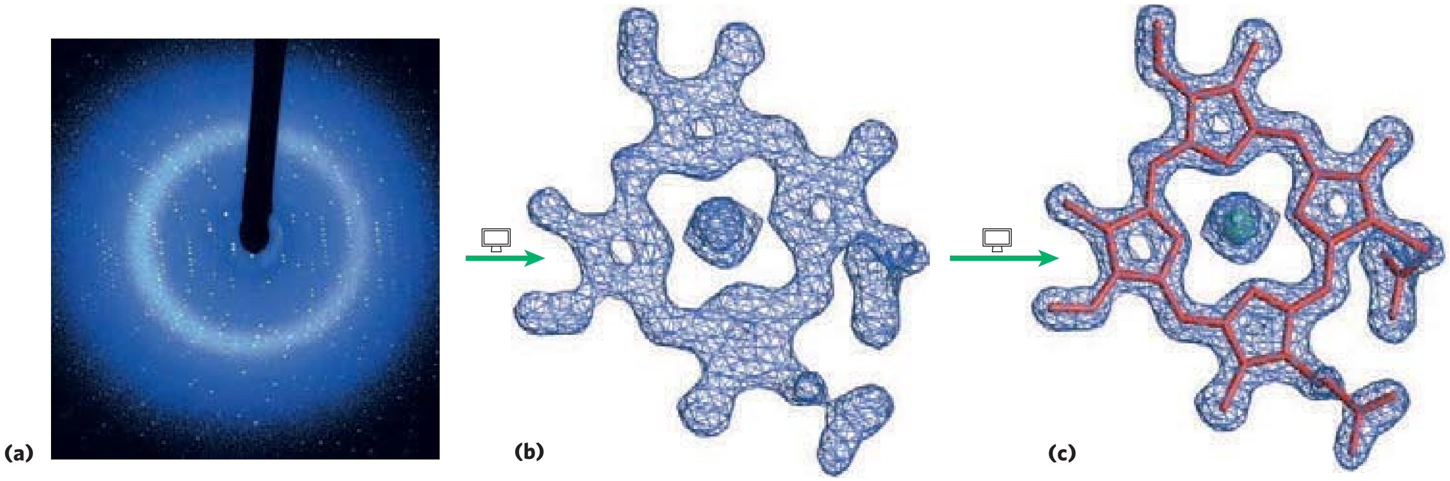
TABLE 4–4 Approximate Proportion of α Helix and β Conformation in Some Single-Chain Proteins

	Residues (%)*	
Protein (total residues)	α Helix	β Conformation
Chymotrypsin (247)	14	45
Ribonuclease (124)	26	35
Carboxypeptidase (307)	38	17
Cytochrome *c* (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 not accounted for by α helix or β conformation consist of bends and irregularly coiled or extended stretches. Segments of α helix and β conformation sometimes deviate slightly from their normal dimensions and geometry.

BOX 4–5 METHODS 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 Na^+ and 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.

Consider how images are generated in a light microscope. Light from a point source is focused on an object. The object scatters the light waves, 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 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. The structures of many 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). A crystal is placed in an x-ray beam between the x-ray source and a detector, and a regular array of spots called reflections 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 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

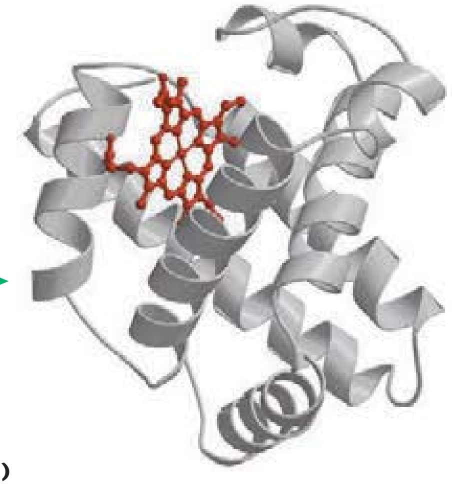


FIGURE 1 Steps in determining the structure of sperm whale myoglobin by x-ray crystallography. **(a)** X-ray diffraction patterns are generated from a crystal of the protein. **(b)** Data extracted from the diffraction patterns are used to calculate a three-dimensional electron-density map. The electron density of only part of the structure, the heme, is shown here. **(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).

myoglobin, have since been determined to a similar level of resolution.

The physical environment in 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 advantage of nuclear magnetic resonance (NMR) studies is that they are 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}H, ^{13}C, ^{15}N, ^{19}F, and ^{31}P, have 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 (~10 μs) 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 on a sample are averaged, increasing the signal-to-noise ratio, and an NMR spectrum such as that in Figure 2 is generated.

^{1}H is particularly important in NMR experiments because of its high sensitivity and natural abundance. For macromolecules, ^{1}H NMR spectra can become quite complicated. Even a small protein has hundreds of ^{1}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).

(*Continued on next page*)

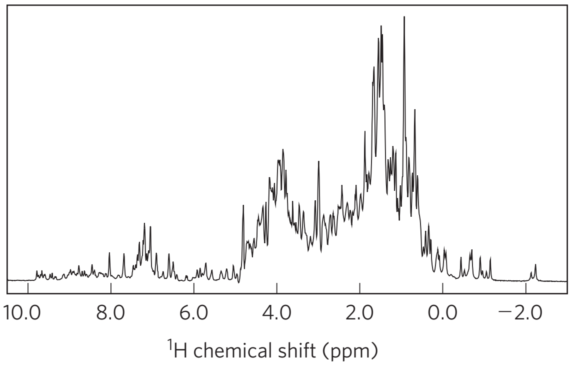


FIGURE 2 One-dimensional NMR spectrum of a globin from a marine blood worm. This protein and sperm whale myoglobin are very close structural analogs, belonging to the same protein structural family and sharing an oxygen-transport function.

BOX 4–5 METHODS Methods for Determining the Three-Dimensional Structure of a Protein (*Continued*)

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 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 α helices. Genetic engineering (Chapter 9) can be used to prepare proteins that contain the rare isotopes ^{13}C or ^{15}N. The new NMR signals produced by these atoms, and the coupling with ^{1}H signals resulting from these substitutions, help in the assignment of individual ^{1}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 (known as breathing) within a protein structure in solution, discussed in more detail in Chapter 5. Normal experimental uncertainty can also play a role.

Protein structures determined by both x-ray crystallography and NMR 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.

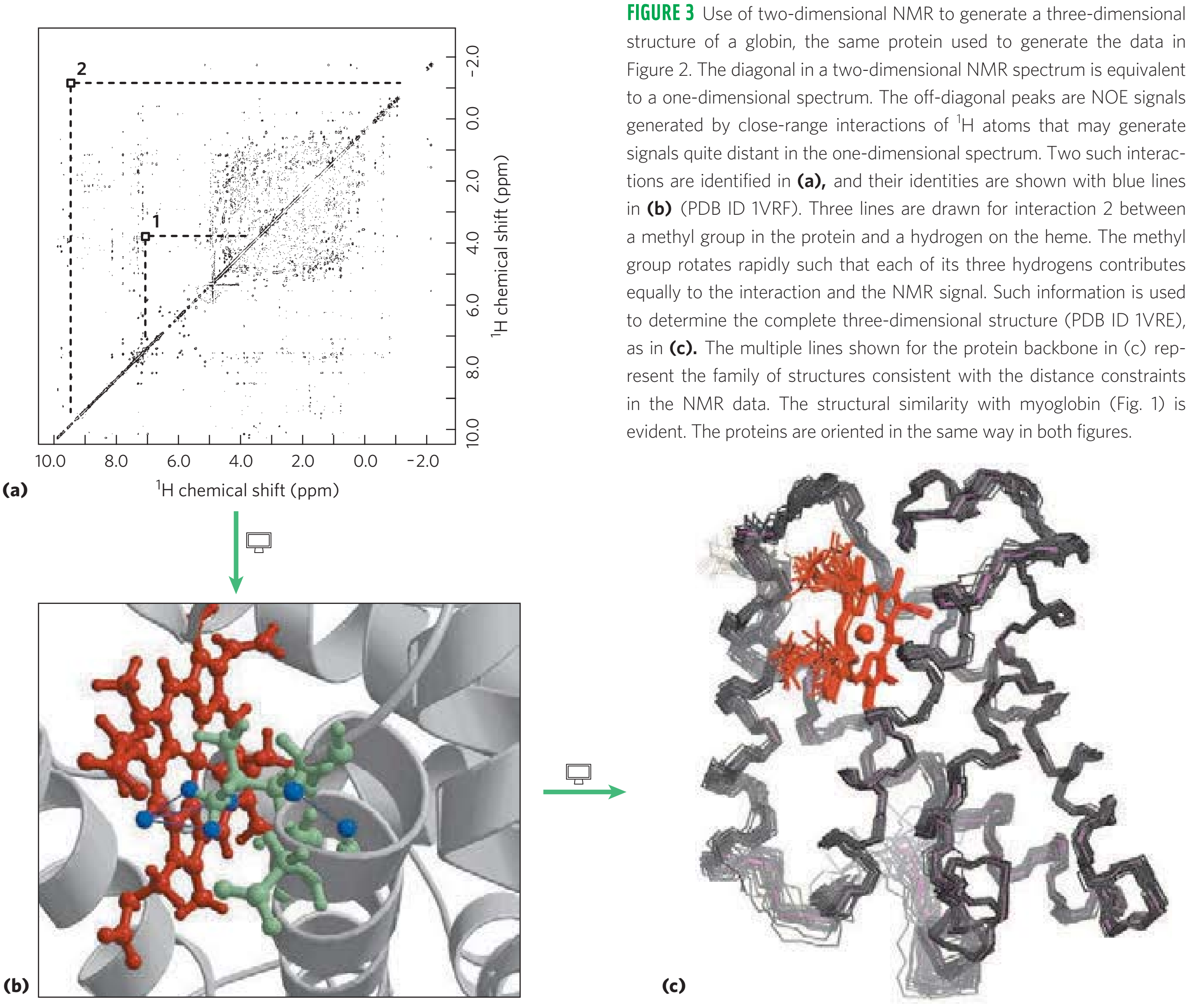


FIGURE 3 Use of two-dimensional NMR to generate a three-dimensional structure of a globin, the same protein used to generate the data in Figure 2. The diagonal in a two-dimensional NMR spectrum is equivalent to a one-dimensional spectrum. The off-diagonal peaks are NOE signals generated by close-range interactions of ^{1}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 in (c) 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.

The first term is **motif**, also called a **fold** or (more rarely) **supersecondary structure**. *A motif or fold is a recognizable folding pattern involving two or more elements of secondary structure and the connection(s) between them.* A motif can be very simple, such as two elements of secondary structure folded against each other, and represent only a small part of a protein. An example is a ***β-α-β* loop** **(Fig. 4–18a)**. A motif can also be a very elaborate structure involving scores of protein segments folded together, such as the *β* barrel (Fig. 4–18b). In some cases, a single large motif may comprise the entire protein. The terms "motif" and "fold" are often used interchangeably, although "fold" is applied more commonly to somewhat more complex folding patterns. The terms encompass any advantageous folding pattern and are useful for describing such patterns. The segment defined as a motif or fold may or may not be independently stable. We have already encountered a well-studied motif, the coiled coil of *α*-keratin, which is also found in some other proteins. The distinctive arrangement of eight *α* helices in myoglobin is replicated in all globins and is called the globin fold. Note that a motif is not a hierarchical structural element falling between secondary and tertiary structure. It is simply a folding pattern. The synonymous term "supersecondary structure" is thus somewhat misleading because it suggests hierarchy.

The second term for describing structural patterns is **domain**. A domain, as defined by Jane Richardson in 1981, is a part of a polypeptide chain that is independently stable or could undergo movements as a single entity with respect to the entire protein. Polypeptides with more than a few hundred amino acid residues often fold into two or more domains, sometimes with different functions. In many cases, a domain from a large protein will retain its native three-dimensional structure even when separated (for example, by proteolytic cleavage) from the remainder of the polypeptide chain. In a protein with multiple domains, each domain may appear as a distinct globular lobe **(Fig. 4–19)**; 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).

