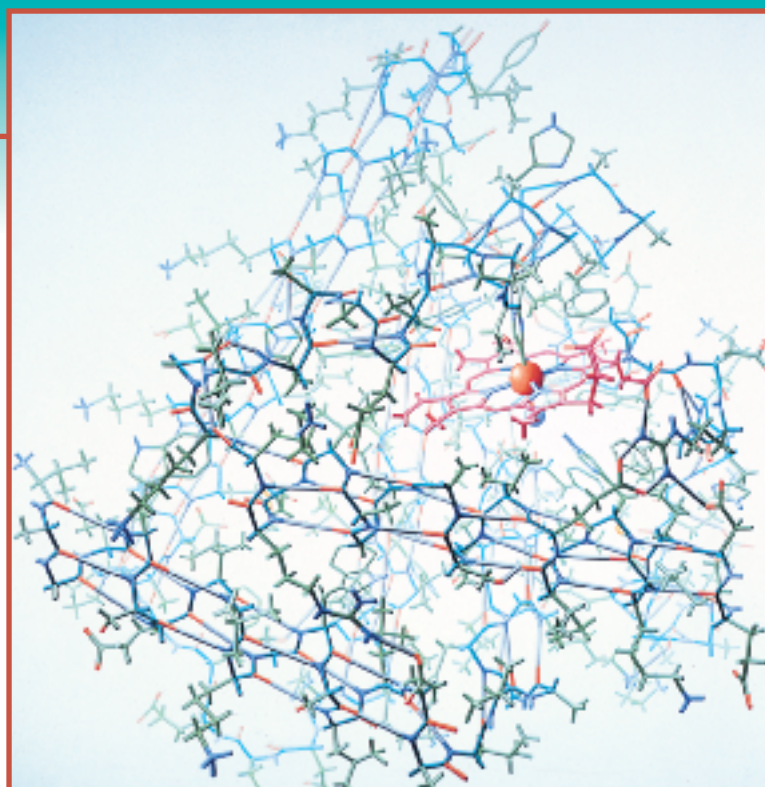


CHAPTER

6



The atomic structure of myoglobin, an oxygen binding protein, is drawn here as a stick model. The overall conformation of a protein such as myoglobin is a function of its amino acid sequence. How do noncovalent forces act on a polypeptide chain to stabilize its unique three-dimensional arrangement of atoms? [Illustrations, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.]

Proteins: Three-Dimensional Structure

1. Secondary Structure
 - A. The Peptide Group
 - B. Regular Secondary Structure: The α Helix and the β Sheet
 - C. Fibrous Proteins
 - D. Nonrepetitive Protein Structure
2. Tertiary Structure
 - A. Determining Protein Structures
 - B. Side Chain Location and Polarity
 - C. Supersecondary Structure and Domains
 - D. Protein Families
3. Quaternary Structure and Symmetry
4. Protein Stability
 - A. Forces That Stabilize Protein Structure
 - B. Protein Dynamics
 - C. Protein Denaturation and Renaturation
5. Protein Folding
 - A. Protein Folding Pathways
 - B. Protein Disulfide Isomerase
 - C. Molecular Chaperones
 - D. Diseases Caused by Protein Misfolding
6. Structural Bioinformatics

For many years, it was thought that proteins were colloids of random structure and that the enzymatic activities of certain crystallized proteins were due to unknown entities associated with an inert protein carrier. In 1934, J.D. Bernal and Dorothy Crowfoot Hodgkin showed that a crystal of the protein **pepsin** yielded a discrete diffraction pattern when placed in an X-ray beam. This result provided the first evidence that pepsin was not a random colloid but an ordered array of atoms organized into a large yet uniquely structured molecule.

Even relatively small proteins contain thousands of atoms, almost all of which occupy definite positions in space. The first X-ray structure of a protein, that of sperm whale myoglobin, was reported in 1958 by John Kendrew and co-workers. At the time—only 5 years after James Watson and Francis Crick had elucidated the simple and elegant structure of DNA (Section 3-2B)—protein chemists were chagrined by the complexity and apparent lack of regularity in the structure of myoglobin. In retrospect, such irregularity seems essential for proteins to fulfill their diverse biological roles. However, comparisons of the nearly 30,000 protein structures now known have revealed that proteins actually exhibit a remarkable degree of structural regularity.

As we saw in Section 5-1, the primary structure of a protein is its linear sequence of amino acids. In discussing protein structure, three further levels of structural complexity are customarily invoked:

- **Secondary structure** is the local spatial arrangement of a polypeptide's backbone atoms without regard to the conformations of its side chains.
- **Tertiary structure** refers to the three-dimensional structure of an entire polypeptide, including that of its side chains.
- Many proteins are composed of two or more polypeptide chains, loosely referred to as subunits. A protein's **quaternary structure** refers to the spatial arrangement of its subunits.

The four levels of protein structure are summarized in Fig. 6-1.

In this chapter, we explore secondary through quaternary structure, including examples of proteins that illustrate each of these levels. We also discuss the process of protein folding and the forces that stabilize folded proteins.

1 Secondary Structure

Protein secondary structure includes the regular polypeptide folding patterns such as helices, sheets, and turns. However, before we discuss these basic structural elements, we must consider the geometric properties of peptide groups, which underlie all higher order structures.

A The Peptide Group

Recall from Section 4-1B that a polypeptide is a polymer of amino acid residues linked by amide (peptide) bonds. In the 1930s and 1940s, Linus Pauling and Robert Corey determined the X-ray structures of several amino acids and dipeptides in an effort to elucidate the conformational constraints on a polypeptide chain. These studies indicated that *the peptide group has*

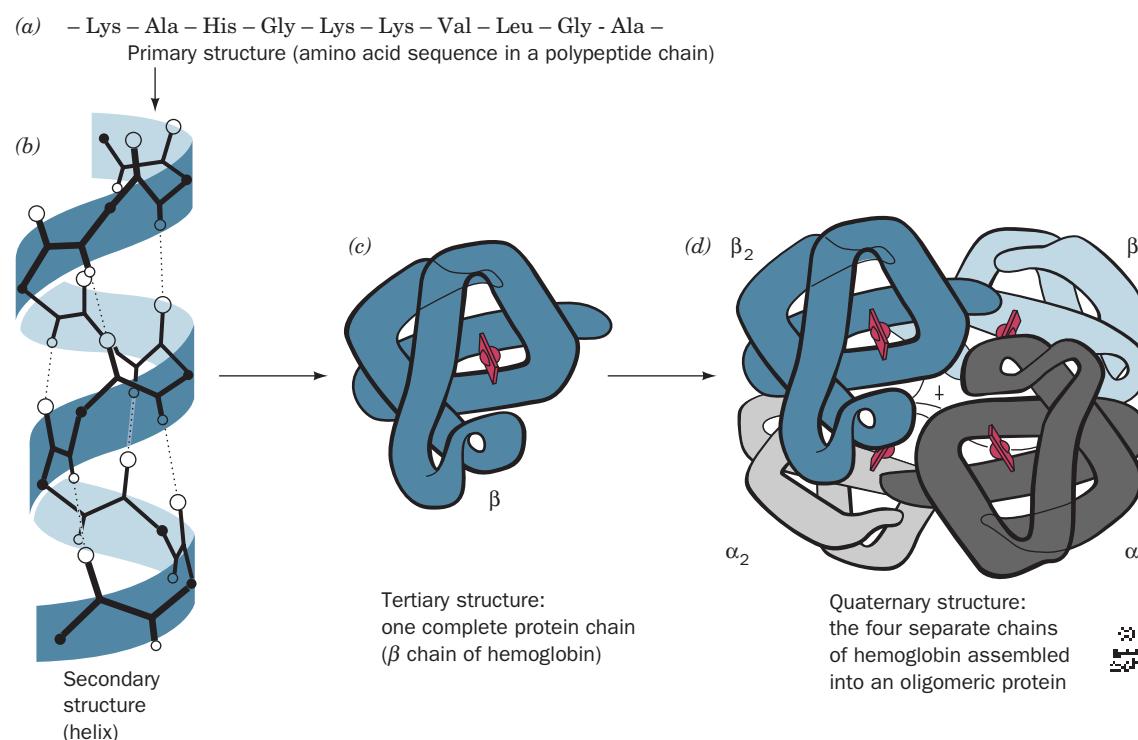
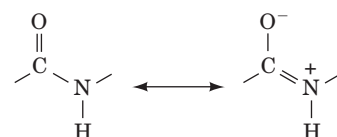


Figure 6-1 Levels of protein structure. (a) Primary structure, (b) secondary structure, (c) tertiary structure, and (d) quaternary structure. [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.]

a rigid, planar structure as a consequence of resonance interactions that give the peptide bond $\sim 40\%$ double-bond character:



This explanation is supported by the observations that a peptide group's C—N bond is 0.13 \AA shorter than its N—C $_{\alpha}$ single bond and that its C=O bond is 0.02 \AA longer than that of aldehydes and ketones. The planar conformation maximizes π -bonding overlap, which accounts for the peptide group's rigidity.

Peptide groups, with few exceptions, assume the **trans conformation**, in which successive C $_{\alpha}$ atoms are on opposite sides of the peptide bond joining them (Fig. 6-2). The **cis conformation**, in which successive C $_{\alpha}$ atoms are on the same side of the peptide bond, is $\sim 8 \text{ kJ} \cdot \text{mol}^{-1}$ less stable than the trans conformation because of steric interference between neighboring side chains. However, this steric interference is reduced in peptide bonds to Pro residues, so $\sim 10\%$ of the Pro residues in proteins follow a cis peptide bond.

Torsion Angles between Peptide Groups Describe Polypeptide Chain Conformations. The **backbone** or **main chain** of a protein refers to the atoms that participate in peptide bonds, ignoring the side chains of the amino acid residues. The backbone can be drawn as a linked sequence of rigid planar peptide groups (Fig. 6-3). The conformation of the backbone can therefore

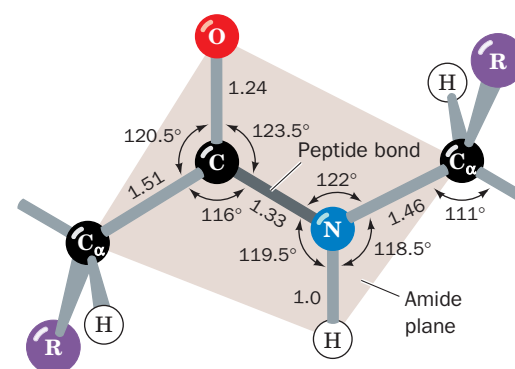


Figure 6-2 The trans peptide group. The bond lengths (in angstroms) and angles (in degrees) are derived from X-ray crystal structures. [After Marsh, R.E. and Donohue, J., *Adv. Protein Chem.* 22, 249 (1967).] See Kinemage Exercise 3-1.

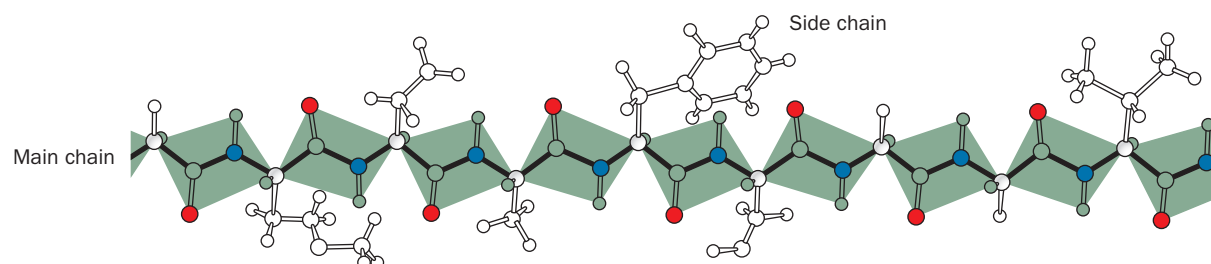


Figure 6-3 Extended conformation of a polypeptide. The backbone is shown as a series of planar peptide groups. [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.]

be described by the **torsion angles** (also called **dihedral angles** or rotation angles) around the C_{α} —N bond (ϕ) and the C_{α} —C bond (ψ) of each residue (Fig. 6-4). These angles, ϕ and ψ , are both defined as 180° when the polypeptide chain is in its fully extended conformation and increase clockwise when viewed from C_{α} .

The conformational freedom and therefore the torsion angles of a polypeptide backbone are sterically constrained. Rotation around the C_{α} —N and C_{α} —C bonds to form certain combinations of ϕ and ψ angles will cause the amide hydrogen, the carbonyl oxygen, or the substituents of C_{α} of adjacent residues to collide (e.g., Fig. 6-5). Certain conformations of

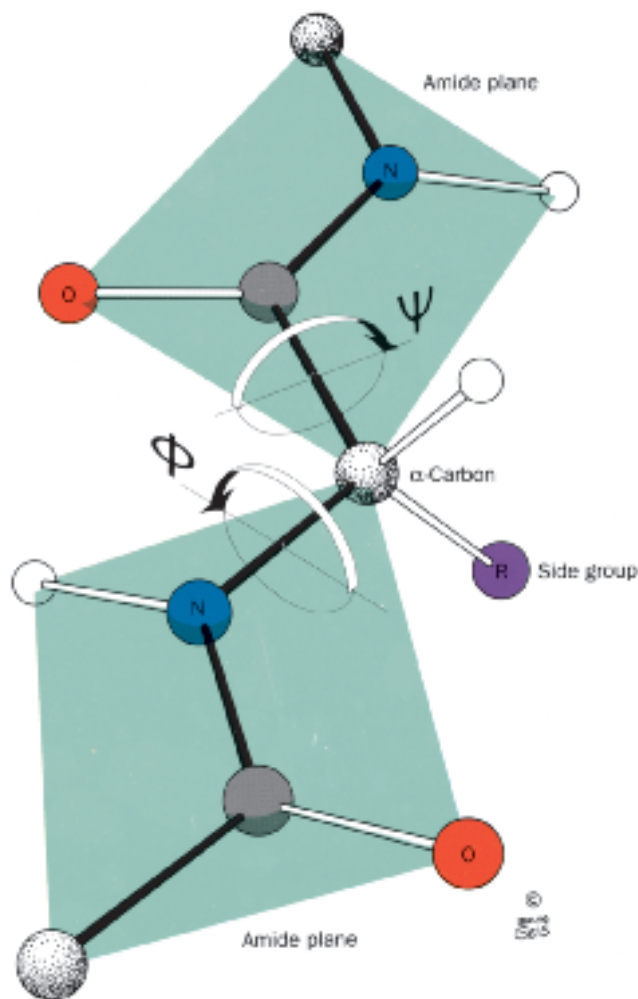


Figure 6-4 Torsion angles of the polypeptide backbone. Two planar peptide groups are shown. The only reasonably free movements are rotations around the C_{α} —N bond (measured as ϕ) and the C_{α} —C bond (measured as ψ). By convention, both ϕ and ψ are 180° in the conformation shown and increase, as indicated, when the peptide plane is rotated in the clockwise direction as viewed from C_{α} . [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.] See Kinemage Exercise 3-1.

longer polypeptides can similarly produce collisions between residues that are far apart in sequence.

The Ramachandran Diagram Indicates Allowed Conformations of Polypeptides. The sterically allowed values of ϕ and ψ can be calculated. Sterically forbidden conformations, such as the one shown in Fig. 6-5, have ϕ and ψ values that would bring atoms closer than the corresponding van der Waals distance (the distance of closest contact between nonbonded atoms). Such information is summarized in a **Ramachandran diagram** (Fig. 6-6), which is named after its inventor, G. N. Ramachandran.

Most areas of the Ramachandran diagram (most combinations of ϕ and ψ) represent forbidden conformations of a polypeptide chain. Only three small regions of the diagram are physically accessible to most residues. The observed ϕ and ψ values of accurately determined structures nearly always fall within these allowed regions of the Ramachandran plot. There are, however, some notable exceptions:

1. The cyclic side chain of Pro limits its range of ϕ values to angles of around -60° , making it, not surprisingly, the most conformationally restricted amino acid residue.
2. Gly, the only residue without a C_β atom, is much less sterically hindered than the other amino acid residues. Hence, its permissible range of ϕ and ψ covers a larger area of the Ramachandran diagram. At Gly residues, polypeptide chains often assume conformations that are forbidden to other residues.

B Regular Secondary Structure: The α Helix and the β Sheet

A few elements of protein secondary structure are so widespread that they are immediately recognizable in proteins with widely differing amino acid sequences. Both the **α helix** and the **β sheet** are such elements; they are called **regular secondary structures** because they are composed of sequences of residues with repeating ϕ and ψ values.