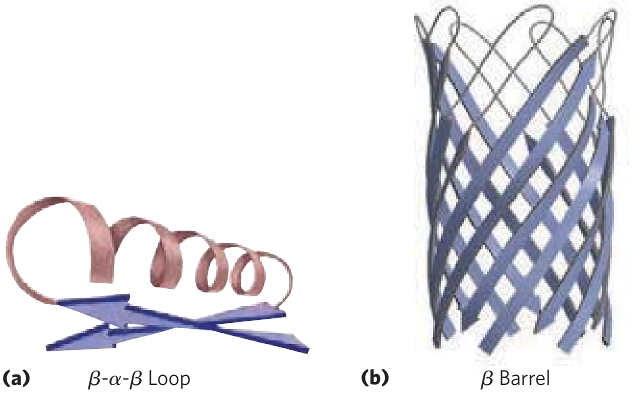


FIGURE 4–18 Motifs. (a) A simple motif, the *β-α-β* loop. **(b)** A more elaborate motif, the *β* barrel. This *β* barrel is a single domain of *α*-hemolysin (a toxin that kills a cell by creating a hole in its membrane) from the bacterium *Staphylococcus aureus* (derived from PDB ID 7AHL).

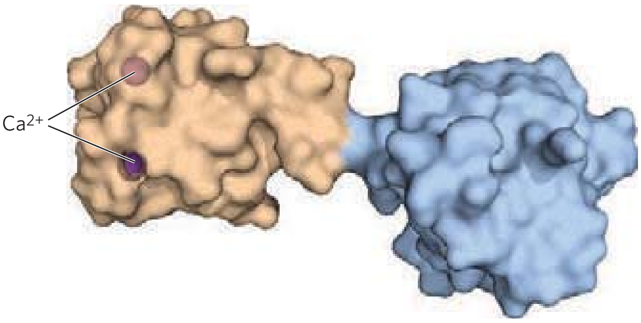


FIGURE 4–19 Structural domains in the polypeptide troponin C. (PDB ID 4TNC) This calcium-binding protein, associated with muscle, has two separate calcium-binding domains, shown here in brown and blue.

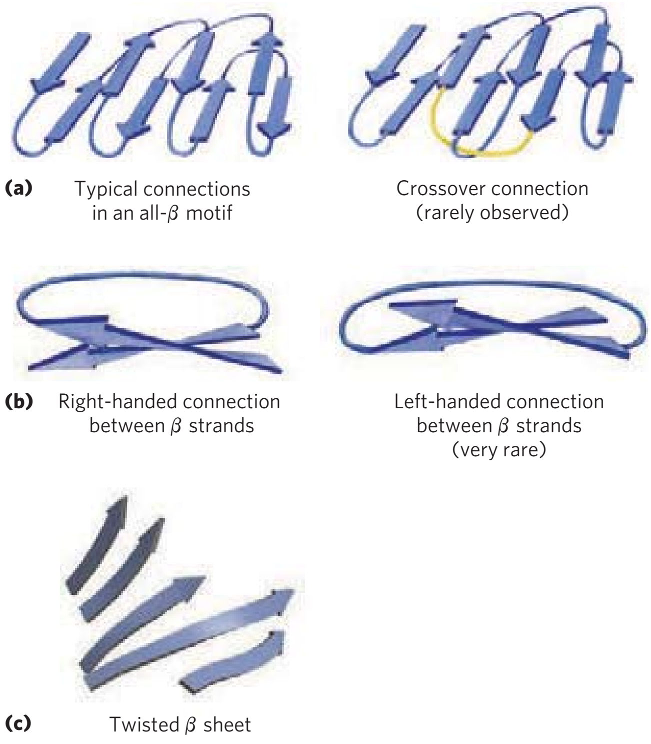
Folding of polypeptides is subject to an array of physical and chemical constraints, and several rules have emerged from studies of common protein folding patterns.

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. Simple motifs, such as the *β-α-β* loop (Fig. 4–18a), create two such layers.

2. Where they occur together in a protein, *α* helices and *β* sheets generally are found in different structural layers. This is because the backbone of a polypeptide segment in the *β* conformation (Fig. 4–6) cannot readily hydrogen-bond to an *α* helix that is adjacent to it.

3. Segments adjacent to each other in the amino acid sequence are usually stacked adjacent to each other in the folded structure. Distant segments of a polypeptide may come together in the tertiary structure, but this is not the norm.

4. The *β* conformation is most stable when the individual segments are twisted slightly in a right-handed sense. This influences both the arrangement of *β* sheets derived from the twisted segments and the path of the polypeptide connections between them. Two parallel *β* strands, for example, must be connected by a crossover strand **(Fig. 4–20a)**. 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



(a) Typical connections in an all-β motif

Crossover connection (rarely observed)

(b) Right-handed connection between β strands

Left-handed connection between β strands (very rare)

(c) Twisted β sheet

FIGURE 4–20 Stable folding patterns in proteins. (a) Connections between β strands in layered β sheets. The strands here are viewed from one end, with no twisting. Thick lines represent connections at the ends nearest the viewer; thin lines are connections at the far ends of the β strands. The connections at a given end (e.g., near the viewer) rarely cross one another. An example of such a rare crossover is illustrated by the yellow strand in the structure on the right. **(b)** Because of the right-handed twist in β strands, connections between strands are generally right-handed. Left-handed connections must traverse sharper angles and are harder to form. **(c)** This twisted β sheet is from a domain of photolyase (a protein that repairs certain types of DNA damage) from *E. coli* (derived from PDB ID 1DNP). Connecting loops have been removed so as to focus on the folding of the β sheet.

form. The twisting of β sheets also leads to a characteristic twisting of the structure formed by many such segments together, as seen in the β barrel (Fig. 4–18b) and twisted β sheet (Fig. 4–20c), 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 β-α-β loops arranged so that the β strands form a barrel creates a particularly stable and common motif, the ***α/β* barrel (Fig. 4–21)**. In this structure, each parallel β segment is attached to its neighbor by an α-helical segment. All connections are right-handed. The α/β barrel is found in many enzymes, often with a binding site (for a cofactor or substrate) in the form of a pocket near one end of the barrel. Note that domains with similar folding patterns are said to have the same motif even though their constituent α helices and β sheets may differ in length.

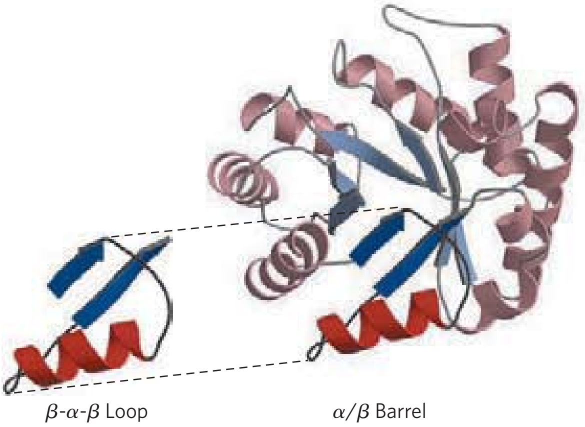


FIGURE 4–21 Constructing large motifs from smaller ones. The α/β barrel is a commonly occurring motif constructed from repetitions of the β-α-β loop motif. This α/β barrel is a domain of pyruvate kinase (a glycolytic enzyme) from rabbit (derived from PDB ID 1PKN).

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, understanding the complexities of tertiary structure is made easier by considering substructures. Taking this idea further, researchers have organized the complete contents of protein databases according to hierarchical levels of structure. All of these databases rely on data and information deposited in the Protein Data Bank. The Structural Classification of Proteins (SCOP) database is a good example of this 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, with four classes of protein structure: **all *α***, **all *β***, ***α/β*** (with α and β segments interspersed or alternating), and ***α* + *β*** (with α and β regions somewhat segregated). Each class includes tens to hundreds of different folding arrangements (motifs), 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** shows a variety of motifs arrayed among the four classes of protein structure; this is 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. 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 (see color key in Fig. 4–22), categorization is based on evolutionary relationships.

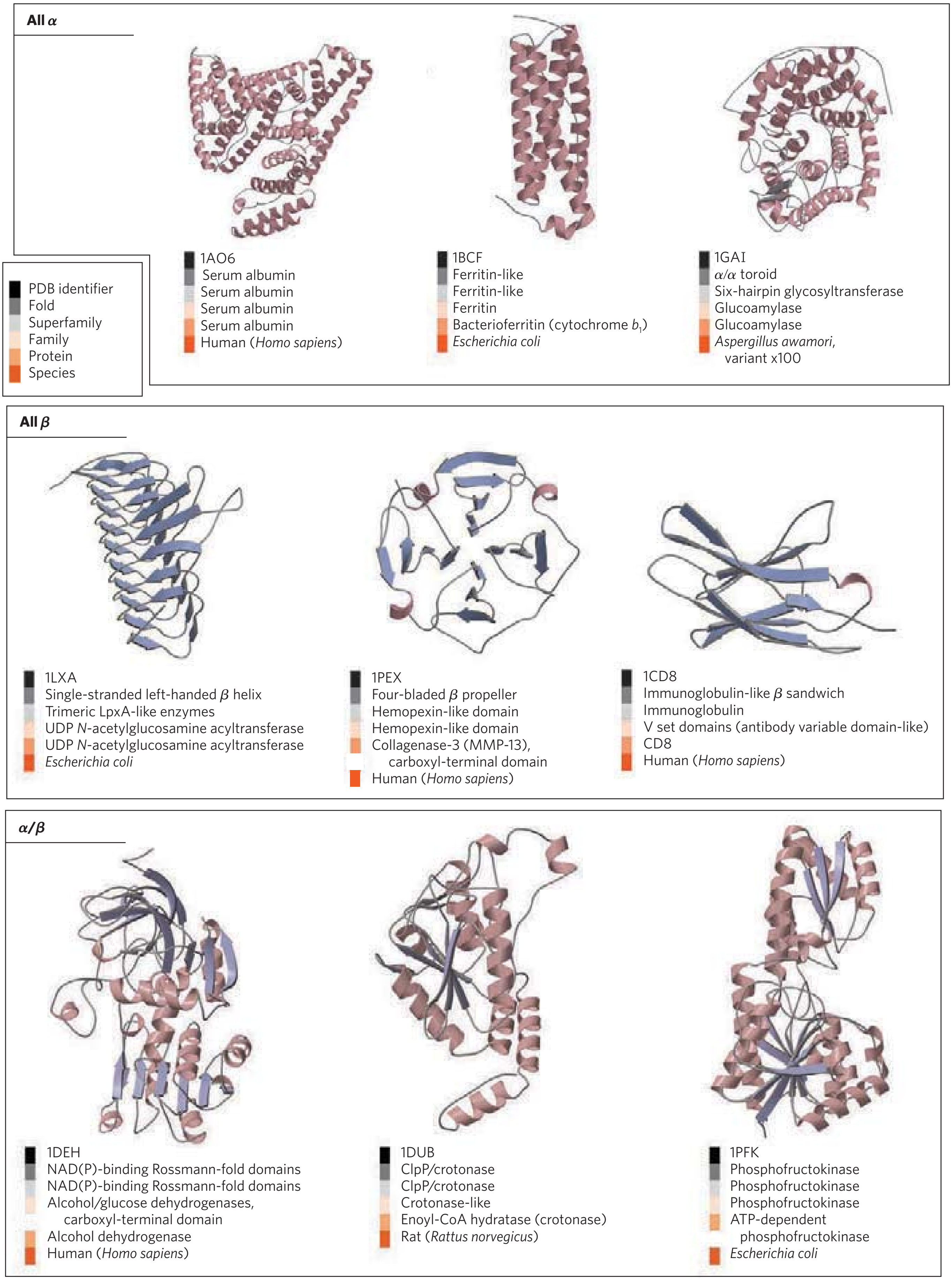

All α
1AO6
Serum albumin
Serum albumin
Serum albumin
Serum albumin
Human (*Homo sapiens*)
1BCF
Ferritin-like
Ferritin-like
Ferritin
Bacterioferritin (cytochrome b_1)
Escherichia coli
1GAI
α/α toroid
Six-hairpin glycosyltransferase
Glucoamylase
Glucoamylase
Aspergillus awamori, variant x100
PDB identifier
Fold
Superfamily
Family
Protein
Species
All β
1LXA
Single-stranded left-handed β helix
Trimeric LpxA-like enzymes
UDP *N*-acetylglucosamine acyltransferase
UDP *N*-acetylglucosamine acyltransferase
Escherichia coli
1PEX
Four-bladed β propeller
Hemopexin-like domain
Hemopexin-like domain
Collagenase-3 (MMP-13), carboxyl-terminal domain
Human (*Homo sapiens*)
1CD8
Immunoglobulin-like β sandwich
Immunoglobulin
V set domains (antibody variable domain-like)
CD8
Human (*Homo sapiens*)
α/β
1DEH
NAD(P)-binding Rossmann-fold domains
NAD(P)-binding Rossmann-fold domains
Alcohol/glucose dehydrogenases, carboxyl-terminal domain
Alcohol dehydrogenase
Human (*Homo sapiens*)
1DUB
ClpP/crotonase
ClpP/crotonase
Crotonase-like
Enoyl-CoA hydratase (crotonase)
Rat (*Rattus norvegicus*)
1PFK
Phosphofructokinase
Phosphofructokinase
Phosphofructokinase
ATP-dependent phosphofructokinase
Escherichia coli

FIGURE 4–22 Organization of proteins based on motifs. Shown here are a few of the hundreds of known stable motifs divided into four classes: all α, all β, α/β, 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 (see the color key). The PDB identifier (listed first for each structure) is the unique accession code given to each structure archived in the Protein Data Bank (www.pdb.org). The α/β, barrel (see Fig. 4–21) is another particularly common α/β motif.

(Continued on next page)

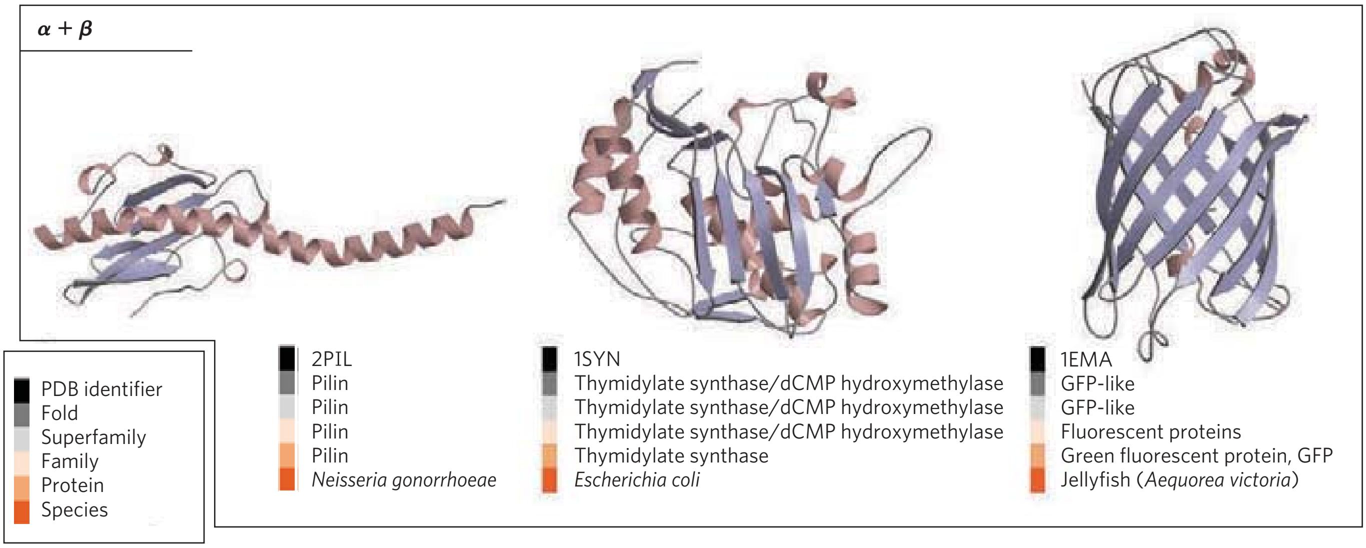

α + β
PDB identifier
Fold
Superfamily
Family
Protein
Species
2PIL
Pilin
Pilin
Pilin
Pilin
Neisseria gonorrhoeae
1SYN
Thymidylate synthase/dCMP hydroxymethylase
Thymidylate synthase/dCMP hydroxymethylase
Thymidylate synthase/dCMP hydroxymethylase
Thymidylate synthase
Escherichia coli
1EMA
GFP-like
GFP-like
Fluorescent proteins
Green fluorescent protein, GFP
Jellyfish (*Aequorea victoria*)

FIGURE 4–22 ***(Continued)***

Many examples of recurring domain or motif structures are available, and these reveal that protein tertiary structure is more reliably conserved than amino acid sequence. The comparison of protein structures can thus provide much information about evolution. Proteins with significant similarity in primary structure and/or with similar tertiary 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 4–5 and in Chapter 5). Two or more families with little similarity in amino acid sequence sometimes make use of the same major structural motif and have functional similarities; these families are grouped as **superfamilies**. An evolutionary relationship among families in a superfamily is considered probable, even though time and functional distinctions—that is, 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 an ancient origin. Many proteins involved in intermediary metabolism and the metabolism of nucleic acids and proteins fall into this category. 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 evolutionary relationships. The SCOP database is curated manually, with the objective of placing proteins in the correct evolutionary framework based on conserved structural features.

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 (from two to hundreds). 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 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. For example, each ribosome, the site of protein synthesis, incorporates dozens of protein subunits along with a number of RNA molecules.

A multisubunit protein is also referred to as a **multimer**. A multimer with just a few subunits is often called an **oligomer**. If a multimer has 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 a single subunit or a group of subunits, is called a

protomer. Greek letters are sometimes used to distinguish the individual subunits that make up a protomer.

The first oligomeric protein to have its three-dimensional structure 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 (Fe^{2+}) state (Fig. 4–17). The protein portion, the globin, consists of two α chains (141 residues each) and two β chains (146 residues each). Note that in this case, α and β do not refer to secondary structures. In a practice that can be confusing to the beginning student, the Greek letters α and β (and γ and δ, and others) are often used to distinguish two different kinds of subunits within a multisubunit protein, regardless of what kinds of secondary structure may predominate in the subunits. 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 α and one β subunit. Hemoglobin can therefore be described either as a tetramer or as a dimer of $\alpha\beta$ protomers. The role these distinct subunits play in hemoglobin function is discussed extensively in Chapter 5.

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

Some Proteins or Protein Segments Are Intrinsically Disordered

In spite of decades of progress in the understanding of protein structure, many proteins cannot be crystallized, making it difficult to determine their three-dimensional structure by methods now considered classical (see Box 4–5). Even where crystallization succeeds, parts of the protein are often sufficiently disordered within the crystal that the determined structure does not include those parts. Sometimes, this is due to subtle features of the structure that render crystallization difficult. However, the reason can be more straightforward: some proteins or protein segments lack an ordered structure in solution.

The concept that some proteins function in the absence of a definable structure is a product of the reassessment of data involving many different proteins. As many as a third of all human proteins may be unstructured or have significant unstructured segments. All organisms have some proteins that fall into this category. **Intrinsically disordered proteins** have properties that are distinct from classical structured proteins. They lack a hydrophobic core, and instead are characterized by high densities of charged amino acid residues such as Lys, Arg, and Glu. Pro residues are also prominent, as they tend to disrupt ordered structures.

Structural disorder and high charge density can facilitate the function of some proteins as spacers, insulators, or linkers in larger structures. Other disordered proteins are scavengers, binding up ions and small molecules in solution and serving as reservoirs or garbage dumps. However, many intrinsically disordered proteins are at the heart of important protein interaction networks. The lack of an ordered structure can facilitate a kind of functional promiscuity, allowing one protein to interact with multiple partners. Some intrinsically disordered proteins act to inhibit the action of other proteins by an unusual mechanism: wrapping around their protein targets. One disordered protein may have several or even dozens of protein partners. The structural disorder allows the inhibitor protein to wrap around the multiple

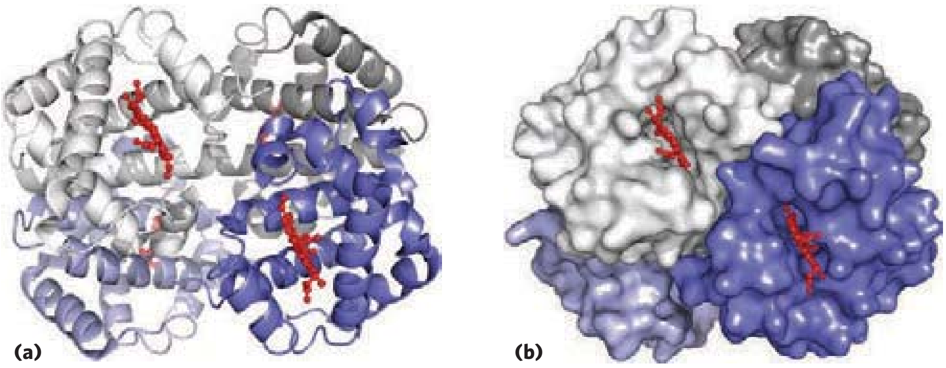


FIGURE 4–23 Quaternary structure of deoxyhemoglobin. (PDB ID 2HHB) X-ray diffraction analysis of deoxyhemoglobin (hemoglobin without oxygen molecules bound to the heme groups) shows how the four polypeptide subunits are packed together. **(a)** A ribbon representation reveals the secondary structural elements of the structure and the positioning of all the heme cofactors. **(b)** A surface contour model shows the pockets in which the heme cofactors are bound and helps to visualize subunit packing. The α subunits are shown in shades of gray; the β subunits in shades of blue. Note that the heme groups (red) are relatively far apart.

targets in different ways. The intrinsically disordered protein p27 plays a key role in controlling mammalian cell division. This protein lacks definable structure in solution. It wraps around and thus inhibits the action of several enzymes called protein kinases (see Chapter 6) that facilitate cell division. The flexible structure of p27 allows it to accommodate itself to its different target proteins. Human tumor cells, which are simply cells that have lost the capacity to control cell division normally, generally have reduced levels of p27; the lower the levels of p27, the poorer the prognosis for the cancer patient. Similarly, intrinsically disordered proteins are often present as hubs or scaffolds at the center of protein networks that constitute signaling pathways. These proteins, or parts of them, may interact with many different binding partners. They often take on a structure when they interact with other proteins, but the structure they assume may vary with different binding partners. The mammalian protein p53 is also critical in the control of cell division. It features both structured and unstructured segments, and the different segments interact with dozens of other proteins. An unstructured region of p53 at the carboxyl terminus interacts with at least four different binding partners and assumes a different structure in each of the complexes **(Fig. 4–24)**.

SUMMARY 4.3 Protein Tertiary and Quaternary Structures

- Tertiary structure is the complete three-dimensional structure of a polypeptide chain. Many proteins fall into one of 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, by x-ray diffraction methods, was that of myoglobin.
- The complex structures of globular proteins can be analyzed by examining folding patterns called motifs (also called folds or supersecondary structures). The thousands of known protein structures are generally assembled from a repertoire of only a few hundred motifs. Domains are regions of a polypeptide chain that can fold stably and independently.

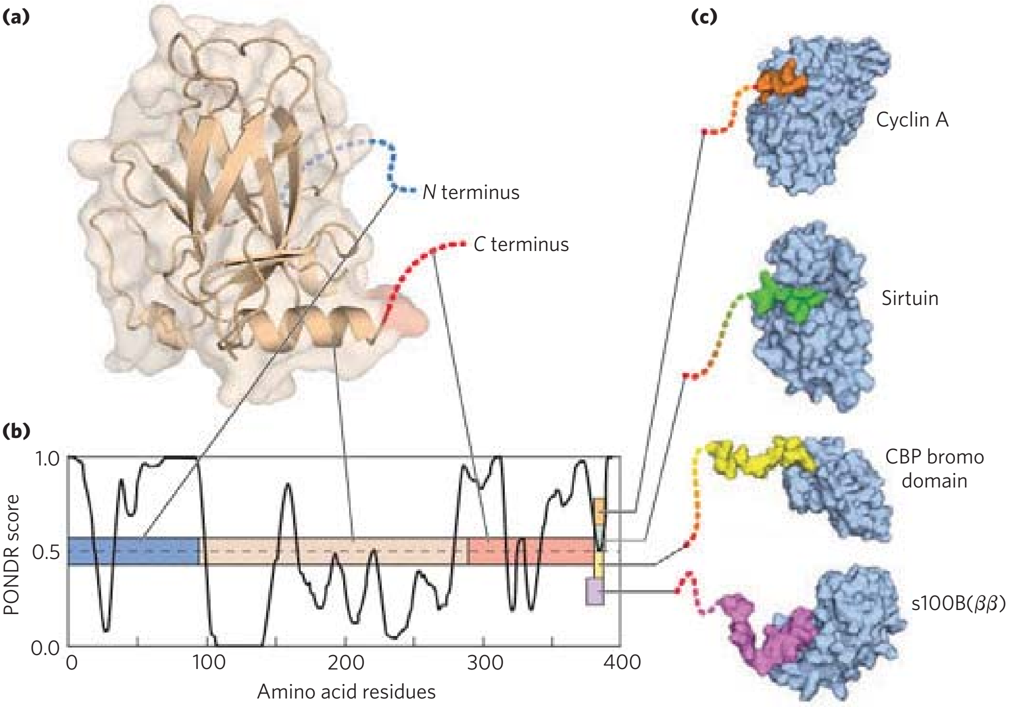


FIGURE 4–24 Binding of the intrinsically disordered carboxyl terminus of p53 protein to its binding partners. (a) The p53 protein is made up of several different segments (PDB ID 1XQH). Only the central domain is well ordered. **(b)** The linear sequence of the p53 protein is depicted as a colored bar. The overlaid graph presents a plot of the PONDR (Predictor of Natural Disordered Regions) score versus the protein sequence. PONDR is one of the best available algorithms for predicting the likelihood that a given amino acid residue is in a region of intrinsic disorder, based on the surrounding amino acid sequence and amino acid composition. A score of 1.0 indicates a probability of 100% that a protein will be disordered. In the actual protein structure, the tan central domain is ordered. The amino-terminal (blue) and carboxyl-terminal (red) regions are disordered. The very end of the carboxyl-terminal region has multiple binding partners and folds when it binds to each of them; however, the three-dimensional structure that is assumed when binding occurs is different for each of the interactions shown, and thus the color of this carboxyl-terminal segment (11 to 20 residues) is shown in a different color in each complex (cyclin A, PDB ID 1H26; sirtuin, PDB ID 1MA3; CBP bromo domain, PDB ID 1JSP; s100B(*ββ*), PDB ID 1DT7).