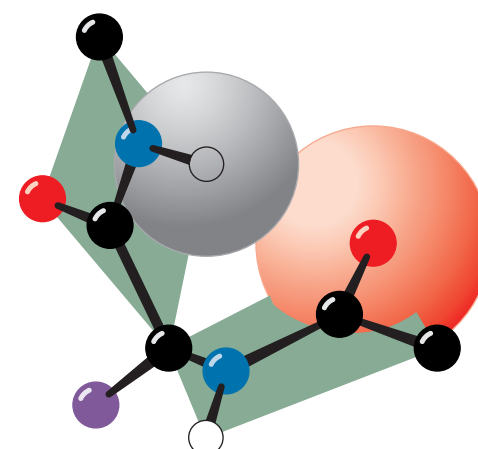


Figure 6-5 Steric interference between adjacent peptide groups. Rotation can result in a conformation in which the amide hydrogen of one residue and the carbonyl oxygen of the next are closer than their van der Waals distance. [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.]

See Guided Exploration 6

Stable helices in proteins: the α helix.

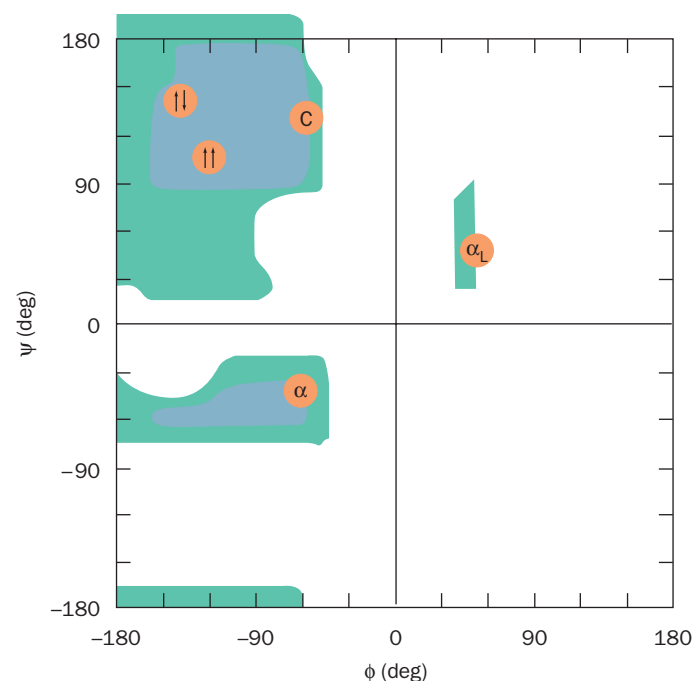


Figure 6-6 The Ramachandran diagram. The blue-shaded regions indicate the sterically allowed ϕ and ψ angles for all residues except Gly and Pro. The green-shaded regions indicate the more crowded (outer limit) ϕ and ψ angles. The orange circles represent conformational angles of several secondary structures: α , right-handed α helix; $\uparrow\uparrow$, parallel β sheet; $\downarrow\downarrow$, antiparallel β sheet; C, collagen helix; α_L , left-handed α helix.



BOX 6-1

*Pathways of Discovery**Linus Pauling and Structural Biochemistry*

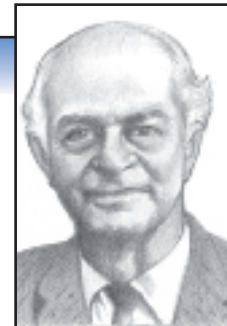
Linus Pauling, the only person to have been awarded two unshared Nobel Prizes, is clearly the dominant figure in twentieth-century chemistry and one of the greatest scientific figures of all time. He received his B.Sc. in chemical engineering from Oregon Agricultural College (now Oregon State University) in 1922 and his Ph.D. in chemistry from the California Institute of Technology in 1925, where he spent much of his career.

The major theme throughout Pauling's long scientific life was the study of molecular structures and the nature of the chemical bond. He began this career by using the then recently invented technique of X-ray crystallography to determine the structures of simple minerals and inorganic salts. At that time, methods for solving the phase problem (see Box 7-2) were unknown, so X-ray structures could only be determined using trial-and-error techniques. This limited the possible molecules that could be effectively studied to those with few atoms and high symmetry such that their atomic coordinates could be fully described by only a few parameters (rather than the three-dimensional coordinates of each of its atoms). Pauling realized that the positions of atoms in molecules were governed by fixed atomic radii, bond distances, and bond angles and used this information to make educated guesses about molecular structures. This greatly extended the complexity of the molecules whose structures could be determined.

In his next major contribution, occurring in 1931, Pauling revolutionized the way that chemists viewed molecules by applying the then infant field of quantum mechanics to chemistry. Pauling formulated the theories of orbital hybridization, electron-pair bonding, and resonance and thereby explained the nature of covalent bonds. This work was summarized in his highly influential monograph *The Nature of the Chemical Bond*, which was first published in 1938.

In the mid-1930s, Pauling turned his attention to biological chemistry. He began these studies in collaboration with his colleague, Robert Corey, by determining the X-ray structures of several amino acids and dipeptides. At that time, the X-ray structural determination of even such small mole-

(Linus Pauling, 1901–1994)



cules required around a year of intense effort, largely because the numerous calculations required to solve a structure had to be made by hand (electronic computers had yet to be invented). Nevertheless, these studies led Pauling and Corey to the conclusions that the peptide bond is planar, which Pauling explained using results from resonance considerations (Section 6-1A), and that hydrogen bonding plays a central role in maintaining macromolecular structures.

In the 1940s, Pauling made several unsuccessful attempts to determine if polypeptides have any preferred conformations. Then, in 1948, while visiting Oxford University, he was confined to bed by a cold. He eventually tired of reading detective stories and science fiction and again turned his attention to proteins. By folding drawings of polypeptides in various ways, he discovered the α helix, whose existence was rapidly confirmed by X-ray studies of α keratin (Section 6-1C). This work was reported in 1951, and later that year Pauling and Corey also proposed both the parallel and antiparallel β pleated sheets. For these ground-breaking insights, Pauling received the Nobel Prize in Chemistry in 1954, although α helices and β sheets were not actually visualized until the first X-ray structures of proteins were determined, five to ten years later.

Pauling made numerous additional pioneering contributions to biological chemistry, most notably that the heme group in hemoglobin changes its electronic state on binding oxygen (Section 7-1A), that vertebrate hemoglobins are $\alpha_2\beta_2$ heterotetramers (Section 7-2A), that the denaturation of proteins is caused by the unfolding of their polypeptide chains, that sickle-cell anemia is caused by a mutation in the β chain of normal adult hemoglobin (the first so-called molecular disease to be characterized; Section 7-4), that molecular complementarity plays an important role in antibody–antigen interactions (Section 28-5C) and by extension all macromolecular interactions, that enzymes catalyze reactions by preferentially binding their transition states (Section 11-3E), and that the

The α Helix. Only one polypeptide helix has both a favorable hydrogen-bonding pattern and ϕ and ψ values that fall within the fully allowed regions of the Ramachandran diagram: the α helix. Its discovery by Linus Pauling in 1951, through model building, ranks as one of the landmarks of structural biochemistry (Box 6-1).

The α helix (Fig. 6-7) is right-handed; that is, it turns in the direction that the fingers of a right hand curl when its thumb points in the direction that

comparison of the sequences of the corresponding proteins in different organisms yields evolutionary insights (Section 5-4).

Pauling was also a lively and stimulating lecturer who for many years taught a general chemistry course [which one of the authors of this textbook (DV) had the privilege of taking]. His textbook, *General Chemistry*, revolutionized the way that introductory chemistry was taught by presenting it as a subject that could be understood in terms of atomic physics and molecular structure. For a book of such generality, an astounding portion of its subject matter had been elucidated by its author. Pauling's amazing grasp of chemistry was demonstrated by the fact that he dictated each chapter of this textbook in a single sitting.

By the late 1940s, Pauling became convinced that the possibility of nuclear war posed an enormous danger to humanity and calculated that the radioactive fallout from each above-ground test of a nuclear bomb would ultimately cause cancer in thousands of people. He therefore began a campaign to educate the public about the hazards of bomb testing and nuclear war. The political climate in the United States at the time was such that the government considered Pauling to be subversive and his passport was revoked (and only returned two weeks before he was to leave for Sweden to receive his first Nobel Prize). Nevertheless, Pauling persisted in this campaign, which culminated, in 1962, with the signing of the first Nuclear Test Ban Treaty. For his efforts, Pauling was awarded the 1962 Nobel Peace Prize.

Pauling saw science as the search for the truth, which included politics and social causes. In his later years, he became a vociferous promoter of what he called orthomolecular medicine, the notion that large doses of vitamins could ward off and cure many human diseases, including cancer. In the best known manifestation of this concept, Pauling advocated taking large doses of vitamin C to prevent the common cold and lessen its symptoms, advice still followed by millions of people, although the medical evidence supporting this notion is scant. It should be noted, however, that Pauling, who followed his own advice, remained active until he died in 1994 at the age of 93.

the helix rises (Fig. 3-7). The α helix has 3.6 residues per turn and a **pitch** (the distance the helix rises along its axis per turn) of 5.4 Å. The α helices of proteins have an average length of ~ 12 residues, which corresponds to over three helical turns, and a length of ~ 18 Å.

In the α helix, the backbone hydrogen bonds are arranged such that the peptide C=O bond of the n th residue points along the helix axis toward the peptide N—H group of the $(n + 4)$ th residue. This results in a strong hy-

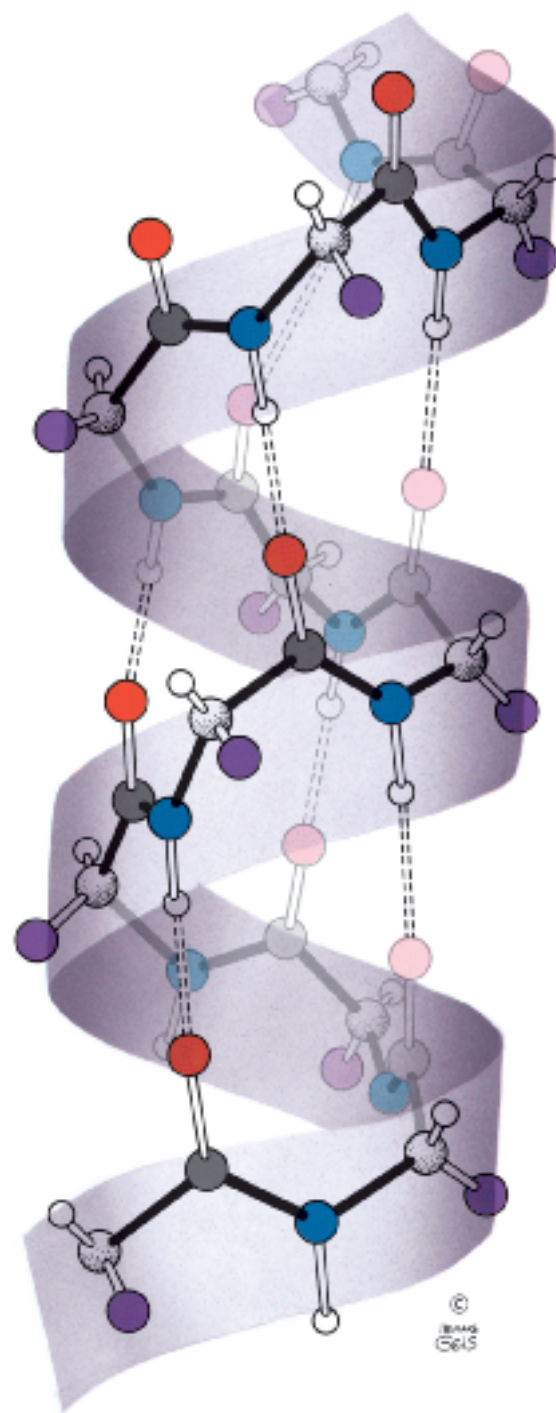


Figure 6-7 Key to Structure. The α helix. This right-handed helical conformation has 3.6 residues per turn. Dashed lines indicate hydrogen bonds between C=O groups and N—H groups that are four residues farther along the polypeptide chain. [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.] See Kinemage Exercise 3-2 and the Animated Figures.



Figure 6-8 Space-filling model of an α helix. The backbone atoms are colored according to type with C green, N blue, O red, and H white. The side chains (*yellow*) project away from the helix. This α helix is a segment of sperm whale myoglobin.

drogen bond that has the nearly optimum $N\cdots O$ distance of 2.8 Å. Amino acid side chains project outward and downward from the helix (Fig. 6-8), thereby avoiding steric interference with the polypeptide backbone and with each other. The core of the helix is tightly packed; that is, its atoms are in van der Waals contact.

β Sheets. In 1951, the same year Pauling proposed the α helix, Pauling and Corey postulated the existence of a different polypeptide secondary structure, the β sheet. Like the α helix, the β sheet uses the full hydrogen-bonding capacity of the polypeptide backbone. *In β sheets, however, hydrogen bonding occurs between neighboring polypeptide chains rather than within one as in an α helix.*

Sheets come in two varieties:

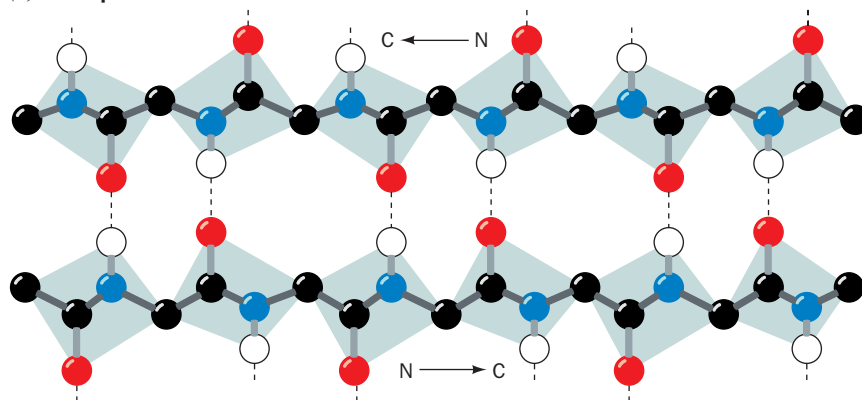
1. The **antiparallel β sheet**, in which neighboring hydrogen-bonded polypeptide chains run in opposite directions (Fig. 6-9a).
2. The **parallel β sheet**, in which the hydrogen-bonded chains extend in the same direction (Fig. 6-9b).

The conformations in which these β structures are optimally hydrogen bonded vary somewhat from that of the fully extended polypeptide shown

See Guided Exploration 7

Hydrogen bonding in β sheets.

(a) Antiparallel



(b) Parallel

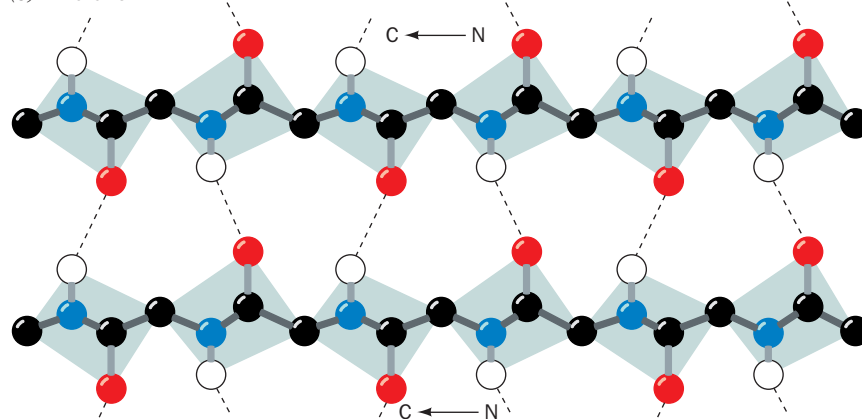



Figure 6-9 Key to Structure. β Sheets. Dashed lines indicate hydrogen bonds between polypeptide strands. Side chains are omitted for clarity. (a) An antiparallel β sheet, (b) A parallel β sheet. [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.]  See Kinemage Exercise 3-3 and the Animated Figures.