- 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, each unit called a protomer.
- Some proteins or protein segments are intrinsically disordered, lacking definable structure. These proteins have distinctive amino acid compositions that allow a more flexible structure. Some of these disordered proteins function as structural components or scavengers; others can interact with many different protein partners, serving as versatile inhibitors or as central components of protein interaction networks.

4.4 Protein Denaturation and Folding

Proteins lead a surprisingly precarious existence. As we have seen, a native protein conformation is only marginally stable. In addition, most proteins must maintain conformational flexibility to function. The continual maintenance of the active set of cellular proteins required under a given set of conditions is called **proteostasis**. Cellular proteostasis requires the coordinated function of pathways for protein synthesis and folding, the refolding of proteins that are partially unfolded, and the sequestration and degradation of proteins that have been irreversibly unfolded. In all cells, these networks involve hundreds of enzymes and specialized proteins.

As seen in **Figure 4–25**, the life of a protein encompasses much more than its synthesis and later degradation. The marginal stability of most proteins can produce a tenuous balance between folded and unfolded states. As proteins are synthesized on ribosomes (Chapter 27), they must fold into their native conformations. Sometimes this occurs spontaneously, but more often it occurs with the assistance of specialized enzymes and complexes called chaperones. Many of these same folding helpers function to refold proteins that become transiently unfolded. Proteins that are not properly folded often have exposed hydrophobic surfaces that render them "sticky," leading to the formation of inactive aggregates. These aggregates may lack their normal function but are not inert; their accumulation in cells lies at the heart of diseases ranging from diabetes to Parkinson and Alzheimer diseases. Not surprisingly, all cells have elaborate pathways for recycling and/or degrading proteins that are irreversibly misfolded.

The transitions between the folded and unfolded states, and the network of pathways that control these transitions, now become our focus.

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.

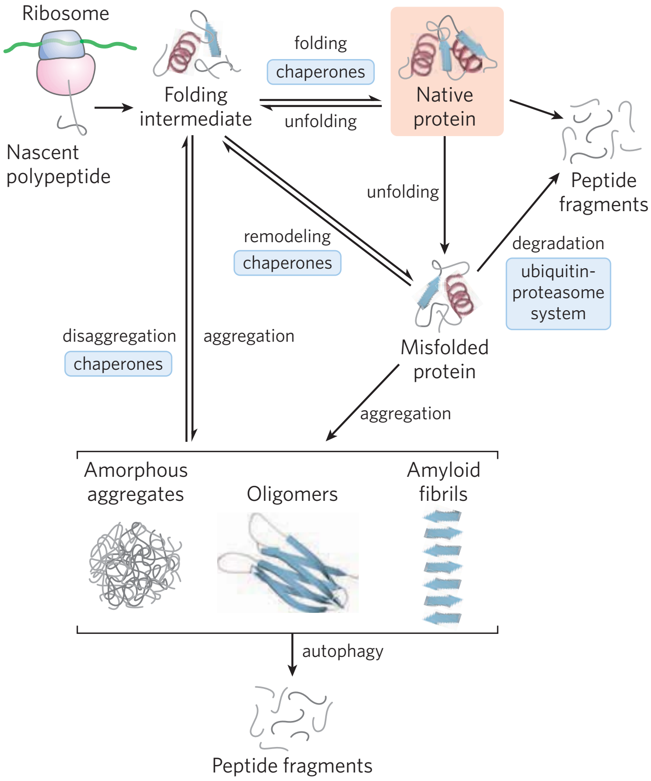


FIGURE 4–25 Pathways that contribute to proteostasis. Three kinds of processes contribute to proteostasis, in some cases with multiple contributing pathways. First, proteins are synthesized on a ribosome. Second, multiple pathways contribute to protein folding, many of which involve the activity of complexes called chaperones. Chaperones (including chaperonins) also contribute to the refolding of proteins that are partially and transiently unfolded. Finally, proteins that are irreversibly unfolded are subject to sequestration and degradation by several additional pathways. Partially unfolded proteins and protein-folding intermediates that escape the quality-control activities of the chaperones and degradative pathways may aggregate, forming both disordered aggregates and ordered amyloid-like aggregates that contribute to disease and aging processes.

Most proteins can be denatured by heat, which has complex effects on many weak interactions in a protein (primarily on the hydrogen bonds). 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 and archaea 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

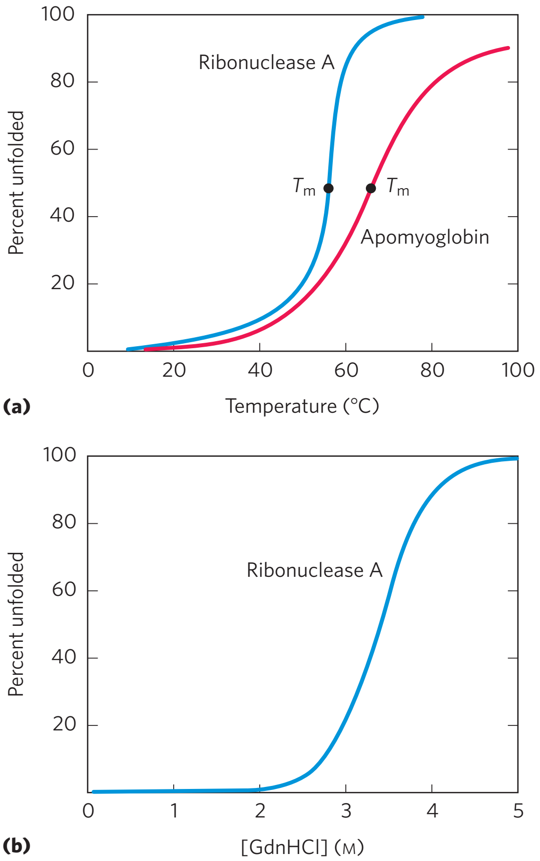


FIGURE 4–26 Protein denaturation. Results are shown for proteins denatured by two different environmental changes. In each case, the transition from the folded to the unfolded state is abrupt, suggesting cooperativity in the unfolding process. **(a)** Thermal denaturation of horse apomyoglobin (myoglobin 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. Denaturation of apomyoglobin was monitored by circular dichroism (see Fig. 4–10), which measures the amount of helical structure in the protein. 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.

homologous proteins derived from bacteria such as *Escherichia coli.* How these small differences promote structural stability at high temperatures is imperfectly understood.

Proteins can also be denatured by extremes of pH, by certain miscible organic solvents 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; urea also disrupts hydrogen bonds; extremes of pH alter the net charge on the protein, causing electrostatic repulsion and the disruption of some hydrogen bonding. The denatured structures resulting from these various treatments are not necessarily the same.

Denaturation often leads to protein precipitation, a consequence of protein aggregate formation as exposed hydrophobic surfaces associate. The aggregates are often highly disordered. The protein precipitate that is seen after boiling an egg white is one example. More-ordered aggregates are also observed in some proteins, as we shall see.

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 A, demonstrated by Christian Anfinsen in the 1950s. Purified ribonuclease A denatures completely in 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. Later, similar results were obtained using chemically synthesized, catalytically active ribonuclease A. 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.

The Anfinsen experiment 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. Subsequent work has shown that only a minority of proteins, many of them small and inherently stable, will fold spontaneously into their native form. Even though all proteins have the potential to fold into their native structure, many require some assistance.

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.

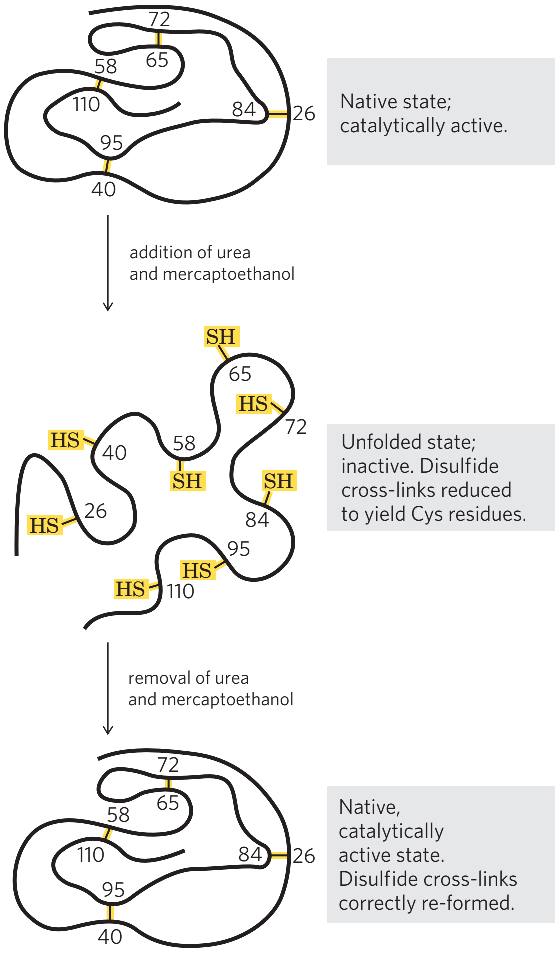

72
65
58
110
84
26
95
40
Native state; catalytically active.
addition of urea and mercaptoethanol
SH
65
HS
72
HS
40
58
SH
SH
HS
26
84
95
HS
HS
110
Unfolded state; inactive. Disulfide cross-links reduced to yield Cys residues.
removal of urea and mercaptoethanol
72
65
58
110
84
26
95
40
Native, catalytically active state. Disulfide cross-links correctly re-formed.

FIGURE 4–27 Renaturation of unfolded, denatured ribonuclease. Urea denatures the ribonuclease, and mercaptoethanol ($HOCH_2CH_2SH$) reduces and thus cleaves the disulfide bonds to yield eight Cys residues. Renaturation involves reestablishing the correct disulfide cross-links.

However, the synthesis of peptide bonds on the ribosome is not enough; the protein must fold.