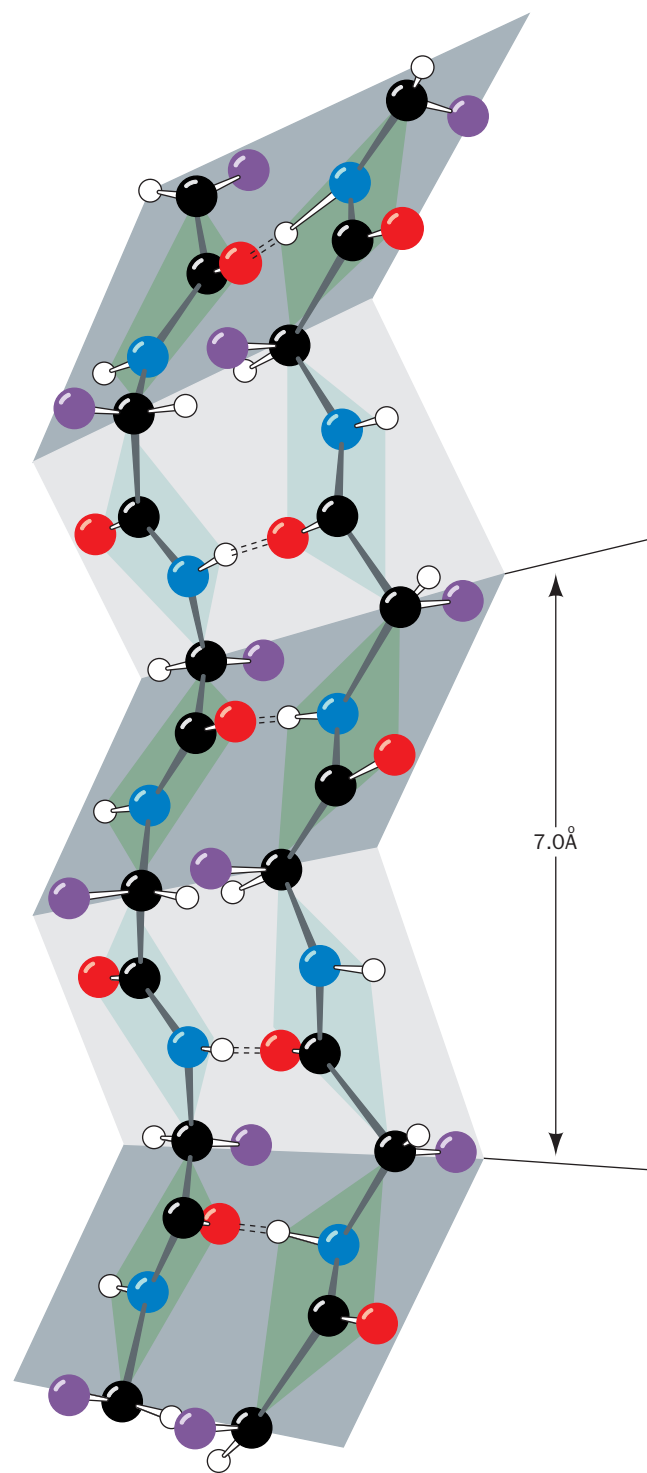



Figure 6-10 Pleated appearance of a β sheet.

Dashed lines indicate hydrogen bonds. The R groups (purple) on each polypeptide chain alternately extend to opposite sides of the sheet and are in register on adjacent chains. [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.]  See Kinemage Exercise 3-3.

in Fig. 6-3. They therefore have a rippled or pleated edge-on appearance (Fig. 6-10) and for that reason are sometimes called “pleated sheets.” Successive side chains of a polypeptide chain in a β sheet extend to opposite sides of the sheet with a two-residue repeat distance of 7.0 \AA .

β Sheets in proteins contain 2 to as many as 22 polypeptide strands, with an average of 6 strands. Each strand may contain up to 15 residues, the average being 6 residues. A six-stranded antiparallel β sheet is shown in Fig. 6-11.

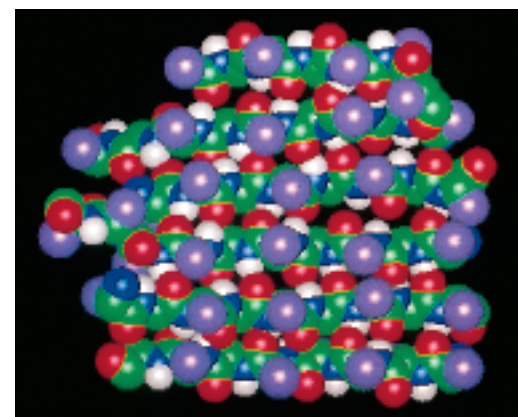


Figure 6-11 Space-filling model of a β sheet. The backbone atoms are colored according to type with C green, N blue, O red, and H white. The R groups are represented by large purple spheres. This six-stranded antiparallel β sheet is from the jack bean protein **concanavalin A**.

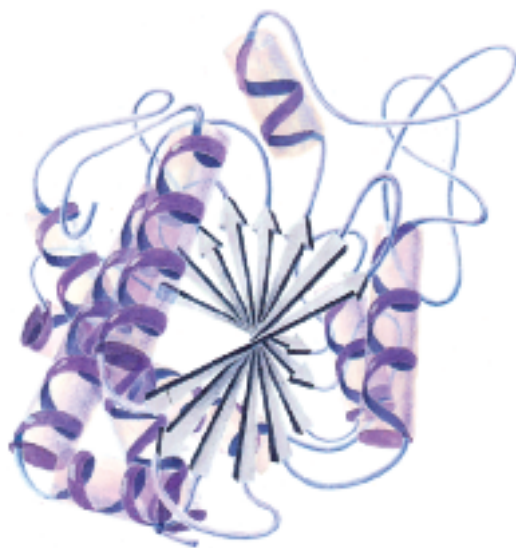


Figure 6-12 Diagram of a β sheet in bovine carboxypeptidase A. The polypeptide backbone is represented by a ribbon with α helices drawn as coils and strands of the β sheet drawn as purple arrows pointing toward the C-terminus. Side chains are not shown. The eight-stranded β sheet forms a saddle-shaped curved surface with a right-handed twist. [After a drawing by Jane Richardson, Duke University. Based on an X-ray structure by William Lipscomb, Harvard University. PDBid 3CPA (for the definition of “PDBid”, see Section 6-6).]

Parallel β sheets containing fewer than five strands are rare. This observation suggests that parallel β sheets are less stable than antiparallel β sheets, possibly because the hydrogen bonds of parallel sheets are distorted compared to those of the antiparallel sheets (Fig. 6-9). β Sheets containing mixtures of parallel and antiparallel strands frequently occur.

β Sheets almost invariably exhibit a pronounced right-handed twist when viewed along their polypeptide strands (Fig. 6-12). Conformational energy calculations indicate that the twist is a consequence of interactions between chiral L-amino acid residues in the extended polypeptide chains. The twist distorts and weakens the β sheet’s interchain hydrogen bonds. The geometry of a particular β sheet is thus a compromise between optimizing the conformational energies of its polypeptide chains and preserving its hydrogen bonding.

The **topology** (connectivity) of the polypeptide strands in a β sheet can be quite complex. The connection between two antiparallel strands may be just a small loop (Fig. 6-13a), but the link between tandem parallel strands must be a crossover connection that is out of the plane of the β sheet (Fig. 6-13b). The connecting link in either case can be extensive, often containing helices (e.g., Fig. 6-12).

C Fibrous Proteins

Proteins have historically been classified as either **fibrous** or **globular**, depending on their overall morphology. This dichotomy predates methods for determining protein structure on an atomic scale and does not do justice to proteins that contain both stiff, elongated, fibrous regions as well as more compact, highly folded, globular regions. Nevertheless, the division helps emphasize the properties of fibrous proteins, which often have a protective, connective, or supportive role in living organisms. The two well-characterized fibrous proteins we discuss here—keratin and collagen—are highly elongated molecules whose shapes are dominated by a single type of secondary structure. They are therefore useful examples of these structural elements.

α Keratin—A Coiled Coil. **Keratin** is a mechanically durable and chemically unreactive protein that occurs in all higher vertebrates. It is the principal component of their horny outer epidermal layer and its related appendages such as hair, horn, nails, and feathers. Keratins have been classified as either α keratins, which occur in mammals, or β keratins, which occur in birds and reptiles. Mammals each have over 30 keratin variants that are expressed in a tissue-specific manner.

The X-ray diffraction pattern of α keratin resembles that expected for an α helix (hence the name α keratin). However, α keratin exhibits a 5.1-Å spacing rather than the 5.4-Å distance corresponding to the pitch of the α helix. This discrepancy is the result of *two α keratin polypeptides, each of which forms an α helix, twisting around each other to form a left-handed coil*. The normal 5.4-Å repeat distance of each α helix in the pair is thereby tilted relative to the axis of this assembly, yielding the observed 5.1-Å spacing. The assembly is said to have a **coiled coil** structure because each α helix itself follows a helical path.

The conformation of α keratin’s coiled coil is a consequence of its primary structure: The central ~ 310 -residue segment of each polypeptide chain has a 7-residue pseudorepeat, *a-b-c-d-e-f-g*, with nonpolar residues predominating at positions *a* and *d*. Since an α helix has 3.6 residues per turn, α keratin’s *a* and *d* residues line up along one side of each α helix (Fig. 6-14a). The hydrophobic strip along one helix associates with the hy-

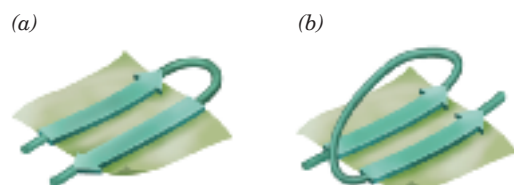


Figure 6-13 Connections between adjacent strands in β sheets. (a) Antiparallel strands may be connected by a small loop. (b) Parallel strands require a more extensive crossover connection. [After Richardson, J.S., *Adv. Protein Chem.* **34**, 196 (1981).]

drophobic strip on another helix. Because the 3.5-residue repeat in α keratin is slightly smaller than the 3.6 residues per turn of a standard α helix, the two keratin helices are inclined about 18° relative to one another, resulting in the coiled coil arrangement. This conformation allows the contacting side chains of each helix to interdigitate (Fig. 6-14b). Coiled coils also occur in numerous, not necessarily fibrous, proteins.

The higher order structure of α keratin is not well understood. The N- and C-terminal domains of each polypeptide facilitate the assembly of coiled coils (dimers) into protofilaments, two of which constitute a protofibril (Fig. 6-15). Four protofibrils constitute a microfibril, which associates with other microfibrils to form a macrofibril. A single mammalian hair consists of layers of dead cells, each of which is packed with parallel microfibrils.

α Keratin is rich in Cys residues, which form disulfide bonds that cross-link adjacent polypeptide chains. The α keratins are classified as “hard” or “soft” according to whether they have a high or low sulfur content. Hard keratins, such as those of hair, horn, and nail, are less pliable than soft keratins, such as those of skin and callus, because the disulfide bonds resist deformation. The disulfide bonds can be reductively cleaved by disulfide interchange with mercaptans (Section 5-3A). Hair so treated can be curled and set in a “permanent wave” by applying an oxidizing agent that reestablishes the disulfide bonds in the new “curled” conformation. Conversely, curly hair can be straightened by the same process.

The springiness of hair and wool fibers is a consequence of the coiled coil’s tendency to recover its original conformation after being untwisted by stretching. If some of its disulfide bonds have been cleaved, however, an α keratin fiber can be stretched to over twice its original length. At this point, the polypeptide chains assume a β sheet conformation. β Keratin, such as that in feathers, exhibits a β -like pattern in its native state.

Collagen—A Triple Helix. Collagen, which occurs in all multicellular animals, is the most abundant vertebrate protein. Its strong, insoluble fibers are the major stress-bearing components of connective tissues such as bone, teeth, cartilage, tendon, and the fibrous matrices of skin and blood vessels. A single collagen molecule consists of three polypeptide chains. Mammals

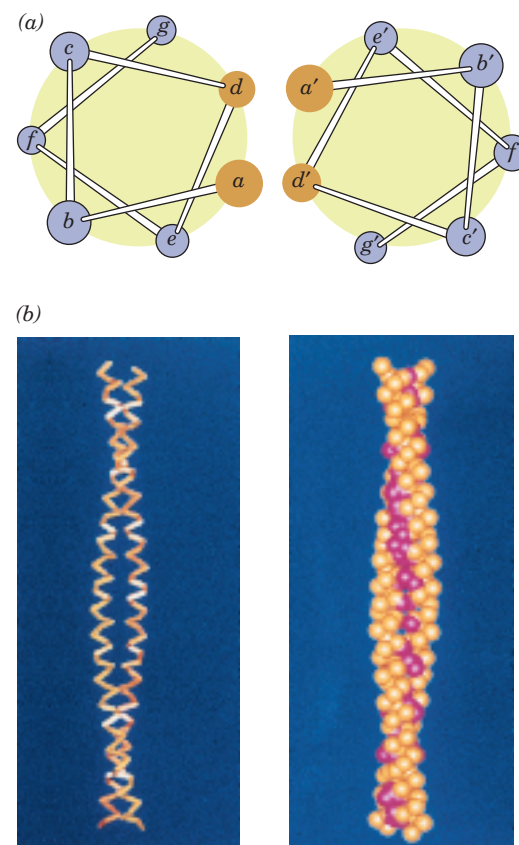


Figure 6-14 A coiled coil. (a) View down the coil axis showing the alignment of nonpolar residues along one side of each α helix. The helices have the pseudorepeating sequence $a-b-c-d-e-f-g$ in which residues a and d are predominately nonpolar. [After McLachlan, A.D. and Stewart, M., *J. Mol. Biol.* **98**, 295 (1975).] (b) Side view of the polypeptide backbones in skeletal (left) and space-filling (right) forms. Note that the side chains (red spheres in the space-filling model) contact each other. This coiled coil is from the protein tropomyosin. [Courtesy of Carolyn Cohen, Brandeis University.] See Kinemage Exercises 4-1 and 4-2.

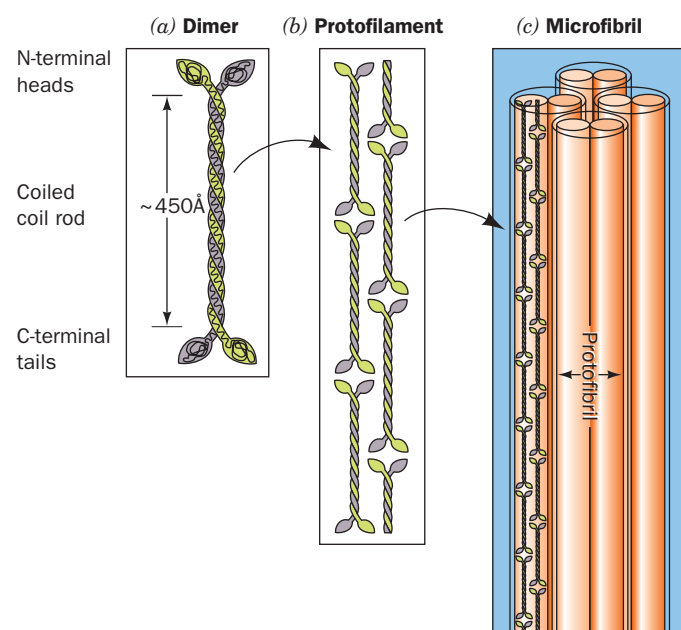
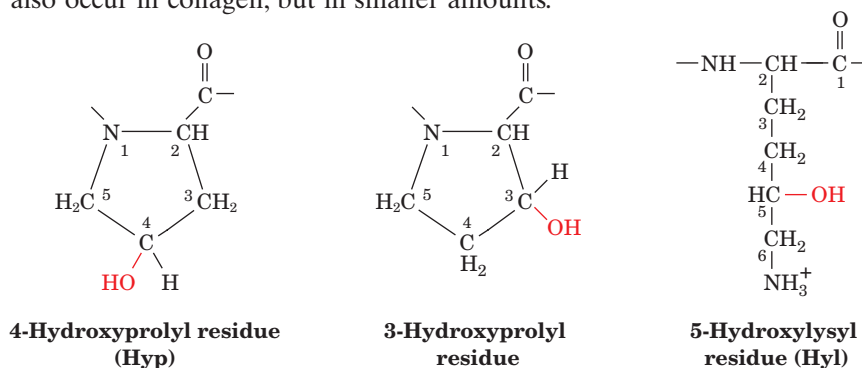


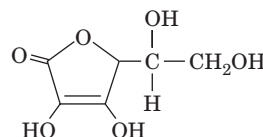
Figure 6-15 Higher order α keratin structure. (a) Two keratin polypeptides form a dimeric coiled coil. (b) Protofilaments are formed from two staggered rows of head-to-tail associated coiled coils. (c) Protofilaments dimerize to form a protofibril, four of which form a microfibril. The structures of the latter assemblies are poorly characterized.

have at least 33 genetically distinct chains that are assembled into at least 20 collagen varieties found in different tissues in the same individual. One of the most common collagens, called Type I, consists of two $\alpha_1(\text{I})$ chains and one $\alpha_2(\text{I})$ chain. It has a molecular mass of ~ 285 kD, a width of ~ 14 Å, and a length of ~ 3000 Å.

Collagen has a distinctive amino acid composition: Nearly one-third of its residues are Gly; another 15 to 30% of its residues are Pro and **4-hydroxyprolyl (Hyp)**. **3-Hydroxyprolyl** and **5-hydroxylysyl (Hyl)** residues also occur in collagen, but in smaller amounts.



These nonstandard residues are formed after the collagen polypeptides are synthesized. For example, Pro residues are converted to Hyp in a reaction catalyzed by **prolyl hydroxylase**. This enzyme requires **ascorbic acid (vitamin C)** to maintain its activity.



Ascorbic acid (vitamin C)

The disease **scurvy** results from the dietary deficiency of vitamin C (see Box 6-2).

The amino acid sequence of a typical collagen polypeptide consists of monotonously repeating triplets of sequence Gly-X-Y over a segment of ~ 1000 residues, where X is often Pro and Y is often Hyp. Hyl sometimes appears at the Y position. Collagen's Pro residues prevent it from forming an α helix (Pro residues cannot assume the α -helical backbone conformation and lack the backbone N—H groups that form the intrahelical hydrogen bonds shown in Fig. 6-7). Instead, *the collagen polypeptide assumes a left-handed helical conformation with about three residues per turn. Three parallel chains wind around each other with a gentle, right-handed, ropelike twist to form the triple-helical structure of a collagen molecule* (Fig. 6-16).

This model of the collagen structure has been confirmed by Barbara Brodsky and Helen Berman, who determined the X-ray crystal structure of a collagen-like model polypeptide. Every third residue of each polypeptide chain passes through the center of the triple helix, which is so crowded that only a Gly side chain can fit there. This crowding explains the absolute requirement for a Gly at every third position of a collagen polypeptide chain. The three polypeptide chains are staggered so that a Gly, X, and Y residue occurs at each position along the triple helix axis (Fig. 6-17a). The peptide groups are oriented such that the N—H of each Gly makes a strong hydrogen bond with the carbonyl oxygen of an X (Pro) residue on a neighboring chain (Fig. 6-17b). The bulky and relatively inflexible Pro and Hyp residues confer rigidity on the entire assembly.



Figure 6-16 The collagen triple helix. Left-handed polypeptide helices are twisted together to form a right-handed superhelical structure. [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.]



BOX 6-2

*Biochemistry in Health and Disease**Collagen Diseases*

Some collagen diseases have dietary causes. In scurvy (caused by vitamin C deficiency), Hyp production decreases because prolyl hydroxylase requires vitamin C. Thus, in the absence of vitamin C, newly synthesized collagen cannot form fibers properly, resulting in skin lesions, fragile blood vessels, poor wound healing, and, ultimately, death. Scurvy was common in sailors on long voyages whose diets were devoid of fresh foods. The introduction of limes to the diet of the British navy by the renowned explorer Captain James Cook alleviated scurvy and led to the nickname “limey” for the British sailor.

The disease **lathyrism** is caused by regular ingestion of the seeds from the sweet pea *Lathyrus odoratus*, which contain a compound that specifically inactivates lysyl oxidase. The resulting reduced cross-linking of collagen fibers produces serious abnormalities of the bones, joints, and large blood vessels.

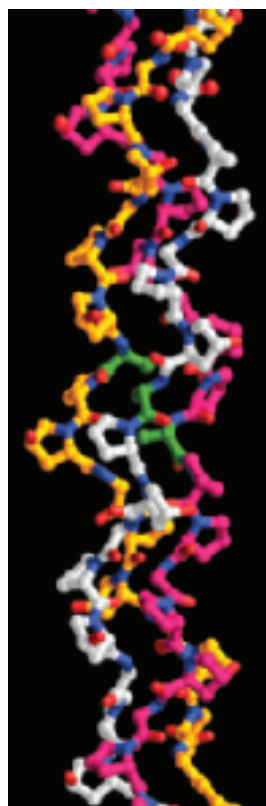
Several rare heritable disorders of collagen are known. Mutations of Type I collagen, which constitutes the major structural protein in most human tissues, usually result in **osteogenesis imperfecta** (brittle bone disease). The severity of this disease varies with the nature and position of the

mutation: Even a single amino acid change can have lethal consequences. For example, the central Gly \rightarrow Ala substitution in the model polypeptide shown in Fig. 6-17 locally distorts the already internally crowded collagen helix. This ruptures the hydrogen bond from the backbone N—H of each Ala (normally Gly) to the carbonyl group of the adjacent Pro in a neighboring chain, thereby reducing the stability of the collagen structure.

Mutations may affect the structure of the collagen molecule or how it forms fibrils. These mutations tend to be dominant because they affect either the folding of the triple helix or fibril formation even when normal chains are also involved.

Many collagen disorders are characterized by deficiencies in the amount of a particular collagen type synthesized, or by abnormal activities of collagen-processing enzymes such as lysyl hydroxylase and lysyl oxidase. One group of at least 10 different collagen-deficiency diseases, the **Ehlers–Danlos syndromes**, are all characterized by the hyperextensibility of the joints and skin. The “India-rubber man” of circus fame had an Ehlers–Danlos syndrome.

(a)



(b)

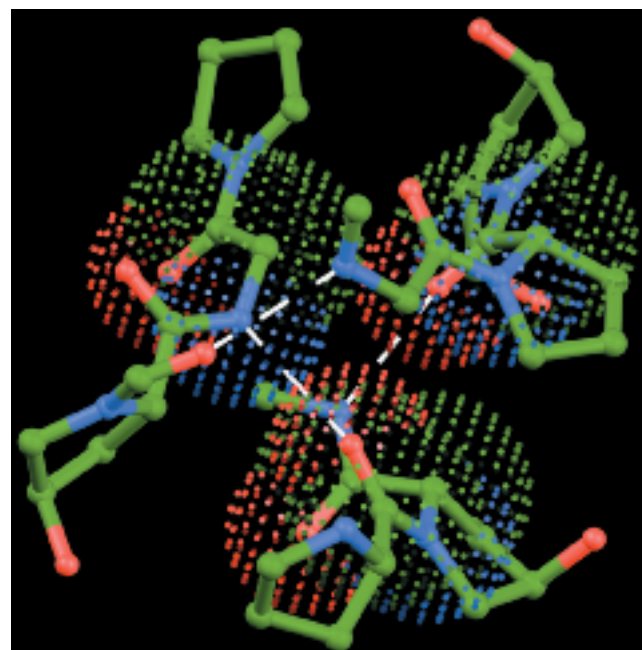



Figure 6-17 Structure of a collagen model peptide. In this X-ray structure of (Pro-Hyp-Gly)₁₀, the fifth Gly of each peptide has been replaced by Ala. (a) A ball-and-stick model of the middle portion of the triple helix oriented with the N-termini at the top. The C atoms of the three chains are colored gold, magenta, and white. The N and O atoms on all chains are blue and red. Note how the replacement of Gly with the bulkier Ala (C atoms in green) distorts the triple helix. (b) This view from the N-terminus down the helix axis shows the interchain hydrogen-bonding associations. Three consecutive residues from each chain are shown in ball-and-stick form. Hydrogen bonds are represented by dashed lines from Gly N atoms to Pro O atoms in adjacent chains. Dots represent the van der Waals surfaces of the backbone atoms of the central residue in each chain. Note the close packing of the atoms along the triple helix axis. [Based on an X-ray structure by Helen Berman, Rutgers University, and Barbara Brodsky, UMDNJ–Robert Wood Johnson Medical School. PDBid 1CAG.]  See Kinemage Exercises 4-3 and 4-4.

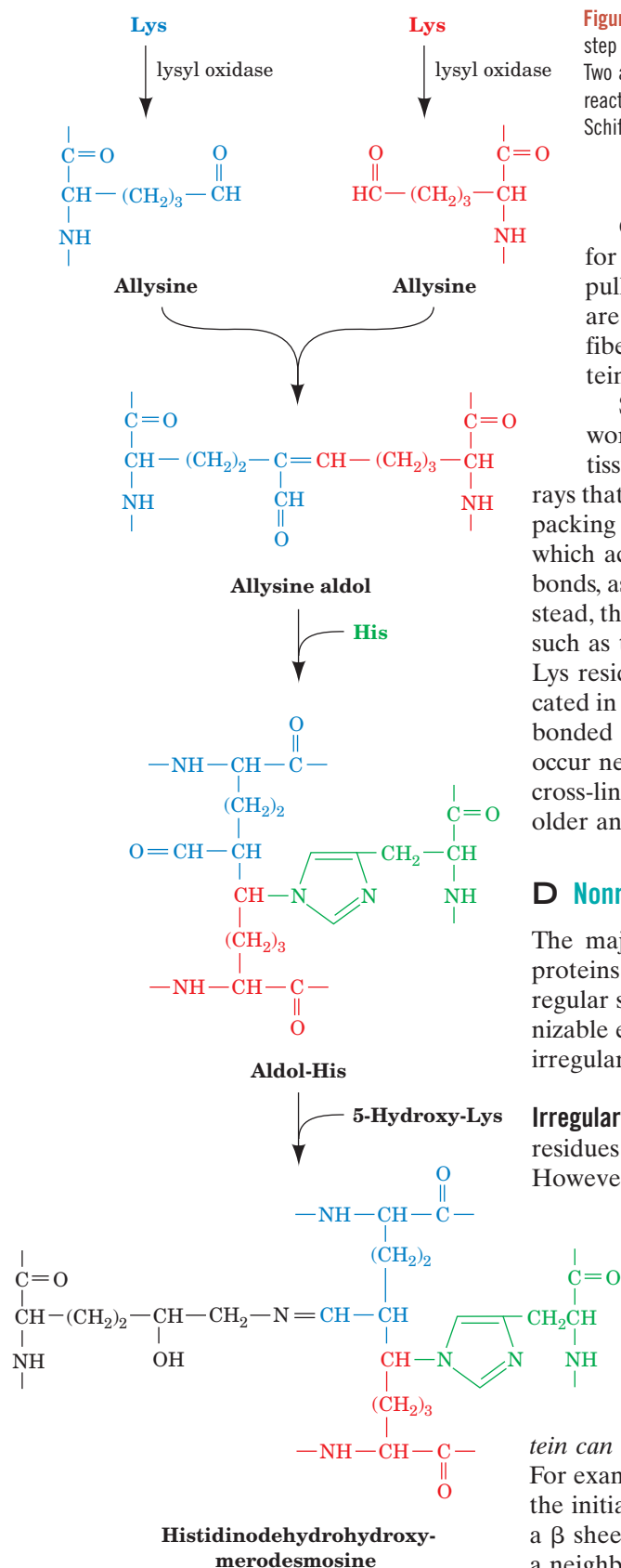


Figure 6-18 A reaction pathway for cross-linking side chains in collagen. The first step is the lysyl oxidase-catalyzed oxidative deamination of Lys to form the aldehyde allysine. Two allysines then undergo an aldol condensation to form allysine aldol. This product can react with His to form aldol histidine, which can in turn react with 5-hydroxylysine to form a Schiff base (an imine bond), thereby cross-linking four side chains.

Collagen's well-packed, rigid, triple-helical structure is responsible for its characteristic tensile strength. The twist in the helix cannot be pulled out under tension because its component polypeptide chains are twisted in the opposite direction (Fig. 6-16). Successive levels of fiber bundles in high-quality ropes and cables, as well as in other proteins such as keratin (Fig. 6-14), are likewise oppositely twisted.

Several types of collagen molecules assemble to form loose networks or thick fibrils arranged in bundles or sheets, depending on the tissue. The collagen molecules in fibrils are organized in staggered arrays that are stabilized by hydrophobic interactions resulting from the close packing of triple-helical units. Collagen is also covalently cross-linked, which accounts for its poor solubility. The cross-links cannot be disulfide bonds, as in keratin, because collagen is almost devoid of Cys residues. Instead, the cross-links are derived from Lys and His side chains in reactions such as those shown in Fig. 6-18. **Lysyl oxidase**, the enzyme that converts Lys residues to those of the aldehyde **allysine**, is the only enzyme implicated in this cross-linking process. Up to four side chains can be covalently bonded to each other. The cross-links do not form at random but tend to occur near the N- and C-termini of the collagen molecules. The degree of cross-linking in a particular tissue increases with age. This is why meat from older animals is tougher than meat from younger animals.

D Nonrepetitive Protein Structure

The majority of proteins are globular proteins that, unlike the fibrous proteins discussed in the preceding section, may contain several types of regular secondary structure, including α helices, β sheets, and other recognizable elements. A significant portion of a protein's structure may also be irregular or unique.

Irregular Structures. Segments of polypeptide chains whose successive residues do not have similar ϕ and ψ values are sometimes called coils. However, you should not confuse this term with the appellation **random coil**, which refers to the totally disordered and rapidly fluctuating conformations assumed by **denatured** (fully unfolded) proteins in solution. In **native** (folded) proteins, *nonrepetitive structures are no less ordered than are helices or β sheets; they are simply irregular and hence more difficult to describe.*

Variations in Standard Secondary Structure. Variations in amino acid sequence as well as the overall structure of the folded protein can distort the regular conformations of secondary structural elements. For example, the α helix frequently deviates from its ideal conformation in the initial and final turns of the helix. Similarly, a strand of polypeptide in a β sheet may contain an "extra" residue that is not hydrogen bonded to a neighboring strand, producing a distortion known as a **β bulge**.

Many of the limits on amino acid composition and sequence (Section 5-1) may be due in part to conformational constraints in the three-dimensional

structure of proteins. For example, a Pro residue produces a kink in an α helix or β sheet. Similarly, steric clashes between several sequential amino acid residues with large branched side chains (e.g., Ile and Tyr) can destabilize α helices.

Analysis of known protein structures by Peter Chou and Gerald Fasman revealed the propensity P of a residue to occur in an α helix or a β sheet (Table 6-1). Chou and Fasman also discovered that certain residues not only have a high propensity for a particular secondary structure but they tend to disrupt or break other secondary structures. Such data are useful for predicting the secondary structures of proteins with known amino acid sequences.

The presence of certain residues outside of α helices or β sheets may also be nonrandom. For example, α helices are often flanked by residues such as Asn and Gln, whose side chains can fold back to form hydrogen bonds with one of the four terminal residues of the helix, a phenomenon termed **helix capping**. Recall that the four residues at each end of an α helix are not fully hydrogen bonded to neighboring backbone segments (Fig. 6-7).

Turns and Loops. Segments with regular secondary structure such as α helices or the strands of β sheets are typically joined by stretches of polypeptide that abruptly change direction. Such **reverse turns** or **β bends** (so named because they often connect successive strands of antiparallel β sheets) almost always occur at protein surfaces. Most reverse turns involve four successive amino acid residues more or less arranged in one of two ways, Type I and Type II, that differ by a 180° flip of the peptide unit linking residues 2 and 3 (Fig. 6-19). Both types of turns are stabilized by a hydrogen bond, although deviations from these ideal conformations often disrupt this hydrogen bond. In Type II turns, the oxygen atom of residue 2 crowds the C_β atom of residue 3, which is therefore usually Gly. Residue 2 of either type of turn is often Pro since it can assume the required conformation.

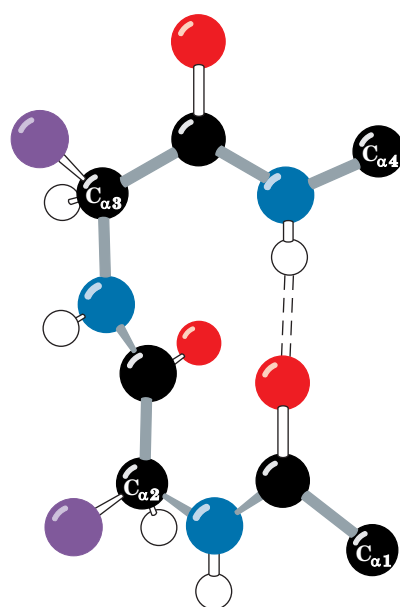
Almost all proteins with more than 60 residues contain one or more loops of 6 to 16 residues, called **Ω loops**. These loops, which have the

Table 6-1 Propensities of Amino Acid Residues for α Helical and β Sheet Conformations

Residue	P_α	P_β
Ala	1.42	0.83
Arg	0.98	0.93
Asn	0.67	0.89
Asp	1.01	0.54
Cys	0.70	1.19
Gln	1.11	1.10
Glu	1.51	0.37
Gly	0.57	0.75
His	1.00	0.87
Ile	1.08	1.60
Leu	1.21	1.30
Lys	1.16	0.74
Met	1.45	1.05
Phe	1.13	1.38
Pro	0.57	0.55
Ser	0.77	0.75
Thr	0.83	1.19
Trp	1.08	1.37
Tyr	0.69	1.47
Val	1.06	1.70

Source: Chou, P.Y. and Fasman, G.D., *Annu. Rev. Biochem.* **47**, 258 (1978).