How does the polypeptide chain arrive at its native conformation? Let's assume conservatively that each of the amino acid residues could take up 10 different conformations on average, giving 10^{100} different conformations for the polypeptide. Let's also assume that the protein folds 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 ($\sim10^{-13}$ second, or the time required for a single molecular vibration), it would take about 10^{77} years to sample all possible conformations. Clearly, protein folding is not 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. However, rapid progress has been made in this field, sufficient to produce robust algorithms that can often predict the structure of smaller proteins on the basis of their amino acid sequences. The major folding pathways are hierarchical. Local secondary structures form first. Certain amino acid sequences fold readily into α helices or β sheets, guided by constraints such as those reviewed in our discussion of secondary structure. Ionic interactions, involving charged groups that are often near one another in the linear sequence of the polypeptide chain, can play an important role in guiding these early folding steps. Assembly of local structures is followed by longer-range interactions between, say, two elements of secondary structure that come together to form stable folded structures. Hydrophobic interactions play a significant role throughout the process, as the aggregation of nonpolar amino acid side chains provides an entropic stabilization to intermediates and, eventually, to the final folded structure. The process continues until complete domains form and the entire polypeptide is folded **(Fig. 4–28)**. Notably, proteins dominated by close-range interactions (between pairs of residues generally located near each other in the polypeptide sequence) tend to fold faster than proteins with more complex folding patterns and many long-range interactions between different segments. As larger proteins with multiple domains are synthesized, domains near the amino terminus (which are synthesized

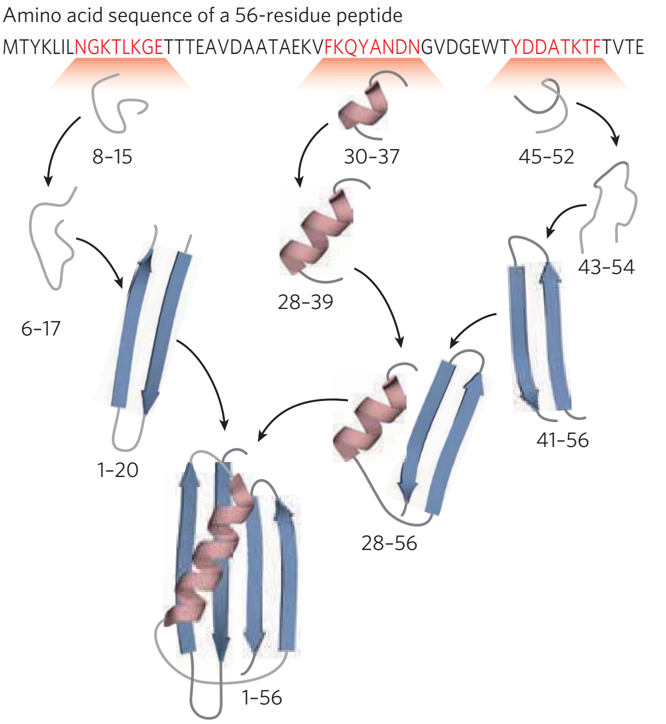

Amino acid sequence of a 56-residue peptide
MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
8–15
30–37
45–52
6–17
28–39
43–54
1–20
41–56
28–56
1–56

FIGURE 4–28 A protein-folding pathway as defined for a small protein. A hierarchical pathway is shown, based on computer modeling. Small regions of secondary structure are assembled first and then gradually incorporated into larger structures. The program used for this model has been highly successful in predicting the three-dimensional structure of small proteins from their amino acid sequence. The numbers indicate the amino acid residues in this 56 residue peptide that have acquired their final structure in each of the steps shown.

first) may fold before the entire polypeptide has been assembled.

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 reflects the decrease in the conformational space that must be searched as the protein approaches its native state. 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). The funnels can have a variety of shapes depending on the complexity of the folding pathway, the existence of semistable intermediates, and the potential for particular intermediates to assemble into aggregates of misfolded proteins (Fig. 4–29).

Thermodynamic stability is not evenly distributed over the structure of a protein—the molecule has regions of relatively high stability and others of low or negligible stability. For example, a protein may have two stable domains joined by a segment that is entirely disordered. Regions of low stability may 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 protein are often essential to protein function. Intrinsically disordered proteins or protein segments do not fold at all.

As our understanding of protein folding and protein structure improves, increasingly sophisticated computer programs for predicting the structure of proteins from their amino acid sequence are being developed. Prediction of protein structure is a specialty field of bioinformatics, and progress in this area is monitored with a biennial test called the CASP (Critical Assessment of Structural Prediction) competition. Entrants from around the world vie to predict the structure of an assigned protein (whose structure has been determined but not yet published). The most successful teams are invited to present their results at a CASP conference. The success of these efforts is improving rapidly.

Some Proteins Undergo Assisted Folding

Not all proteins fold spontaneously as they are synthesized in the cell. Folding for many proteins requires **chaperones**, proteins that interact with partially folded or improperly folded polypeptides, facilitating correct folding pathways or providing microenvironments in which folding can occur. Several types of molecular chaperones are found in organisms ranging from bacteria to humans. Two major families of chaperones, both well studied, are the **Hsp70** family and the **chaperonins**.

The Hsp70 family of proteins generally have a molecular weight near 70,000 and are more abundant in

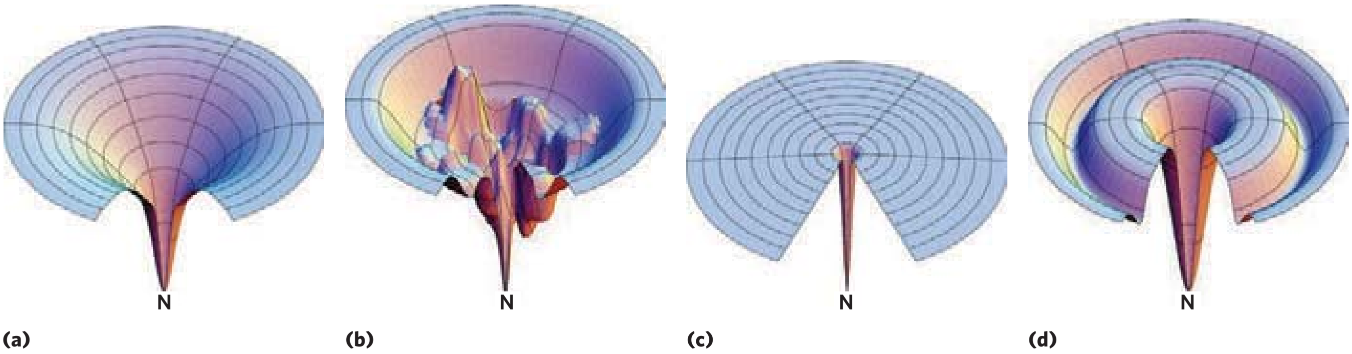


FIGURE 4–29 **The thermodynamics of protein folding depicted as free-energy funnels.** As proteins fold, the conformational space that can be explored by the structure is constrained. This is modeled as a three-dimensional thermodynamic funnel, with ΔG represented as depth and with the native structure (N) at the bottom (lowest free-energy point) of the funnel. The funnel for a given protein can have a variety of shapes, depending on the number and types of folding intermediates in the folding pathways. Any folding intermediate with significant stability and a finite lifetime would be represented as a local free-energy minimum—a depression on the surface of the funnel. **(a)** A simple but relatively wide and smooth funnel represents a protein that has multiple folding pathways (that is, the order in which different parts of the protein fold would be somewhat random), but it assumes its three-dimensional structure with no folding intermediates that have significant stability. **(b)** This funnel represents a more typical protein that has multiple possible folding intermediates with significant stability on the multiple pathways leading to the native structure. **(c)** A protein with one stable native structure, essentially no other folded intermediates with significant stability, and only one or a very few productive folding pathways is shown as a funnel with one narrow depression leading to the native form. **(d)** A protein with folding intermediates of substantial stability on virtually every pathway leading to the native state (that is, a protein in which a particular motif or domain always folds quickly, but other parts of the protein fold more slowly and in a random order) is depicted by a funnel with a major depression surrounding the depression leading to the native form.

cells stressed by elevated temperatures (hence, *h*eat *s*hock *p*roteins of M_r 70,000, or Hsp70). Hsp70 proteins bind to regions of unfolded polypeptides that are rich in hydrophobic residues. These chaperones thus "protect" both proteins subject to denaturation by heat and new peptide molecules being synthesized (and not yet folded). Hsp70 proteins also block the folding of certain proteins that must remain unfolded until they have been translocated across a membrane (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 uses energy from ATP hydrolysis and involves several other proteins (including a class called Hsp40). **Figure 4–30** illustrates chaperone-assisted folding as elucidated for the eukaryotic Hsp70 and Hsp40 chaperones. The binding of an unfolded polypeptide by an Hsp70 chaperone may break up a protein aggregate or prevent the formation of a new one. When the bound polypeptide is released, it has a chance to resume folding to its native structure. If folding does not occur rapidly enough, the polypeptide may be bound again and the process repeated. Alternatively, the Hsp70-bound polypeptide may be delivered to a chaperonin.

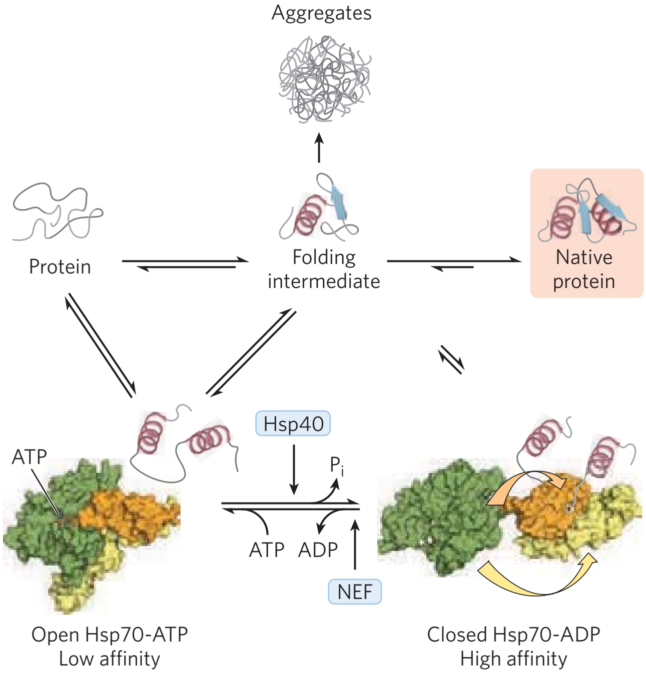


FIGURE 4–30 Chaperones in protein folding. The pathway by which Hsp70-class chaperones bind and release polypeptides is illustrated for the eukaryotic chaperones Hsp70 and Hsp40. The chaperones do not actively promote the folding of the substrate protein, but instead prevent aggregation of unfolded peptides. The unfolded or partly folded proteins bind first to the open, ATP-bound form of Hsp70 (PDB ID 2QXL). Hsp40 then interacts with this complex and triggers ATP hydrolysis that produces the closed form of the complex (derived from PDB IDs 2KHO and 1DKZ), where the domains colored orange and yellow come together like the two parts of a jaw, trapping parts of the unfolded protein inside. Dissociation of ADP and recycling of the Hsp70 requires interaction with another protein, nucleotide-exchange factor (NEF). For a population of polypeptide molecules, some fraction of the molecules released after the transient binding of partially folded proteins by Hsp70 will take up the native conformation. The remainder are rebound by Hsp70 or diverted to the chaperonin system (Hsp60; see Fig. 4–31). In bacteria, the Hsp70 and Hsp40 chaperones are called DnaK and DnaJ, respectively. DnaK and DnaJ were first identified as proteins required for in vitro replication of certain viral DNA molecules (hence the "Dna" designation).

Chaperonins are elaborate protein complexes required for the folding of some 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). The analogous chaperonin system in eukaryotes is called Hsp60. The chaperonins first became known when they were found to be necessary for the growth of certain bacterial viruses (hence the designation "Gro"). This family of proteins is structured as a series of multisubunit rings, forming two chambers oriented back to back. An unfolded protein is first bound to an exposed hydrophobic surface near the apical end of one GroEL chamber. The protein is then trapped within the chamber when it is capped transiently by the GroES "lid" **(Fig. 4–31)**. GroEL undergoes substantial conformational changes, coupled to slow ATP hydrolysis, which also regulates the binding and release of GroES. Inside the chamber, a protein has about 10 seconds to fold—the time required for the bound ATP to hydrolyze. Constraining a protein within the chamber prevents inappropriate protein aggregation and also restricts the conformational space that a polypeptide chain can explore as it folds. The protein is released when the GroES cap dissociates but can rebound rapidly for another round if folding has not been completed. The two chambers in a GroEL complex alternate in binding and releasing unfolded polypeptide substrates. In eukaryotes, the Hsp60 system utilizes a similar process to fold proteins. However, in place of the GroES lid, protrusions from the apical domains of the subunits flex and close over the chamber. The ATP hydrolytic cycle is also slower in the Hsp60 complexes, giving the proteins constrained inside more time to fold.