(a) Type I



(b) Type II

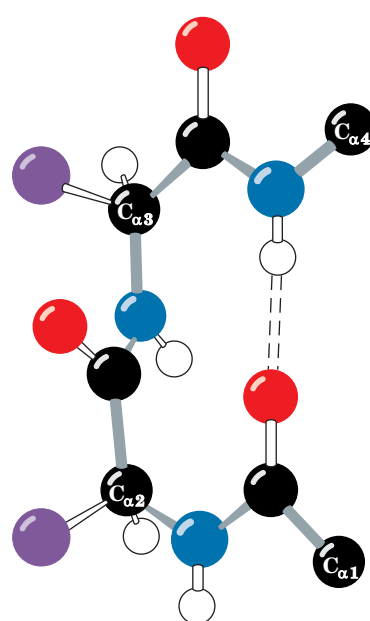


Figure 6-19 Reverse turns in polypeptide chains. Dashed lines represent hydrogen bonds. (a) Type I. (b) Type II. [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.]

See Kinemage Exercise 3-4.

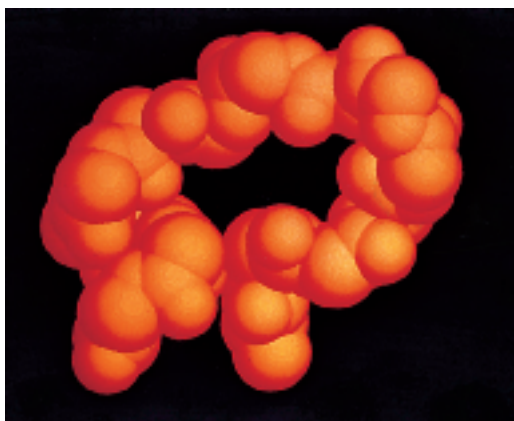


Figure 6-20 Space-filling model of an Ω loop. Only backbone atoms are shown; the side chains would fill the loop. This structure is residues 40 to 54 from cytochrome c. [Courtesy of George Rose, The Johns Hopkins University.]

necked-in shape of the Greek uppercase letter omega (Fig. 6-20), are compact globular entities because their side chains tend to fill in their internal cavities. Since Ω loops are almost invariably located on the protein surface, they may have important roles in biological recognition processes.

2 Tertiary Structure

The tertiary structure of a protein describes the folding of its secondary structural elements and specifies the positions of each atom in the protein, including those of its side chains. The known protein structures have come to light through **X-ray crystallographic** and **nuclear magnetic resonance (NMR)** studies. The atomic coordinates of most of these structures are deposited in a database known as the Protein Data Bank (PDB). These data are readily available via the Internet (<http://www.rcsb.org>), which allows the tertiary structures of a variety of proteins to be analyzed and compared. The common features of protein tertiary structures reveal much about the biological functions of proteins and their evolutionary origins.

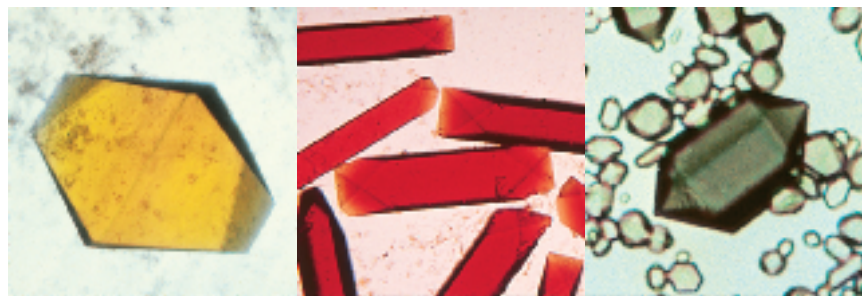
A Determining Protein Structures

X-Ray crystallography is a technique that directly images molecules. X-Rays must be used to do so because, according to optical principles, the uncertainty in locating an object is approximately equal to the wavelength of the radiation used to observe it (covalent bond distances and the wavelengths of the X-rays used in structural studies are both ~ 1.5 Å; individual molecules cannot be seen in a light microscope because visible light has a minimum wavelength of 4000 Å). There is, however, no such thing as an X-ray microscope because there are no X-ray lenses. Rather, a crystal of the molecule to be imaged (e.g., Fig. 6-21) is exposed to a collimated beam of X-rays and the resulting **diffraction pattern**, which arises from the regularly repeating positions of atoms in the crystal, is recorded by a radiation counter



(a) (b) (c)

Figure 6-21 Protein crystals. (a) Azurin from *Pseudomonas aeruginosa*, (b) flavodoxin from *Desulfovibrio vulgaris*, (c) rubredoxin from *Clostridium pasteurianum*, (d) azidomet myohemerythrin from the marine worm *Siphonosoma funafuti*, (e) lamprey hemoglobin, and (f) bacteriochlorophyll *a* protein from *Prosthecochloris aestuarii*. These crystals are colored because the proteins contain light-absorbing groups; proteins are colorless in the absence of such groups. [Parts a–c courtesy of Larry Siecker, University of Washington; parts d and e courtesy of Wayne Hendrikson, Columbia University; and part f courtesy of John Olsen, Brookhaven National Laboratories, and Brian Matthews, University of Oregon.]



(d) (e) (f)

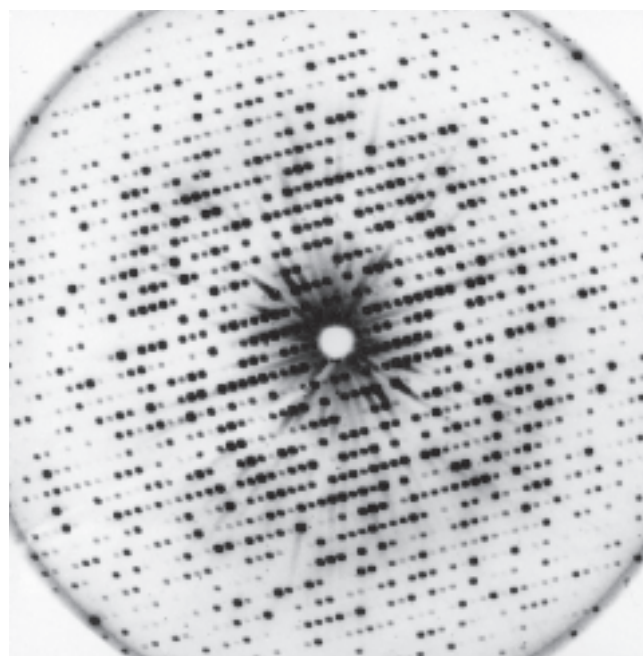


Figure 6-22 An X-ray diffraction photograph of a crystal of sperm whale myoglobin. The intensity of each diffraction maximum (the darkness of each spot) is a function of the crystal's electron density. [Courtesy of John Kendrew, Cambridge University, U.K.]

or, now infrequently, on photographic film (Fig. 6-22). The X-rays used in structural studies are produced by laboratory X-ray generators or, increasingly often, by a **synchrotron**, a type of particle accelerator that produces X-rays of far greater intensity. The intensities of the diffraction maxima (darkness of the spots on a film) are then used to construct mathematically the three-dimensional image of the crystal structure through methods that are beyond the scope of this text. In what follows, we discuss some of the special problems associated with interpreting the X-ray crystal structures of proteins.

X-Rays interact almost exclusively with the electrons in matter, not the nuclei. An X-ray structure is therefore an image of the **electron density** of the object under study. Such **electron density maps** are usually presented with the aid of a graphics computer as one or more sets of contours (e.g., Fig. 6-23) in which a contour represents a specific level of electron density

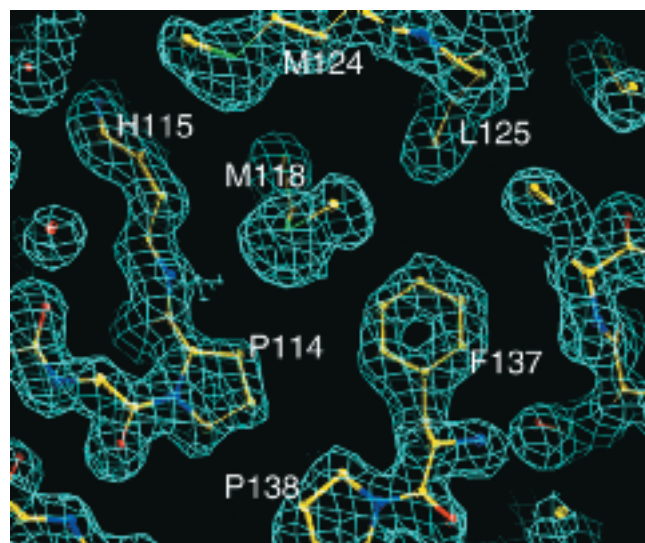


Figure 6-23 A thin section through a 1.5-Å-resolution electron density map of a protein that is contoured in three dimensions. Only a single contour level (cyan) is shown, together with a ball-and-stick model of the corresponding polypeptide segments colored according to atom type with C yellow, N blue, and O red. A water molecule is represented by a red sphere. [Courtesy of Xinhua Ji, NCI-Frederick Cancer Research and Development Center, Frederick, Maryland.]

in the same way that a contour on a topographic map indicates locations that have a particular altitude.

Most Protein Crystal Structures Exhibit Less Than Atomic Resolution. The molecules in protein crystals, as in other crystalline substances, are arranged in regularly repeating three-dimensional lattices. Protein crystals, however, differ from those of most small organic and inorganic molecules in being highly hydrated; they are typically 40 to 60% water by volume. The aqueous solvent of crystallization is necessary for the structural integrity of the protein crystals, as J. D. Bernal and Dorothy Crowfoot Hodgkin first noted in 1934 when they carried out the original X-ray studies of protein crystals. This is because water is required for the structural integrity of native proteins themselves (Section 6-4).

The large solvent content of protein crystals gives them a soft, jellylike consistency so that their molecules usually lack the rigid order characteristic of crystals of small molecules such as NaCl or glycine. The molecules in a protein crystal are typically disordered by more than an angstrom so that the corresponding electron density map lacks information concerning structural details of smaller size. The crystal is therefore said to have a resolution limit of that size. Protein crystals typically have resolution limits in the range 1.5 to 3.0 Å, although some are better ordered (have higher resolution, that is, a lesser resolution limit) and many are less ordered (have lower resolution).

Since an electron density map of a protein must be interpreted in terms of its atomic positions, the accuracy and even the feasibility of a crystal structure analysis depends on the crystal's resolution limit. Indeed, the inability to obtain crystals of sufficiently high resolution is a major limiting factor in determining the X-ray crystal structure of a protein or other macromolecule. Figure 6-24 indicates how the quality (degree of focus) of an electron density map varies with its resolution limit. At 6-Å resolution, the presence of a molecule the size of diketopiperazine is difficult to discern. At 2.0-Å resolution, its individual atoms cannot yet be distinguished, although its molecular shape has become reasonably evident. At 1.5-Å resolution, which roughly corresponds to a bond distance, individual atoms become partially resolved. At 1.1-Å resolution, atoms are clearly visible.

Most protein crystal structures are too poorly resolved for their electron density maps to reveal clearly the positions of individual atoms (e.g., Fig. 6-24). Nevertheless, the distinctive shape of the polypeptide backbone usually permits it to be traced, which, in turn, allows the positions and orientations of its side chains to be deduced (e.g., Fig. 6-23). Yet side chains of comparable size and shape, such as those of Leu, Ile, Thr, and Val, cannot always be differentiated with a reasonable degree of confidence (hydrogen atoms, having but one electron, are only visible in the few macromolecular X-ray structures with resolution limits less than ~1.2 Å), so that a protein structure cannot be elucidated from its electron density map alone. Rather, the primary structure of the protein must be known, thereby permitting the sequence of amino acid residues to be fitted to its electron density map. Mathematical refinement can then reduce the uncertainty in the crystal structure's atomic positions to as little as 0.1 Å (in contrast, positional errors in the most accurately determined small molecule X-ray structures are as little as 0.001 Å).

Most Crystalline Proteins Maintain Their Native Conformations. What is the relationship between the structure of a protein in a crystal and that in solu-

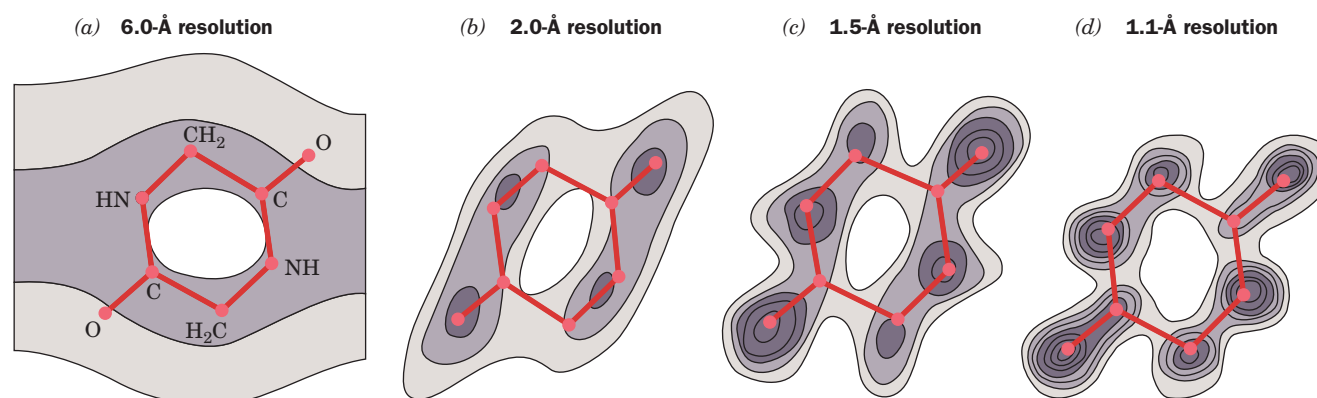


Figure 6-24 Section through the electron density map of diketopiperazine calculated at the indicated resolution levels. Hydrogen atoms are not visible in these maps because of their low electron density. [After Hodgkin, D.C., *Nature* **188**, 445 (1960).]

tion, where globular proteins normally function? Several lines of evidence indicate that *crystalline proteins assume very nearly the same structures that they have in solution*:

1. A protein molecule in a crystal is essentially in solution because it is bathed by solvent of crystallization over all of its surface except for the few, generally small patches that contact neighboring protein molecules. In fact, the 40 to 60% water content of typical protein crystals is similar to that of many cells (e.g., see Fig. 1-6).
2. A protein may crystallize in one of several forms or “habits,” depending on crystallization conditions, that differ in how the protein molecules are arranged in space relative to each other. In the numerous cases in which different crystal forms of the same protein have been independently analyzed, the molecules have virtually identical conformations. Similarly, in the several cases for which both the X-ray crystal structure and the solution NMR structure of the same protein have been determined, the two structures are, for the most part, identical to within experimental error (see below). Evidently, crystal packing forces do not greatly perturb the structures of protein molecules.
3. The most compelling evidence that crystalline proteins have biologically relevant structures is the observation that many enzymes are catalytically active in the crystalline state. The catalytic activity of an enzyme, as we shall see, is very sensitive to the relative orientations of the groups involved in binding and catalysis (Chapter 11). Active crystalline enzymes must therefore have conformations that closely resemble their solution conformations.

Protein Structure Determination by NMR. The determination of the three-dimensional structures of small globular proteins in aqueous solution has become possible, since the mid-1980s, through the development of **two-dimensional (2D) NMR spectroscopy** (and, more recently, of 3D and 4D techniques), in large part by Kurt Wüthrich. Such NMR measurements, whose description is beyond the scope of this text, yield the interatomic distances between specific protons that are $<5 \text{ \AA}$ apart in a protein of known sequence. The interproton distances may be either through space, as determined by nuclear Overhauser effect spectroscopy (NOESY,

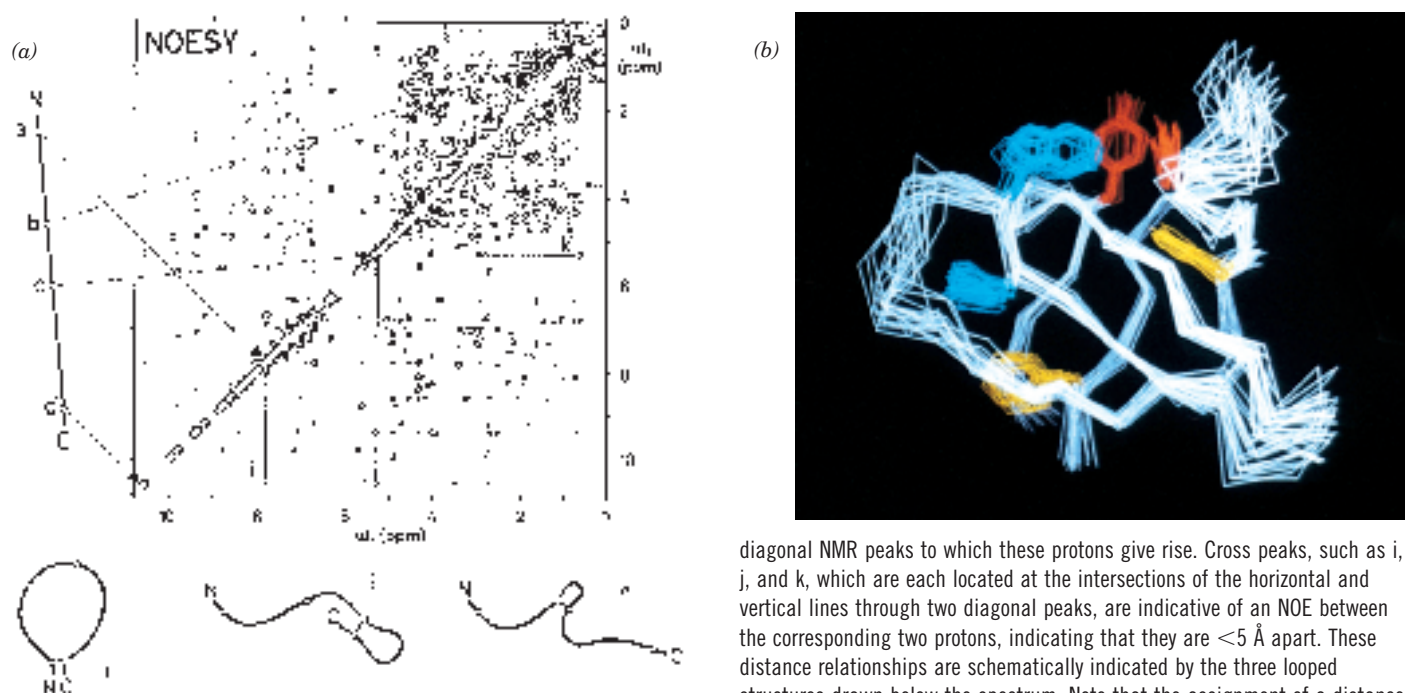


Figure 6-25 2D proton NMR structures of proteins. (a) A NOESY spectrum of a protein presented as a contour plot with two frequency axes, ω_1 and ω_2 . The conventional 1D-NMR spectrum of the protein, which occurs along the diagonal of the plot ($\omega_1 = \omega_2$), is too crowded with peaks to be directly interpretable (even a small protein has hundreds of protons). The off-diagonal peaks, the so-called cross peaks, each arise from the interaction of two protons that are $<5 \text{ \AA}$ apart in space and whose 1D-NMR peaks are located where the horizontal and vertical lines through the cross peak intersect the diagonal [a **nuclear Overhauser effect (NOE)**]. For example, the line to the left of the spectrum represents the extended polypeptide chain with its N- and C-terminal ends identified by the letters N and C and with the positions of four protons, a to d, represented by small circles. The dashed arrows indicate the

diagonal NMR peaks to which these protons give rise. Cross peaks, such as i, j, and k, which are each located at the intersections of the horizontal and vertical lines through two diagonal peaks, are indicative of an NOE between the corresponding two protons, indicating that they are $<5 \text{ \AA}$ apart. These distance relationships are schematically indicated by the three looped structures drawn below the spectrum. Note that the assignment of a distance relationship between two protons in a polypeptide requires that the NMR peaks to which they give rise and their positions in the polypeptide be known, which requires that the polypeptide's amino acid sequence has been previously determined. [After Wüthrich, K., *Science* **243**, 45 (1989).] (b) The NMR structure of a 64-residue polypeptide comprising the **Src protein SH3 domain** (Section 21-3D). The drawing represents 20 superimposed structures that are consistent with the 2D- and 3D-NMR spectra of the protein (each calculated from a different, randomly generated starting structure). The polypeptide backbone, as represented by its connected C_α atoms, is white and its Phe, Tyr, and Trp side chains are yellow, red, and blue, respectively. It can be seen that the polypeptide backbone folds into two 3-stranded antiparallel β sheets that form a sandwich. [Courtesy of Stuart Schreiber, Harvard University.]

Fig. 6-25a), or through bonds, as determined by correlated spectroscopy (COSY). These distances, together with known geometric constraints such as covalent bond distances and angles, group planarity, chirality, and van der Waals radii, are used to compute the protein's three-dimensional structure. However, since interproton distance measurements are imprecise, they are insufficient to imply a unique structure. Rather, they are consistent with an ensemble of closely related structures. Consequently, an NMR structure of a protein (or any other macromolecule with a well-defined structure) is often presented as a representative sample of structures that are consistent with the constraints (e.g., Figure 6-25b). The "tightness" of a bundle of such structures is indicative both of the accuracy with which the structure is known, which in the most favorable cases is roughly comparable to that of an X-ray crystal structure with a resolution of 2 to 2.5 \AA , and of the conformational fluctuations that the protein undergoes (Section 6-4B). Although present NMR methods are limited to determining the structures of macromolecules with molecular masses no greater than $\sim 40 \text{ kD}$, recent advances in NMR technology suggest that this limit may soon increase to $\sim 100 \text{ kD}$ or more.

In most of the several cases in which both the NMR and X-ray crystal structures of a particular protein have been determined, the two structures

are in good agreement. There are, however, a few instances in which there are real differences between the corresponding X-ray and NMR structures. These, for the most part, involve surface residues that, in the crystal, participate in intermolecular contacts and are thereby perturbed from their solution conformations. NMR methods, besides providing mutual cross-checks with X-ray techniques, can determine the structures of proteins and other macromolecules that fail to crystallize. Moreover, since NMR can probe motions over time scales spanning 10 orders of magnitude, it can also be used to study protein folding and dynamics (Section 6-4).

Visualizing Proteins. The huge number of atoms in proteins makes it difficult to visualize them using the same sorts of models employed for small organic molecules. Ball-and-stick representations showing all or most atoms in a protein (as in Figs. 6-7 and 6-10) are exceedingly cluttered, and space-filling models (as in Figs. 6-8 and 6-11) obscure the internal details of the protein. Accordingly, computer-generated or artistic renditions (e.g., Fig. 6-12) are often more useful for representing protein structures. The course of the polypeptide chain can be followed by tracing the positions of its C_α atoms or by representing helices as helical ribbons or cylinders and β sheets as sets of flat arrows pointing from the N- to the C-termini.

B Side Chain Location and Polarity

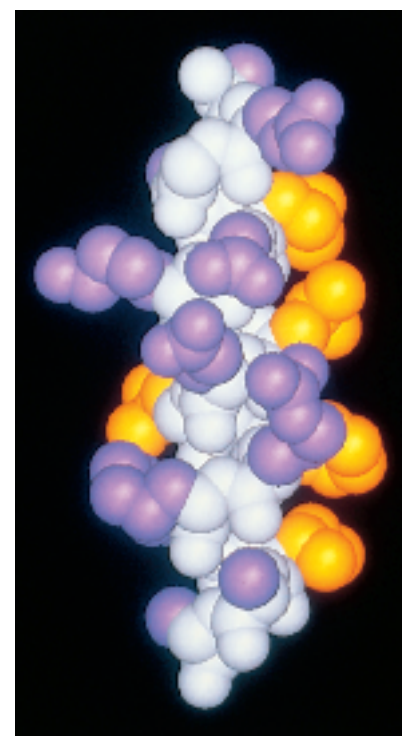
In the years since Kendrew solved the structure of myoglobin, nearly 30,000 protein structures have been reported. No two are exactly alike, but they exhibit remarkable consistencies.

Side Chain Location Varies with Polarity. The primary structures of globular proteins generally lack the repeating sequences that support the regular conformations seen in fibrous proteins. However, *the amino acid side chains in globular proteins are spatially distributed according to their polarities:*

1. The nonpolar residues Val, Leu, Ile, Met, and Phe occur mostly in the interior of a protein, out of contact with the aqueous solvent. The hydrophobic effects that promote this distribution are largely responsible for the three-dimensional structure of native proteins.
2. The charged polar residues Arg, His, Lys, Asp, and Glu are usually located on the surface of a protein in contact with the aqueous solvent. This is because immersing an ion in the virtually anhydrous interior of a protein is energetically unfavorable.
3. The uncharged polar groups Ser, Thr, Asn, Gln, and Tyr are usually on the protein surface but also occur in the interior of the molecule. When buried in the protein, these residues are almost always hydrogen bonded to other groups; in a sense, the formation of a hydrogen bond “neutralizes” their polarity. This is also the case with the polypeptide backbone.

These general principles of side chain distribution are evident in individual elements of secondary structure (Fig. 6-26) as well as in whole proteins

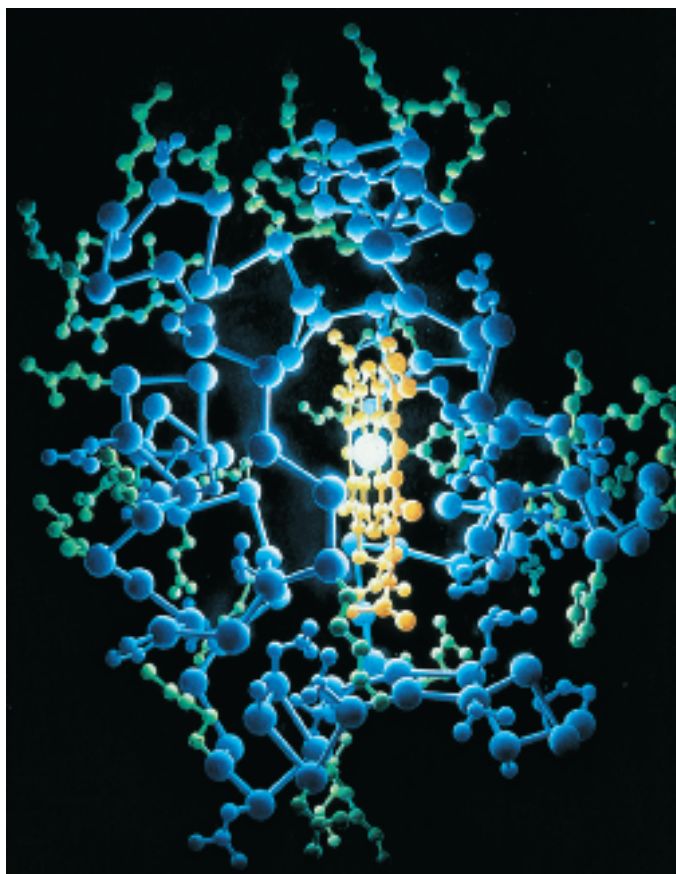
Figure 6-26 Side chain locations in an α helix and a β sheet. In these space-filling models, the main chain is white, nonpolar side chains are yellow or brown, and polar side chains are purple. (a) An α helix from sperm whale myoglobin. Note that the nonpolar residues are primarily on one side of the helix. (b) An antiparallel β sheet from concanavalin A (*side view*). The protein interior is to the right and the exterior is to the left.



(a)

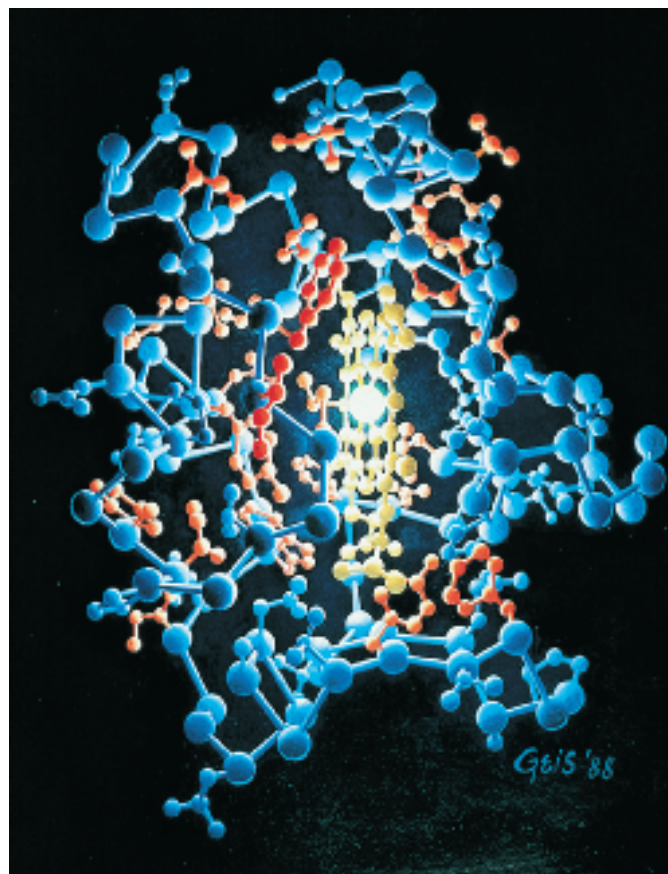


(b)




(a)

Figure 6-27 Side chain distribution in horse heart cytochrome *c*. In these paintings, based on an X-ray structure determined by Richard Dickerson, the protein is illuminated by its single iron atom centered in a heme group. Hydrogen atoms are not shown. In (a) the hydrophilic side chains are green,



(b)

and in (b) the hydrophobic side chains are orange. [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.]  See Kinemage Exercise 5.

See Guided Exploration 8

Secondary structures in proteins

(Fig. 6-27). Polar side chains tend to extend toward—and thereby help form—the protein's surface, whereas nonpolar side chains largely extend toward—and thereby occupy—its interior.

Most proteins are quite compact, with their interior atoms packed together even more efficiently than the atoms in a crystal of small organic molecules. Nevertheless, the atoms of protein side chains almost invariably have low-energy arrangements. Evidently, interior side chains adopt relaxed conformations despite the profusion of intramolecular interactions. Closely packed protein interiors generally exclude water. When water molecules are present, they often occupy specific positions where they can form hydrogen bonds, sometimes acting as a bridge between two hydrogen-bonding protein groups.

C Supersecondary Structures and Domains

The major types of secondary structural elements, α helices and β sheets, occur in globular proteins in varying proportions and combinations. Some proteins, such as *E. coli* **cytochrome *b*₅₆₂** (Fig. 6-28a), consist only of α helices

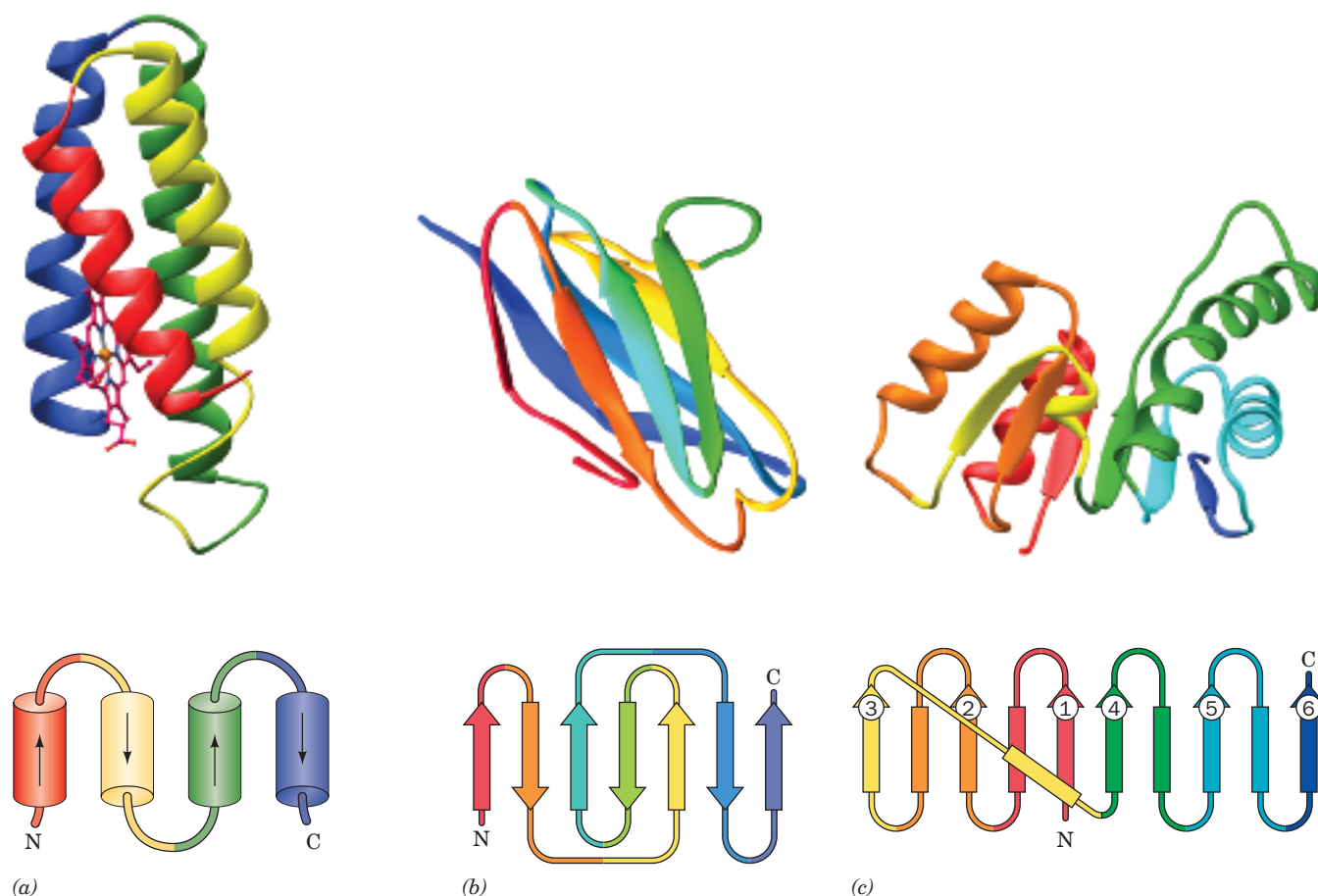


Figure 6-28 A selection of protein structures. The proteins are represented by their peptide backbones, drawn in ribbon form, with β strands shown as flat arrows pointing from N- to C-terminus. The polypeptide chain is colored, from N- to C-terminus, in rainbow order, from red to blue. Below each drawing is the corresponding topological diagram indicating the connectivity of its helices (represented by cylinders or rectangles) and β strands (represented by flat arrows) (a) The X-ray structure of the 106-residue *E. coli* **cytochrome b_{562}** , which forms an up-down-up-down 4-helix bundle. Its bound heme group is shown in ball-and-stick form with C magenta, N blue, O red, and Fe orange. (b) The X-ray structure of the N-terminal domain of the 103-residue human immunoglobulin fragment **Fab New** showing its immunoglobulin fold. The polypeptide chain is folded into a sandwich of 3- and 4-stranded antiparallel β sheets. (c) The X-ray structure of the 163-residue N-terminal domain of dogfish lactate dehydrogenase. It contains a 6-stranded parallel β sheet in which the crossovers between β strands all contain an α helix that forms a right-handed helical turn with its flanking β strands. [Based on X-ray structures by (a) F. Scott Matthews, Washington University School of Medicine; (b) Roberto Poljak, The Johns Hopkins School of Medicine; and (c) Michael Rossmann, Purdue University. PDBids (a) 256B, (b) 7FAB, and (c) 6LDH.]