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

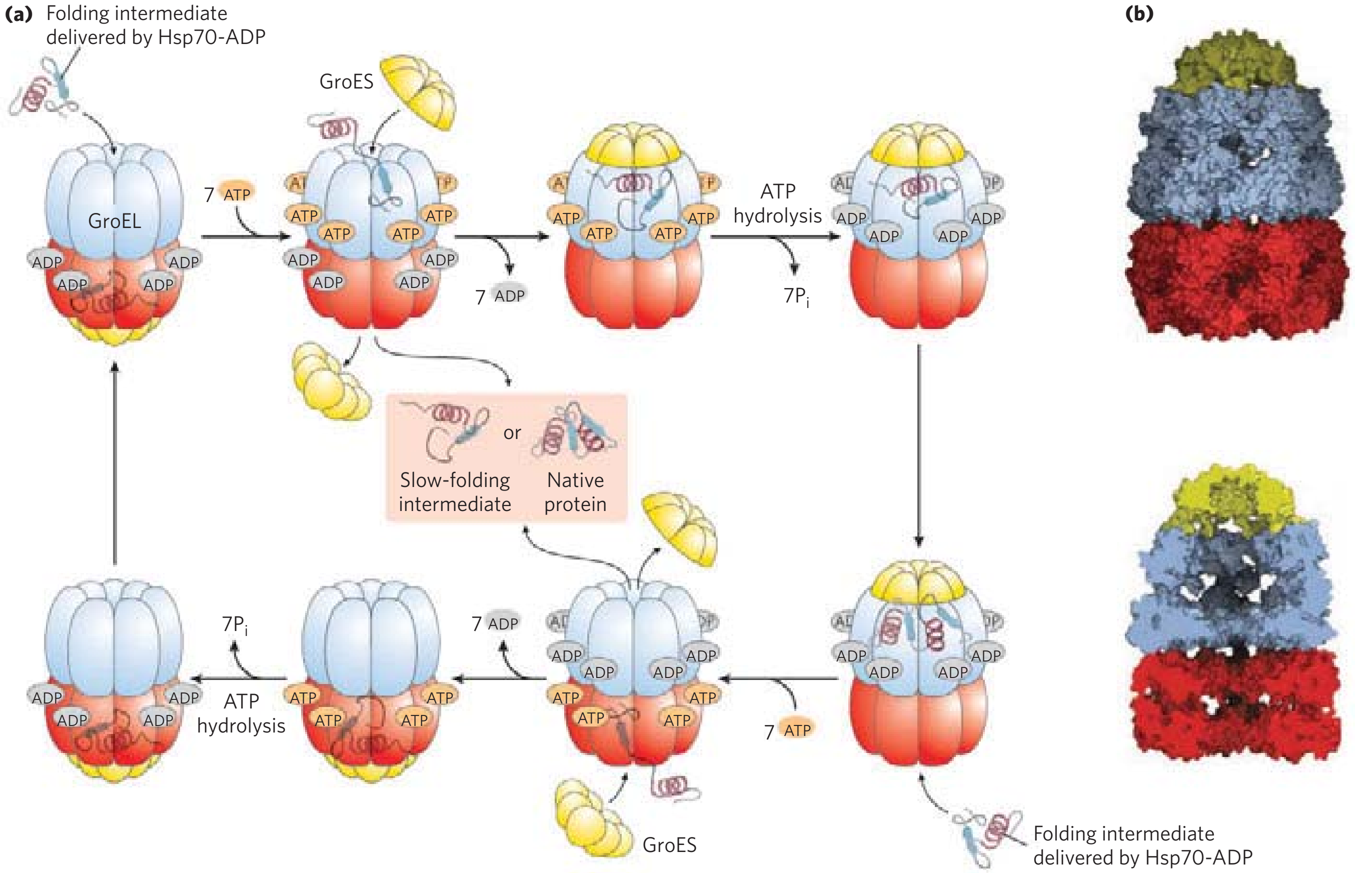

(a) Folding intermediate delivered by Hsp70-ADP
GroES
GroEL
7 ATP
ATP
ADP
7 ADP
ATP hydrolysis
7P$_i$
or
Slow-folding intermediate
Native protein
7 ADP
7P$_i$
ATP hydrolysis
7 ATP
GroES
Folding intermediate delivered by Hsp70-ADP
(b)

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 chambers formed by two heptameric rings (each subunit M_r 57,000). GroES, also a heptamer (subunit M_r 10,000), blocks one of the GroEL chambers after an unfolded protein is bound inside. The chamber with the unfolded protein is referred to as cis; the opposite one is trans. Folding occurs within the cis chamber, during the time it takes to hydrolyze the 7 ATP bound to the subunits in the heptameric ring. The GroES and the ADP molecules then dissociate, and the protein is released. The two chambers of the GroEL/Hsp60 systems alternate in the binding and facilitated folding of client proteins. **(b)** Surface and cutaway images of the GroEL/GroES complex (PDB ID 1AON). The cutaway (below) illustrates the large interior space within which other proteins are bound.

Defects in Protein Folding Provide the Molecular Basis for a Wide Range of Human Genetic Disorders

Despite the many processes that assist in protein folding, misfolding does occur. In fact, protein misfolding is a substantial problem in all cells, and a quarter or more of all polypeptides synthesized may be destroyed because they do not fold correctly. In some cases, the misfolding causes or contributes to the development of serious disease.

Many conditions, including type 2 diabetes, Alzheimer disease, Huntington disease, and Parkinson disease, are associated with a misfolding mechanism: a soluble protein that is normally secreted from the cell is secreted in a misfolded state and converted into an insoluble extracellular **amyloid** fiber. The diseases are collectively referred to as **amyloidoses**. The fibers are highly ordered and unbranched, with a diameter of 7 to 10 nm and a high degree of β-sheet structure. The β segments are oriented perpendicular to the axis of the fiber. In some amyloid fibers the overall structure features two layers of β sheet, such as that shown for amyloid-β peptide in **Figure 4–32**.

Many proteins can take on the amyloid fibril structure as an alternative to their normal folded conformations, and most of these proteins have a concentration of aromatic amino acid residues in a core region of β sheet or α helix. The proteins are secreted in an incompletely folded conformation. The core (or some part of it) folds into a β sheet before the rest of the protein folds correctly, and the β sheets from two or more incompletely folded protein molecules associate to begin forming an amyloid fibril. The fibril grows in the extracellular space. Other parts of the protein then fold differently, remaining on the outside of the β-sheet core in the growing fibril. The effect of aromatic residues in stabilizing the structure is shown in Figure 4–32c. Because most of the protein molecules fold normally, the onset of symptoms in the amyloidoses is often very slow. If a person inherits a mutation such as substitution with an aromatic residue at a position that favors formation of amyloid fibrils, disease symptoms may begin at an earlier age.

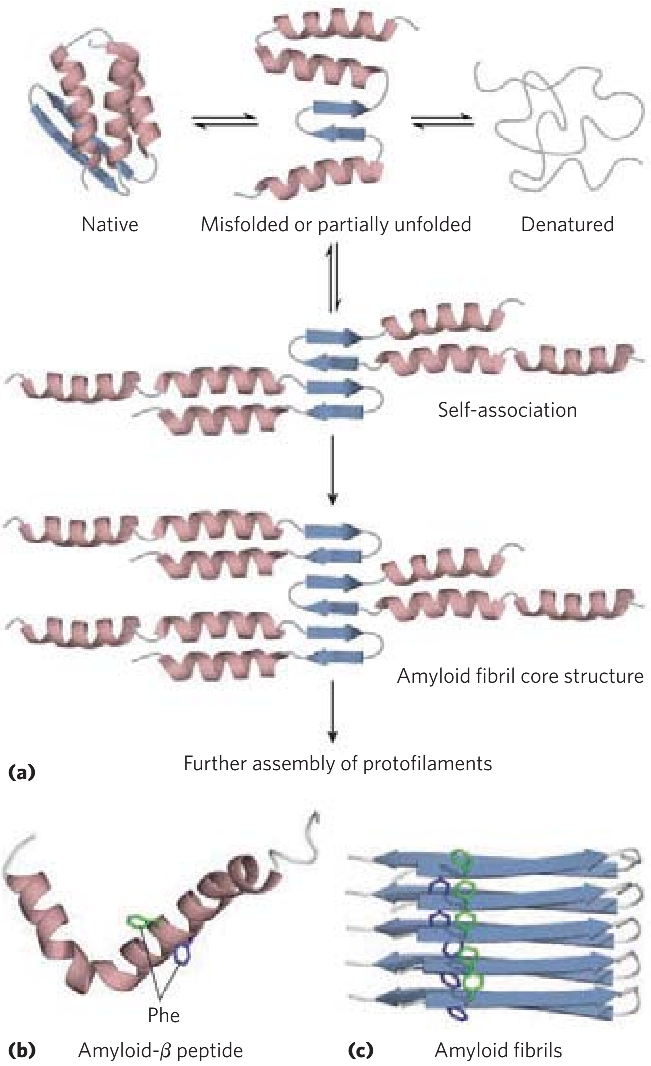

Native
Misfolded or partially unfolded
Denatured
Self-association
Amyloid fibril core structure
(a)
Further assembly of protofilaments
Phe
(b) Amyloid-β peptide
(c) Amyloid fibrils

FIGURE 4–32 Formation of disease-causing amyloid fibrils. (a) Protein molecules whose normal structure includes regions of β sheet undergo partial folding. In a small number of the molecules, before folding is complete, the β-sheet regions of one polypeptide associate with the same region in another polypeptide, forming the nucleus of an amyloid. Additional protein molecules slowly associate with the amyloid and extend it to form a fibril. **(b)** The amyloid-β peptide begins as two α-helical segments of a larger protein. Proteolytic cleavage of this larger protein leaves the relatively unstable amyloid-β peptide, which loses its α-helical structure. It can then assemble slowly into amyloid fibrils **(c)**, which contribute to the characteristic plaques on the exterior of nervous tissue in people with Alzheimer disease. The aromatic side chains shown here play a significant role in stabilizing the amyloid structure. Amyloid is rich in β sheet, with the β strands arranged perpendicular to the axis of the amyloid fibril. Amyloid-β peptide takes the form of two layers of extended parallel β sheet. Some amyloid-forming peptides may fold to form left-handed β-helices (see Fig. 4–22).

In eukaryotes, proteins destined for secretion undergo their initial folding in the endoplasmic reticulum (ER; see pathway in Chapter 27). When stress conditions arise, or when protein synthesis threatens to overwhelm the protein-folding capacity of the ER, unfolded proteins can accumulate. These conditions trigger the unfolded protein response (UPR). A set of transcriptional regulators that constitute the UPR bring the various systems into alignment by increasing the concentration of chaperones in the ER or decreasing the rate of overall protein synthesis, or both. Amyloid aggregates that form before the UPR can come into play may be removed. Some are degraded by **autophagy**. In this process, they are first encapsulated in a membrane, then the contents of the resulting vesicle are degraded after the vesicle docks with a cytosolic lysosome. Alternatively, misfolded proteins can be degraded by a system of proteases called the ubiquitin-proteasome system (described in Chapter 27). Defects in any of these systems decrease the capacity to deal with misfolded proteins and increase the propensity for development of amyloid-related diseases.

Some amyloidoses are systemic, involving many tissues. Primary systemic amyloidosis is caused by deposition of fibrils consisting of misfolded immunoglobulin light chains (see Chapter 5), or fragments of light chains derived from proteolytic degradation. The mean age of onset is about 65 years. Patients have symptoms including fatigue, hoarseness, swelling, and weight loss, and many die within the first year after diagnosis. The kidneys or heart are often most affected. Some amyloidoses are associated with other types of disease. People with certain chronic infectious or inflammatory diseases such as rheumatoid arthritis, tuberculosis, cystic fibrosis, and some cancers can experience a sharp increase in secretion of an amyloid-prone polypeptide called serum amyloid A (SAA) protein. This protein, or fragments of it, deposits in the connective tissue of the spleen, kidney, and liver, and around the heart. People with this condition, known as secondary systemic amyloidosis, have a wide range of symptoms, depending on the organs initially affected. The disease is generally fatal within a few years. More than 80 amyloidoses are associated with mutations in transthyretin (a protein that binds to and transports thyroid hormones, distributing them throughout the body and brain). A variety of mutations in this protein lead to amyloid deposition concentrated around different tissues, thus producing different symptoms. Amyloidoses are also associated with inherited mutations in the proteins lysozyme, fibrinogen A α chain, and apolipoproteins A-I and A-II; all of these proteins are described in later chapters.