spanned by short connecting links. Others, such as the **immunoglobulin fold** (Fig. 6-28b), have a large proportion of β sheets and are devoid of α helices. Most proteins, such as dogfish **lactate dehydrogenase** (Fig. 6-28c) and carboxypeptidase A (Fig. 6-12), have significant amounts of both types of secondary structure (on average, $\sim 31\%$ α helix and $\sim 28\%$ β sheet).

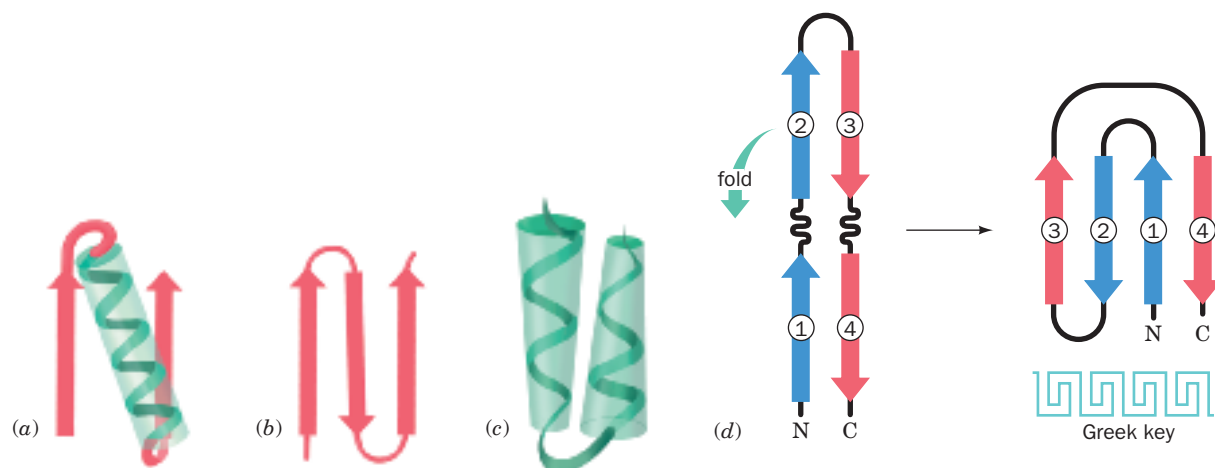


Figure 6-29 Schematic diagrams of supersecondary structures. (a) a $\beta\alpha\beta$ motif, (b) a β hairpin motif, (c) an $\alpha\alpha$ motif, and (d) a Greek key motif, showing how it is constructed from a folded-over β hairpin.

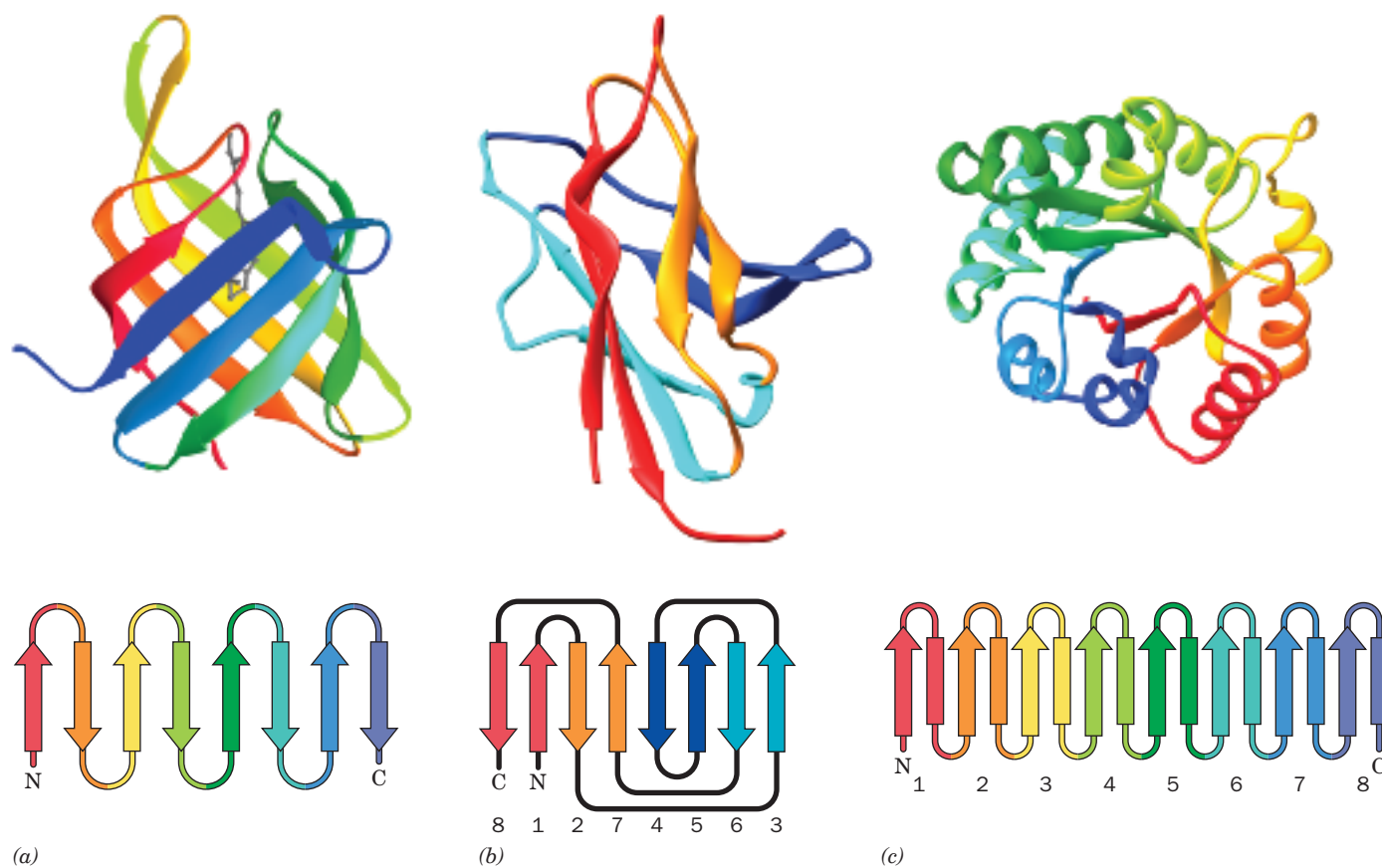


Figure 6-30 X-Ray structures of β barrels. Each polypeptide is drawn and colored and accompanied by its corresponding topological diagram as is described in the legend to Fig. 6-28. (a) Human **retinol binding protein** showing its 8-stranded up-and-down β barrel (residues 1–142 of this 182-residue protein). Note that each β strand is linked via a short loop to its clockwise-adjacent strand as seen from the top. The protein's bound retinol molecule is represented by a gray ball-and-stick model. (b) **Peptide- N^4 -(N -acetyl- β -D-glucosaminy)asparagine amidase F** from *Flavobacterium meningosepticum* (residues 1–140 of this 340-residue enzyme). Note how its 8-stranded β barrel is formed by rolling up a 4-segment β hairpin. Here the two β strands in each segment of the β hairpin are colored alike with strands 1 and 8 (the N- and C-terminal strands) red, strands

2 and 7 orange, strands 3 and 6 cyan, and strands 4 and 5 blue. This motif, which is known as a **jelly roll** or **Swiss roll barrel**, is so named because of its topological resemblance to these rolled-up pastries. (c) Chicken muscle **triose phosphate isomerase (TIM; 247 residues)** forms a so-called α/β barrel in which 8 pairs of alternating β strands and α helices roll up to form an inner barrel of 8 parallel β strands surrounded by an outer barrel of 8 parallel α helices. The protein is viewed approximately along the axis of the α/β barrel. Note that the α/β barrel is essentially a series of linked $\beta\alpha\beta$ motifs. [Based on X-ray structures by (a) T. Alwyn Jones, Biomedical Center, Uppsala, Sweden; (b) Patrick Van Roey, New York State Department of Health, Albany, New York; and (c) David Phillips, Oxford University, Oxford, U.K. PDBids (a) 1RBP, (b) 1PNG, and (c) 1TIM.]

Certain groupings of secondary structural elements, called **supersecondary structures** or **motifs**, occur in many unrelated globular proteins:

1. The most common form of supersecondary structure is the **$\beta\alpha\beta$ motif**, in which an α helix connects two parallel strands of a β sheet (Fig. 6-29a).
2. Another common supersecondary structure, the **β hairpin** motif, consists of antiparallel strands connected by relatively tight reverse turns (Fig. 6-29b).
3. In an **$\alpha\alpha$ motif**, two successive antiparallel α helices pack against each other with their axes inclined. This permits energetically favorable intermeshing of their contacting side chains (Fig. 6-29c). Similar associations stabilize the coiled coil conformation of α keratin, although its helices are parallel rather than antiparallel (Section 6-1C).
4. In the **Greek key motif** (Fig. 6-29d; named after an ornamental design commonly used in ancient Greece; see inset), a β hairpin is folded over to form a 4-stranded antiparallel β sheet. Of the 10 possible ways of connecting the strands of a 4-stranded antiparallel β sheet, the two that form Greek key motifs are, by far, the most common in proteins of known structure.
5. Extended β sheets often roll up to form **β barrels**. Three different types of β barrels are shown in Fig. 6-30.

Motifs may have functional as well as structural significance. For example, Michael Rossmann has shown that a $\beta\alpha\beta\alpha\beta$ unit, in which the β strands form a parallel sheet with α helical connections, often acts as a nucleotide-binding site. In most proteins that bind dinucleotides [such as nicotinamide adenine dinucleotide (NAD^+); Fig. 11-3], two such $\beta\alpha\beta\alpha\beta$ units combine to form a motif known as a **dinucleotide-binding fold**, or **Rossmann fold**. Lactate dehydrogenase (Fig. 6-28c) is an example of such a protein.

Large Polypeptides Form Domains. Polypeptide chains containing more than ~ 200 residues usually fold into two or more globular clusters known as **domains**, which give these proteins a bi- or multilobal appearance. Most domains consist of 100 to 200 amino acid residues and have an average diameter of ~ 25 Å. Each subunit of the enzyme **glyceraldehyde-3-phosphate dehydrogenase**, for example, has two distinct domains (Fig. 6-31). A polypeptide chain wanders back and forth within a domain, but neighboring domains are usually connected by only one or two polypeptide segments. *Consequently, many domains are structurally independent units that have the characteristics of small globular proteins.* Nevertheless, the domain structure of a protein is not necessarily obvious since its domains may make such extensive contacts with each other that the protein appears to be a single globular entity.

An inspection of the various protein structures diagrammed in this chapter reveals that domains consist of two or more layers of secondary structural elements. The reason for this is clear: At least two such layers are required to seal off a domain's hydrophobic core from its aqueous environment.

Domains often have a specific function such as the binding of a small molecule. In Fig. 6-31, for example, NAD^+ binds to the first domain of glyceraldehyde-3-phosphate dehydrogenase (note its dinucleotide-binding

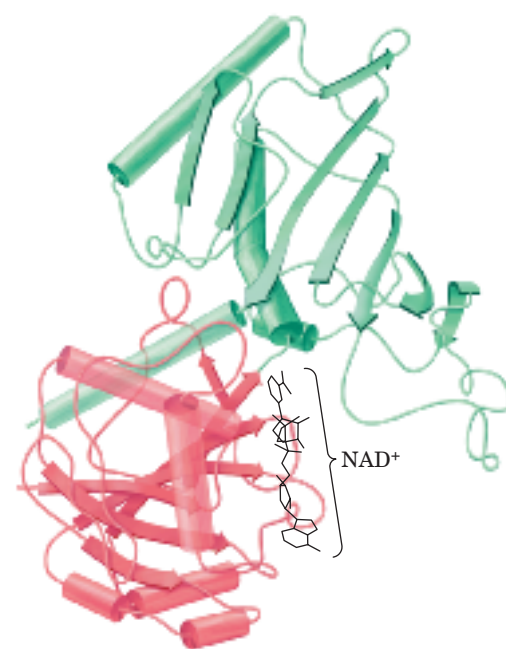


Figure 6-31 The two-domain protein glyceraldehyde-3-phosphate dehydrogenase. The first domain (red) binds NAD^+ (black), and the second domain (green) binds glyceraldehyde-3-phosphate (not shown). [After Biesecker, G., Harris, J.I., Thierry, J.C., Walker, J.E., and Wonacott, A., *Nature* 266, 331 (1977).] See the Interactive Exercises.

fold). In multidomain proteins, binding sites often occupy the clefts between domains; that is, small molecules are bound by groups from two domains. In such cases, the relatively pliant covalent connection between the domains allows flexible interactions between the protein and the small molecule.

D Protein Families

The many thousands of known protein structures, comprising an even greater number of separate domains, can be grouped into families by examining the overall paths followed by their polypeptide chains. When folding patterns are compared without regard to the amino acid sequence or the presence of surface loops, the number of unique structural domains drops to only a few hundred. (Although not all protein structures are known, estimates place an upper limit of about 1000 on the total number of unique protein domains in nature.) Surprisingly, a few dozen folding patterns account for about half of all known protein structures.

There are several possible reasons for the limited number of known domain structures. The numbers may reflect database bias; that is, the collection of known protein structures may not be a representative sample of all protein structures. However, the rapidly increasing number of proteins whose structures have been determined makes this possibility less and less plausible. More likely, the common protein structures may be evolutionary sinks—domains that arose and persisted because of their ability (1) to form stable folding patterns; (2) to tolerate amino acid deletions, substitutions, and insertions, thereby making them more likely to survive evolutionary changes; and/or (3) to support essential biological functions.

Polypeptides with similar sequences tend to adopt similar backbone conformations. This is certainly true for evolutionarily related proteins that carry out similar functions. For example, the cytochromes *c* of different species are highly conserved proteins with closely similar sequences (see Table 5-5) and three-dimensional structures.