Some amyloid diseases are associated with particular organs. The amyloid-prone protein is generally secreted only by the affected tissue, and its locally high concentration leads to amyloid deposition around that tissue (although some of the protein may be distributed systemically). One common site of amyloid deposition is near the pancreatic islet β cells, responsible for insulin secretion and regulation of glucose metabolism (see Fig. 23–26). Secretion by β cells of a small (37 amino acid) peptide called islet amyloid polypeptide (IAPP), or amylin, can lead to amyloid deposits around the islets, gradually destroying the cells. A healthy human adult has 1 to 1.5 million

BOX 4–6 MEDICINE Death by Misfolding: The Prion Diseases

A misfolded brain protein seems to be the causative agent of several rare degenerative brain diseases in mammals. Perhaps the best known of these is bovine spongiform encephalopathy (BSE; also known as mad cow disease). 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). Progressive deterioration of the brain leads to a spectrum of neurological symptoms, including weight loss, erratic behavior, problems with posture, balance, and coordination, and loss of cognitive function. The diseases are fatal.

In the 1960s, investigators found that preparations of the disease-causing agents seemed 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.

The infectious agent has been traced to a single protein (M_r 28,000), which Prusiner dubbed **prion** protein (PrP). The name was derived from *prot*einaceous *in*fectious, but Prusiner thought that "prion" sounded better than "proin." Prion protein is a normal constituent of brain tissue in all mammals. Its role is not known in detail, but it may 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 $\mathrm{PrP}^{\mathrm{C}}$, occurs in an altered conformation called $\mathrm{PrP}^{\mathrm{Sc}}$ (Sc denotes scrapie). The structure

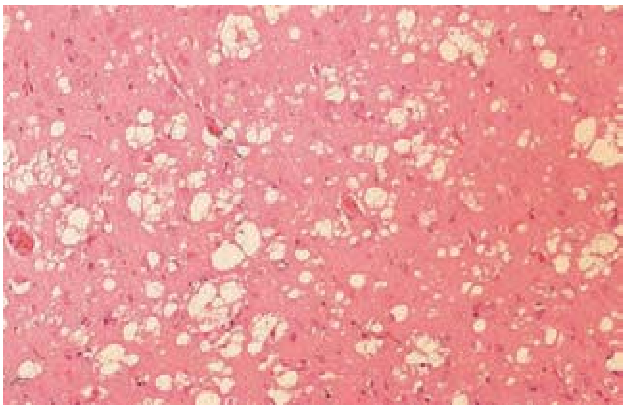


FIGURE 1 Stained section of cerebral cortex from autopsy of a patient with Creutzfeldt-Jakob disease shows spongiform (vacuolar) degeneration, the most characteristic neurohistological feature. The yellowish vacuoles are intracellular and occur mostly in pre- and postsynaptic processes of neurons. The vacuoles in this section vary in diameter from 20 to 100 μm.

pancreatic β cells. With progressive loss of these cells, glucose homeostasis is affected and eventually, when 50% or more of the cells are lost, the condition matures into type 2 (non–insulin-dependent) diabetes mellitus.

The amyloid deposition diseases that trigger neurodegeneration, particularly in older adults, are a special class of localized amyloidoses. Alzheimer disease is associated with extracellular amyloid deposition by neurons, involving the amyloid-β peptide (Fig. 4–32b), derived from a larger transmembrane protein (amyloid-β precursor protein) found in most human tissues. When it is part of the larger protein, the peptide is composed of two α-helical segments spanning the membrane. When the external and internal domains are cleaved off by dedicated proteases, the relatively unstable amyloid-β peptide leaves the membrane and loses its α-helical structure. It can then take the form of two layers of extended parallel β sheet, which can slowly assemble into amyloid fibrils (Fig. 4–32c). Deposits of these amyloid fibers seem to be the primary cause of Alzheimer disease, but a second type of amyloidlike aggregation, involving a protein called tau, also occurs intracellularly (in neurons) in people with Alzheimer disease. Inherited mutations in the tau protein do not result in Alzheimer disease, but they cause a frontotemporal dementia and parkinsonism (a condition with symptoms resembling Parkinson disease) that can be equally devastating.

Several other neurodegenerative conditions involve intracellular aggregation of misfolded proteins. In Parkinson disease, the misfolded form of the protein α-synuclein aggregates into spherical filamentous masses called Lewy bodies. Huntington disease involves the protein huntingtin, which has a long polyglutamine repeat. In some individuals, the polyglutamine repeat is longer than normal and a more subtle type of intracellular aggregation occurs. Notably, when the mutant human proteins involved in Parkinson and Huntington diseases are expressed in *Drosophila melanogaster*, the flies display neurodegeneration expressed as eye deterioration, tremors, and early death. All of these symptoms are highly suppressed if expression of the Hsp70 chaperone is also increased.

of PrP^C features two α helices. The structure of PrP^{Sc} is very different, with much of the structure converted to amyloid-like β sheets (Fig. 2). The interaction of PrP^{Sc} with PrP^C converts the latter to PrP^{Sc}, 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^{Sc} 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^C to PrP^{Sc} more likely. A complete understanding of prion diseases awaits new information on 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).

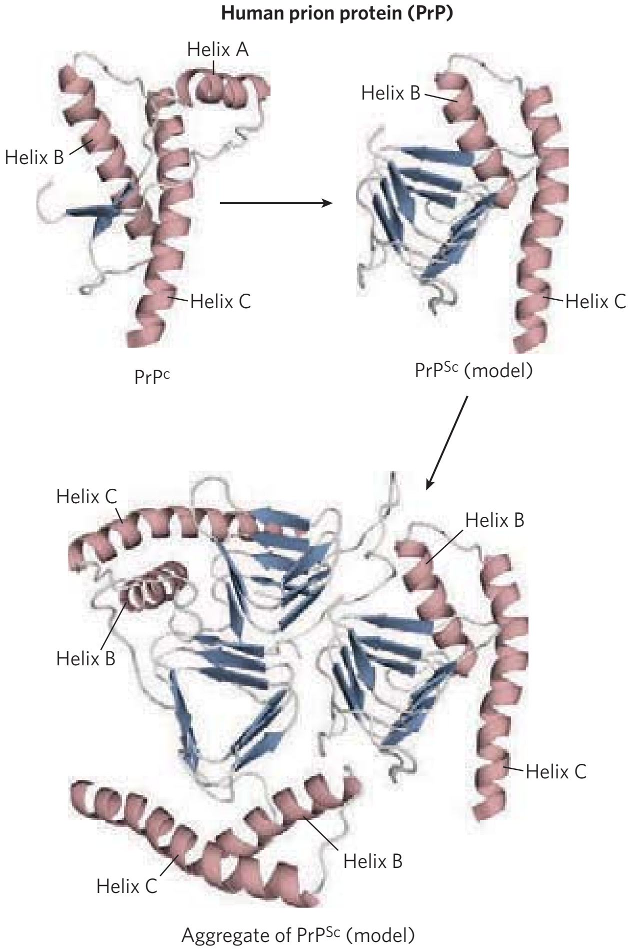


FIGURE 2 Structure of the globular domain of human PrP (PDB ID 1QLX) and models of the misfolded, disease-causing conformation PrP^{Sc}, and an aggregate of PrP^{Sc}. The α helices are labeled to help illustrate the conformation change. Helix A is incorporated into the β-sheet structure of the misfolded conformation.

Protein misfolding need not lead to amyloid formation to cause serious disease. For example, cystic fibrosis is caused by defects in a membrane-bound protein called *c*ystic *f*ibrosis *t*ransmembrane conductance *r*egulator (CFTR), which acts as a channel for chloride ions. The most common cystic fibrosis–causing mutation is the deletion of a Phe residue at position 508 in CFTR, which causes improper protein folding. Most of this protein is then degraded and its normal function is lost (see Box 11–2). Many of the disease-related mutations in collagen (p. 130) also cause defective folding. A particularly remarkable type of protein misfolding is seen in the prion diseases (Box 4–6). ■

SUMMARY 4.4 Protein Denaturation and Folding

- The maintenance of the steady-state collection of active cellular proteins required under a particular set of conditions—called proteostasis—involves an elaborate set of pathways and processes that fold, refold, and degrade polypeptide chains.
- The three-dimensional structure and the function of most 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 tertiary structure is determined by amino acid sequence.
- Protein folding in cells is generally hierarchical. Initially, regions of secondary structure may form, followed by folding into motifs and domains. 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.
- Protein misfolding is the molecular basis of a wide range of human diseases, including the amyloidoses.

Key Terms

Terms in bold are defined in the glossary.

conformation	115
native conformation	116
hydrophobic interactions	116
solvation layer	116
peptide group	118
Ramachandran plot	119
secondary structure	119
α helix	120
β conformation	123
β sheet	123
β turn	123
circular dichroism (CD) spectroscopy	124
tertiary structure	125
quaternary structure	125
fibrous proteins	125
globular proteins	125
α-keratin	126
collagen	127
silk fibroin	130
Protein Data Bank (PDB)	132
motif	137
fold	137
domain	137
protein family	140
multimer	140
oligomer	140
protomer	141
intrinsically disordered proteins	141
proteostasis	143
denaturation	143
renaturation	144
chaperone	146
Hsp70	146
chaperonin	146
protein disulfide isomerase (PDI)	147
peptide prolyl cis-trans isomerase (PPI)	147
amyloid	148
amyloidoses	148
autophagy	149
prion	150

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.

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

A comprehensive and authoritative source.

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. Protein 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

Beeby, M., O'Connor, B.D., Ryttersgaard, C., Boutz, D.R., Perry, L.J., & Yeates, T.O. (2005) The genomics of disulfide bonding and protein stabilization in thermophiles. *PLoS Biol.* **3**, e309.

Brown, J.H. (2006) Breaking symmetry in protein dimers: designs and function. *Protein Sci.* **15**, 1–13.

Dunker, A.K. & Kriwacki, R.W. (2011) The orderly chaos of proteins. *Sci. Am.* **304** (April), 68–73.

A nice summary of the work on proteins that lack intrinsic structure.

Herráez, A. (2006) Biomolecules in the computer. *Biochem. Mol. Biol. Educ.* **34**, 255–261.

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

A description of how macromolecules such as proteins are crystallized.

Milner-White, E.J. (1997) The partial charge of the nitrogen atom in peptide bonds. *Protein Sci.* **6**, 2477–2482.

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.

Protein Denaturation and Folding

Chiti, F. & Dobson, C.M. (2006) Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* **75**, 333–366.

Dill, K.A., Ozkan, S.B., Shell, M.S., & Weikl, T.R. (2008) The protein folding problem. *Annu. Rev. Biophys.* **37**, 289–316.

Gazit, E. (2005) Mechanisms of amyloid fibril self-assembly and inhibition. *FEBS J.* **272**, 5971–5978.

Hartl, F.U., Bracher, A., & Hayer-Hartl, M. (2011) Molecular chaperones in protein folding and proteostasis. *Nature* **475**, 324–332.

Hoppener, J.W.M. & Lips, C.J.M. (2006) Role of islet amyloid in type 2 diabetes mellitus. *Int. J. Biochem. Cell Biol.* **38**, 726–736.

Kapinga, H.H. & Craig, E.A. (2010) The HSP70 chaperone machinery: J proteins as drivers of functional specificity. *Nat. Rev. Mol. Cell. Biol.* **11**, 579–592.

Norrby, E. (2011) Prions and protein-folding diseases. *J. Intern. Med.* **270**, 1–14.

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

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

Selkoe, D.J. (2003) Folding proteins in fatal ways. *Nature* **426**, 900–904.

A good summary of amyloidoses.

Tang, Y., Chang, H., Roeben, A., Wischnewski, D., Wischnewski, N., Kerner, M., Hartl, F., & Hayer-Hartl, M. (2006) Structural features of the GroEL-GroES nanocage required for rapid folding of encapsulated protein. *Cell* **125**, 903–914.

Tyedmers, J., Mogk, A., & Bukau, B. (2010) Cellular strategies for controlling protein aggregation. *Nat. Rev. Mol. Cell Biol.* **11**, 777–788.

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 α-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?

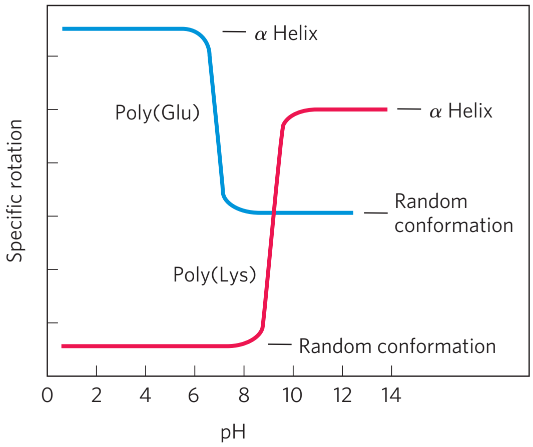
2. Structural and Functional Relationships in Fibrous Proteins William Astbury discovered that the x-ray diffraction pattern of wool shows a repeating structural unit spaced about 5.2 Å along the length of the wool fiber. When he steamed and stretched the wool, the x-ray pattern showed a new repeating 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 *α*-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 *α*-keratin filaments are synthesized inside living epidermal cells and assembled into ropelike structures (see Fig. 4–11). The fundamental structural element of *α*-keratin is the *α* helix, which has 3.6 amino acid residues per turn and a rise of 5.4 Å per turn (see Fig. 4–4a). Assuming that the biosynthesis of *α*-helical keratin chains is the rate-limiting factor in the growth of hair, calculate the rate at which peptide bonds of *α*-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 *α*-Helical Secondary Structures The unfolding of the *α* helix of a polypeptide to a randomly coiled conformation is accompanied by a large decrease in a property called specific rotation, a measure of a solution's capacity to rotate circularly polarized light. Polyglutamate, a polypeptide made up of only L-Glu residues, has the *α*-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 *α* helix at pH 10, but when the pH is lowered to 7 the specific rotation also decreases, as shown by the following graph.



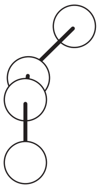
What is the explanation for the effect of the pH changes on the conformations of poly(Glu) and poly(Lys)? Why does the transition occur over such a narrow range of pH?

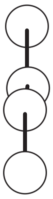
5. Disulfide Bonds Determine the Properties of Many Proteins Some natural proteins are rich in disulfide bonds, and their mechanical properties (tensile strength, viscosity, hardness, etc.) are correlated with the degree of disulfide bonding.

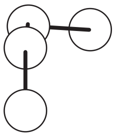
(a) Glutenin, a wheat protein rich in disulfide bonds, is responsible for the cohesive and elastic character of dough made from wheat flour. Similarly, the hard, tough nature of tortoise shell is due to the extensive disulfide bonding in its *α*-keratin. What is the molecular basis for the correlation between disulfide-bond content and mechanical properties of the protein?

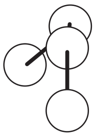
(b) Most globular proteins are denatured and lose their activity when briefly heated to 65 °C. However, globular proteins that contain multiple disulfide bonds often must be heated longer at higher temperatures to denature them. One such protein is bovine pancreatic trypsin inhibitor (BPTI), which has 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?

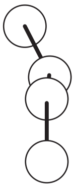
6. Dihedral Angles A series of torsion angles, ϕ and ψ, that might be taken up by the peptide backbone is shown below. Which of these closely correspond to ϕ and ψ for an idealized collagen triple helix? Refer to Figure 4–9 as a guide.

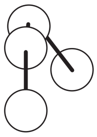

(a)


(b)


(c)


(d)


(e)


(f)

7. 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. Consider the following amino acid sequence.

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) Where might bends or *β* turns 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.)

8. Bacteriorhodopsin in Purple Membrane Proteins Under the proper environmental conditions, the salt-loving archaeon *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 α-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 α 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.)

9. Protein Structure Terminology Is myoglobin a motif, a domain, or a complete three-dimensional structure?

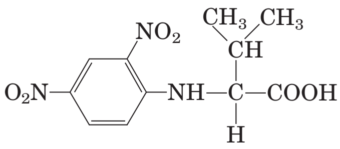
10. Interpreting Ramachandran Plots Examine the two proteins labeled (a) and (b) below. Which of the two Ramachandran plots, labeled (c) and (d), is more likely to be derived from which protein? Why?

11. 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:

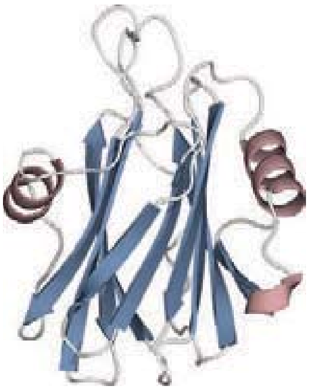
$$\text{—X—Gly—Pro—Y—} \xrightarrow{H_2O} \text{—X—COO}^- + H_3\overset{+}{N}\text{—Gly—Pro—Y—}$$

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?

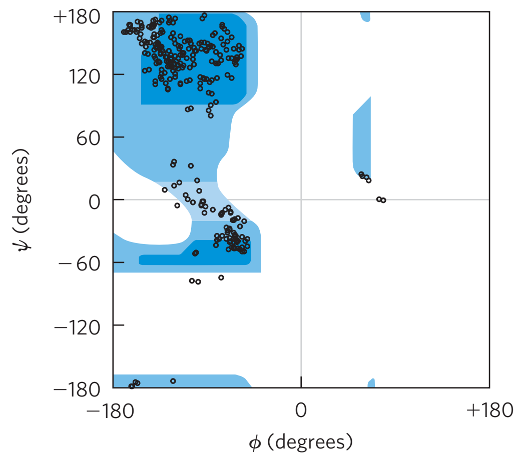
12. Number of Polypeptide Chains in a Multisubunit Protein A sample (660 mg) of an oligomeric protein of M_r 132,000 was treated with an excess of 1-fluoro-2,4-dinitrobenzene (Sanger's reagent) under slightly alkaline conditions until the chemical reaction was complete. The peptide bonds of the protein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



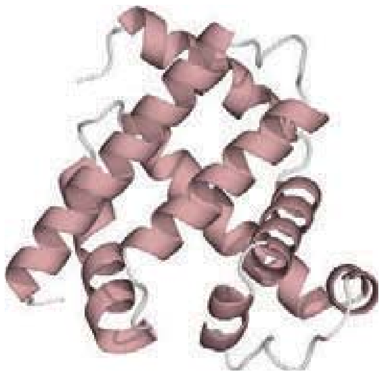
(a)



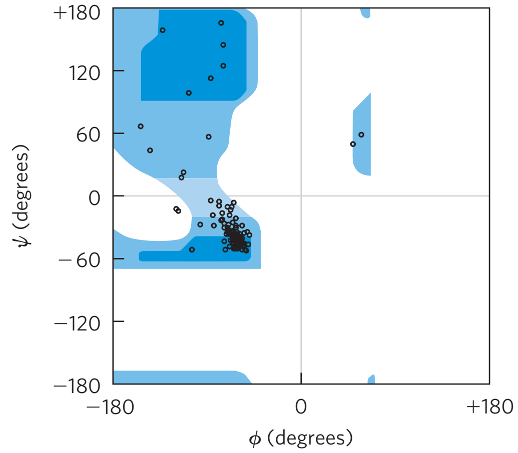
(c)



(b)



(d)



2,4-Dinitrophenyl derivatives of the α-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?

13. Predicting Secondary Structure Which of the following peptides is more likely to take up an α-helical structure, and why?

(a) LKAENDEAARAMSEA

(b) CRAGGFPWDQPGTSN

14. Amyloid Fibers in Disease Several small aromatic molecules, such as phenol red (used as a nontoxic drug model), have been shown to inhibit the formation of amyloid in laboratory model systems. A goal of the research on these small aromatic compounds is to find a drug that would efficiently inhibit the formation of amyloid in the brain in people with incipient Alzheimer disease.

(a) Suggest why molecules with aromatic substituents would disrupt the formation of amyloid.

(b) Some researchers have suggested that a drug used to treat Alzheimer disease may also be effective in treating type 2 (non–insulin-dependent) diabetes mellitus. Why might a single drug be effective in treating these two different conditions?

Using the Web

15. Protein Modeling on the Internet A group of patients with Crohn disease (an inflammatory bowel disease) underwent biopsies of their intestinal mucosa in an attempt to identify the causative agent. Researchers identified a protein that was present at higher levels in patients with Crohn 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
TAELKICRVN  RNSGSCLGGD  EIFLLCDKVQ
KEDIEVYFTG  PGWEARGSFS  QADVHRQVAI
VFRTPPYADP  SLQAPVRVSM  QLRRPSDREL
SEPMEFQYLP  DTDDRHRIEE  KRKRTYETFK
SIMKKSPFSG  PTDPRPPPRR  IAVPSRSSAS
VPKPAPQPYP
```

(a) You can identify this protein using a protein database on the Internet. Some good places to start include Protein Information Resource (PIR; http://pir.georgetown.edu), Structural Classification of Proteins (SCOP; http://scop.mrc-lmb.cam.ac.uk/scop), and Prosite (http://prosite.expasy.org).

At your selected database site, follow links to the sequence comparison engine. Enter about 30 residues from the protein sequence 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 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.pdb.org) or SCOP.

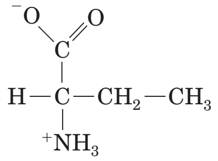
(d) In the course of your Web searches, what did you learn about the cellular function of the protein?

Data Analysis Problem

16. Mirror-Image Proteins As noted in Chapter 3, "The amino acid residues in protein molecules are exclusively L stereoisomers." It is not clear whether this selectivity is necessary for proper protein function or is an accident of evolution. To explore this question, Milton and colleagues (1992) published a study of an enzyme made entirely of D stereoisomers. The enzyme they chose was HIV protease, a proteolytic enzyme made by HIV that converts inactive viral preproteins to their active forms.

Previously, Wlodawer and coworkers (1989) had reported the complete chemical synthesis of HIV protease from L-amino acids (the L-enzyme), using the process shown in Figure 3–32. Normal HIV protease contains two Cys residues at positions 67 and 95. Because chemical synthesis of proteins containing Cys is technically difficult, Wlodawer and colleagues substituted the synthetic amino acid L-α-amino-n-butyric acid (Aba) for the two Cys residues in the protein. In the authors' words, this was done to "reduce synthetic difficulties associated with Cys deprotection and ease product handling."

(a) The structure of Aba is shown below. Why was this a suitable substitution for a Cys residue? Under what circumstances would it not be suitable?



L-α-Amino-n-butyric acid

Wlodawer and coworkers denatured the newly synthesized protein by dissolving it in 6 M guanidine HCl and then allowed it to fold slowly by dialyzing away the guanidine against a neutral buffer (10% glycerol, 25 mM $NaPO_4$, pH 7).

(b) There are many reasons to predict that a protein synthesized, denatured, and folded in this manner would not be active. Give three such reasons.

(c) Interestingly, the resulting L-protease was active. What does this finding tell you about the role of disulfide bonds in the native HIV protease molecule?

In their new study, Milton and coworkers synthesized HIV protease from D-amino acids, using the same protocol as the earlier study (Wlodawer et al.). Formally, there are three possibilities for the folding of the D-protease: it would give (1) the same shape as the L-protease, (2) the mirror image of the L-protease, or (3) something else, possibly inactive.

(d) For each possibility, decide whether or not it is a likely outcome and defend your position.

In fact, the D-protease was active: it cleaved a particular synthetic substrate and was inhibited by specific inhibitors. To examine the structure of the D- and L-enzymes, Milton and coworkers tested both forms for activity with D and L forms of a chiral peptide substrate and for inhibition by D and L forms of a chiral peptide-analog inhibitor. Both forms were also tested for inhibition by the achiral inhibitor Evans blue. The findings are given in the table.

			Inhibition		
	Substrate hydrolysis		Peptide inhibitor		Evans blue
HIV Protease	D-substrate	L-substrate	D-inhibitor	L-inhibitor	(achiral)
L-protease	−	+	−	+	+
D-protease	+	−	+	−	+

(e) Which of the three models proposed above is supported by these data? Explain your reasoning.

(f) Why does Evans blue inhibit both forms of the protease?

(g) Would you expect chymotrypsin to digest the D-protease? Explain your reasoning.

(h) Would you expect total synthesis from D-amino acids followed by renaturation to yield active enzyme for any enzyme? Explain your reasoning.

References

Milton, R.C., Milton, S.C., & Kent, S.B. (1992) Total chemical synthesis of a D-enzyme: the enantiomers of HIV-1 protease show demonstration of reciprocal chiral substrate specificity. *Science* **256**, 1445–1448.

Wlodawer, A., Miller, M., Jaskólski, M., Sathyanarayana, B.K., Baldwin, E., Weber, I.T., Selk, L.M., Clawson, L., Schneider, J., & Kent, S.B. (1989) Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science* **245**, 616–621.