Cytochrome *c* occurs only in eukaryotes, but prokaryotes contain proteins, known as **c-type cytochromes**, which perform the same general function (that of an electron carrier). The *c*-type cytochromes from different species exhibit only low degrees of sequence similarity to each other and to eukaryotic cytochromes *c*. Yet their X-ray structures are clearly similar, particularly in polypeptide chain folding and side chain packing in the protein interior (Fig. 6-32). The major structural differences among *c*-type

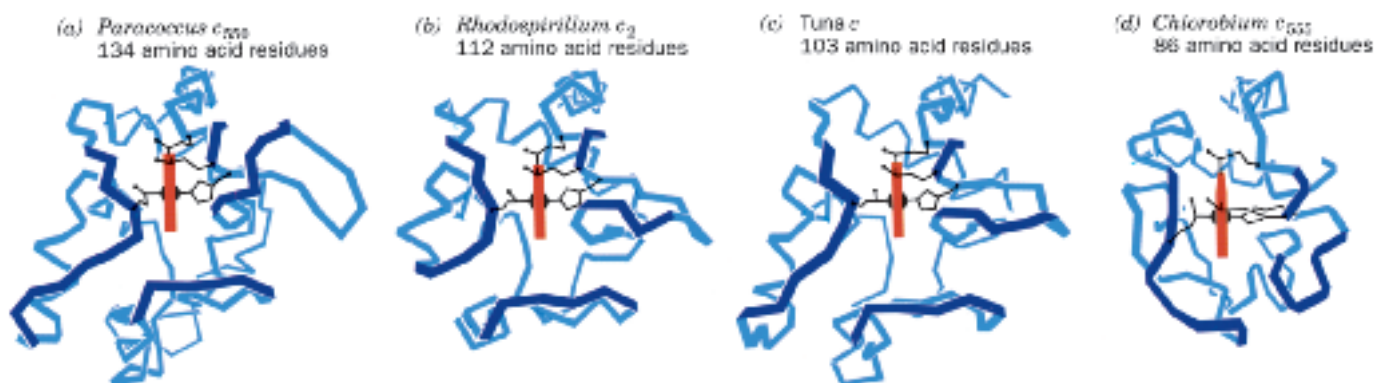


Figure 6-32 Three-dimensional structures of *c*-type cytochromes. The polypeptide backbones (*blue*) are shown in analogous orientations such that their heme groups (*red*) are viewed edge-on. The Cys, Met, and His side chains that covalently link the heme to the protein are also shown. (a) Cytochrome *c*₅₅₀ from *Paracoccus denitrificans* (134 residues), (b) cytochrome *c*₂ from

Rhodospirillum rubrum (112 residues), (c) cytochrome *c* from tuna (103 residues), and (d) cytochrome *c*₅₅₅ from *Chlorobium limicola* (86 residues). [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.]

See Kinemage Exercise 5.

cytochromes lie in the various polypeptide loops on their surfaces. The sequences of the *c*-type cytochromes have diverged so far from one another that, in the absence of their X-ray structures, they can be properly aligned only through the use of mathematically sophisticated computer programs. Thus, *it appears that the essential structural and functional elements of proteins, rather than their amino acid residues, are conserved during evolution.*

Structural similarities in proteins with only distantly related functions are commonly observed. For example, many NAD^+ -binding enzymes that participate in widely different metabolic pathways contain similar dinucleotide-binding folds (e.g., Fig. 6-31) coupled to diverse domains that carry out specific enzymatic reactions.

3 Quaternary Structure and Symmetry

Most proteins, particularly those with molecular masses >100 kD, consist of more than one polypeptide chain. These polypeptide subunits associate with a specific geometry. The spatial arrangement of these subunits is known as a protein's quaternary structure.

There are several reasons why multisubunit proteins are so common. In large assemblies of proteins, such as collagen fibrils, the advantages of subunit construction over the synthesis of one huge polypeptide chain are analogous to those of using prefabricated components in constructing a building: Defects can be repaired by simply replacing the flawed subunit; the site of subunit manufacture can be different from the site of assembly into the final product; and the only genetic information necessary to specify the entire edifice is the information specifying its few different self-assembling subunits. In the case of enzymes, increasing a protein's size tends to better fix the three-dimensional positions of its reacting groups. *Increasing the size of an enzyme through the association of identical subunits is more efficient than increasing the length of its polypeptide chain since each subunit has an active site. More importantly, the subunit construction of many enzymes provides the structural basis for the regulation of their activities* (Sections 7-3B and 12-3).

Subunits Usually Associate Noncovalently. A multisubunit protein may consist of identical or nonidentical polypeptide chains. Hemoglobin, for example, has the subunit composition $\alpha_2\beta_2$ (Fig. 6-33). Proteins with more than one subunit are called **oligomers**, and their identical units are called **protomers**. A protomer may therefore consist of one polypeptide chain or several unlike polypeptide chains. In this sense, hemoglobin is a dimer of $\alpha\beta$ protomers.

The contact regions between subunits resemble the interior of a single-subunit protein: They contain closely packed nonpolar side chains, hydrogen bonds involving the polypeptide backbones and their side chains, and, in some cases, interchain disulfide bonds. However, the subunit interfaces of proteins that dissociate *in vivo* have lesser hydrophobicities than do permanent interfaces.

Subunits Are Symmetrically Arranged. In the vast majority of oligomeric proteins, the protomers are symmetrically arranged; that is, each protomer occupies a geometrically equivalent position in the oligomer. Proteins cannot have inversion or mirror symmetry, however, because bringing the protomers into coincidence would require converting chiral L residues to D residues. Thus, *proteins can have only rotational symmetry.*

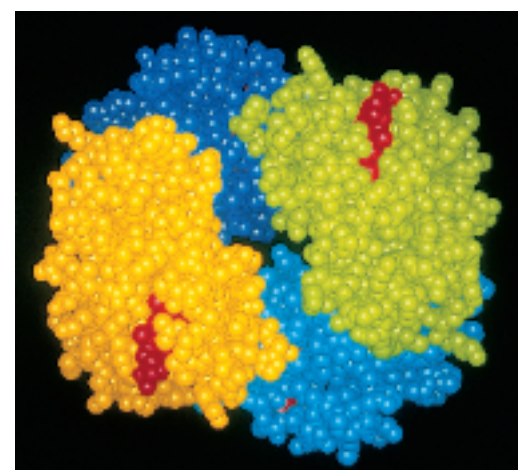


Figure 6-33 Quaternary structure of hemoglobin. In this space-filling model, the α_1 , α_2 , β_1 , and β_2 subunits are colored yellow, green, cyan, and blue, respectively. Heme groups are red. [Based on an X-ray structure by Max Perutz, MRC Laboratory of Molecular Biology, U.K. PDBid 2DHB.]

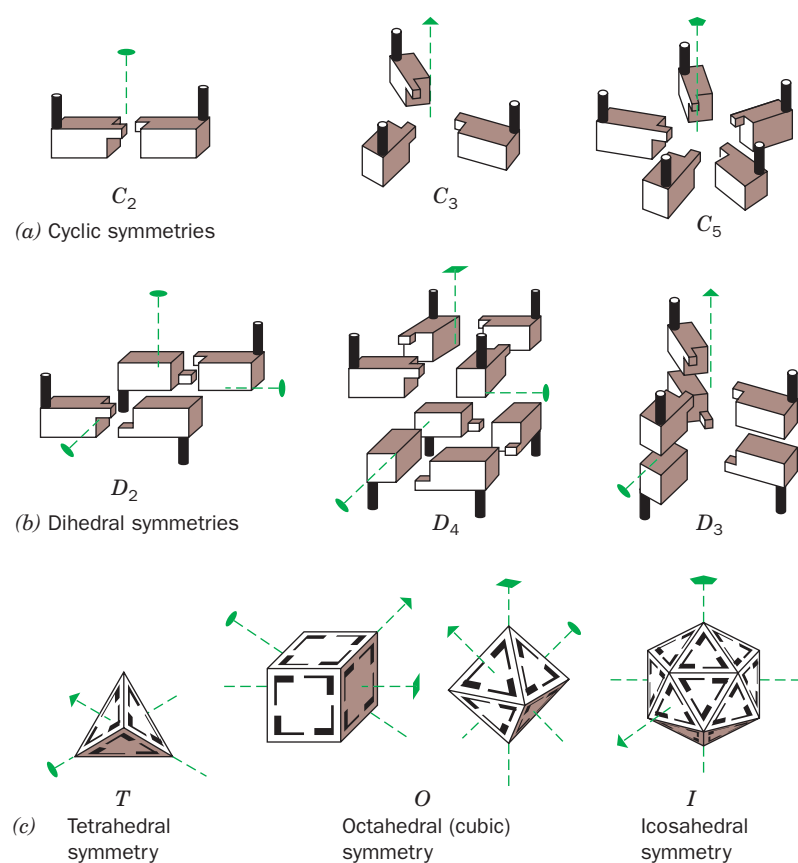


Figure 6-34 Some symmetries of oligomeric proteins. The oval, the triangle, the square, and the pentagon at the ends of the dashed green lines indicate, respectively, the unique twofold, threefold, fourfold, and fivefold rotational axes of the objects shown. (a) Assemblies with cyclic (C) symmetry. (b) Assemblies with dihedral (D) symmetry. In these objects, a 2-fold axis is perpendicular to another rotational axis. (c) Assemblies with the rotational symmetries of a tetrahedron (T), a cube or octahedron (O), and an icosahedron (I). [Illustration, Irving Geis. Image from the Irving Geis Collection/Howard Hughes Medical Institute. Rights owned by HHMI. Reproduction by permission only.] See the Animated Figures.

In the simplest type of rotational symmetry, **cyclic symmetry**, protomers are related by a single axis of rotation (Fig. 6-34a). Objects with two-, three-, or n -fold rotational axes are said to have C_2 , C_3 , or C_n symmetry, respectively. C_2 symmetry is the most common; higher cyclic symmetries are relatively rare.

Dihedral symmetry (D_n), a more complicated type of rotational symmetry, is generated when an n -fold rotation axis intersects a twofold rotation axis at right angles (Fig. 6-34b). An oligomer with D_n symmetry consists of $2n$ protomers. D_2 symmetry is the most common type of dihedral symmetry in proteins.

Other possible types of rotational symmetry are those of a tetrahedron, cube, and icosahedron (Fig. 6-34c). Some multienzyme complexes and spherical viruses are built on these geometric plans.

4 Protein Stability

Incredible as it may seem, thermodynamic measurements indicate that *native proteins are only marginally stable under physiological conditions*. The free energy required to denature them is $\sim 0.4 \text{ kJ} \cdot \text{mol}^{-1}$ per amino acid residue, so a fully folded 100-residue protein is only about $40 \text{ kJ} \cdot \text{mol}^{-1}$ more stable than its unfolded form (for comparison, the energy required to break a typical hydrogen bond is $\sim 20 \text{ kJ} \cdot \text{mol}^{-1}$). The various noncovalent influences on proteins—hydrophobic effects, electrostatic interactions, and hydrogen bonding—each have energies that may total thousands of kilojoules per mole over an entire protein molecule. Consequently, a pro-

tein structure is the result of a delicate balance among powerful counter-vailing forces. In this section, we discuss the forces that stabilize proteins and the processes by which proteins achieve their most stable folded state.

A Forces That Stabilize Protein Structure

Protein structures are governed primarily by hydrophobic effects and, to a lesser extent, by interactions between polar residues, and by other types of bonds.

The Hydrophobic Effect. *The hydrophobic effect, which causes nonpolar substances to minimize their contacts with water (Section 2-1C), is the major determinant of native protein structure.* The aggregation of nonpolar side chains in the interior of a protein is favored by the increase in entropy of the water molecules that would otherwise form ordered “cages” around the hydrophobic groups. The combined hydrophobic and hydrophilic tendencies of individual amino acid residues in proteins can be expressed as **hydropathies** (Table 6-2). The greater a side chain’s hydropathy, the more likely it is to occupy the interior of a protein and vice versa. Hydropathies are good predictors of which portions of a polypeptide chain are inside a protein, out of contact with the aqueous solvent, and which portions are outside (Fig. 6-35).

Site-directed mutagenesis experiments in which individual interior residues have been replaced by a number of others suggest that the factors that affect stability are, in order, the hydrophobicity of the substituted residue, its steric compatibility, and, last, the volume of its side chain.

Electrostatic Interactions. In the closely packed interiors of native proteins, van der Waals forces, which are relatively weak (Section 2-1A), are nevertheless an important stabilizing influence. This is because these forces act over only short distances and hence are lost when the protein is unfolded.

Perhaps surprisingly, *hydrogen bonds, which are central features of protein structures, make only minor contributions to protein stability.* This is because hydrogen-bonding groups in an unfolded protein form hydrogen bonds with water molecules. Thus the contribution of a hydrogen bond to the stability of a native protein is the small difference in hydrogen bonding free energies between the native and unfolded states (-2 to $8 \text{ kJ}\cdot\text{mol}^{-1}$ as determined by site-directed mutagenesis studies). Nevertheless, hydrogen bonds are important determinants of native protein structures, because if a protein folded in a way that prevented a hydrogen bond from forming, the stabilizing energy of that hydrogen bond would be lost. Hydrogen bonding therefore fine-tunes tertiary structure by “selecting” the unique native structure of a protein from among a relatively small number of hydrophobically stabilized conformations.

Table 6-2 Hydropathy Scale for Amino Acid Side Chains

Side Chain	Hydropathy
Ile	4.5
Val	4.2
Leu	3.8
Phe	2.8
Cys	2.5
Met	1.9
Ala	1.8
Gly	-0.4
Thr	-0.7
Ser	-0.8
Trp	-0.9
Tyr	-1.3
Pro	-1.6
His	-3.2
Glu	-3.5
Gln	-3.5
Asp	-3.5
Asn	-3.5
Lys	-3.9
Arg	-4.5

Source: Kyte, J. and Doolittle, R.F., *J. Mol. Biol.* **157**, 110 (1982).

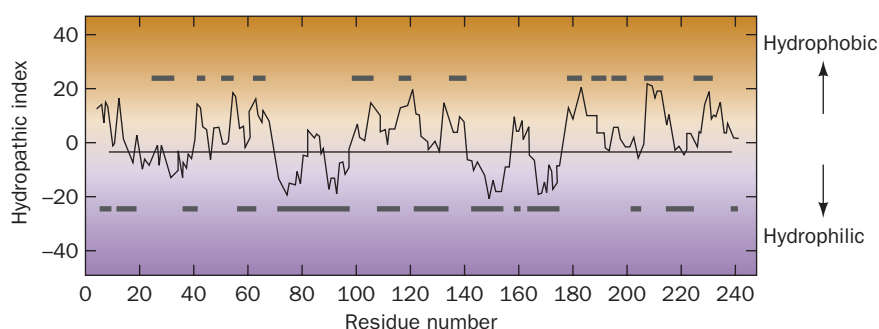


Figure 6-35 A hydropathic index plot for bovine chymotrypsinogen. The sum of the hydropathies of nine consecutive residues is plotted versus residue sequence number. A large positive hydropathic index indicates a hydrophobic region of the polypeptide, whereas a large negative value indicates a hydrophilic region. The upper bars denote the protein’s interior regions, as determined by X-ray crystallography, and the lower bars denote the protein’s exterior regions. [After Kyte, J. and Doolittle, R.F., *J. Mol. Biol.* **157**, 111 (1982).]

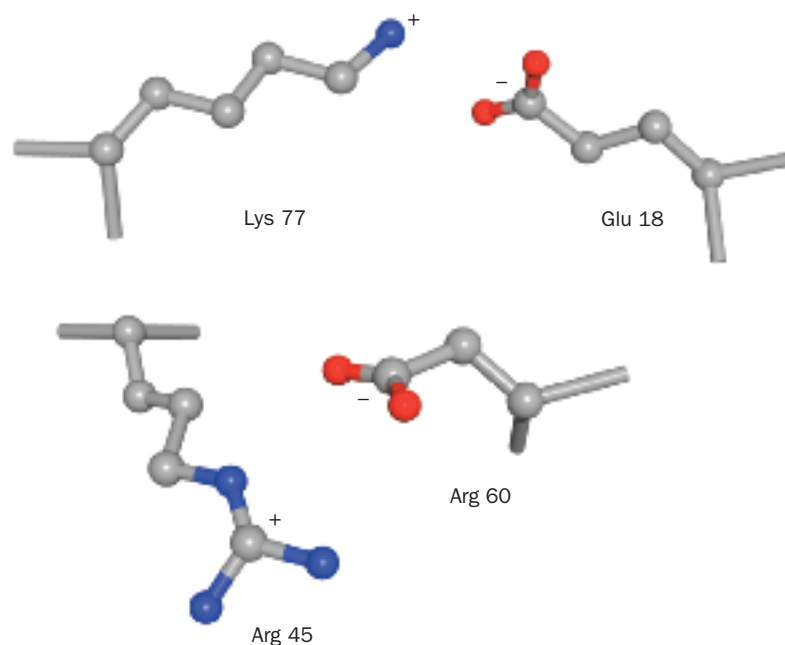


Figure 6-36 Examples of ion pairs in myoglobin. In each case, oppositely charged side chain groups from residues far apart in sequence closely approach each other through the formation of ion pairs.

The association of two ionic protein groups of opposite charge (e.g., Lys and Asp) is known as an **ion pair** or **salt bridge**. About 75% of the charged residues in proteins are members of ion pairs that are located mostly on the protein surface (Fig. 6-36). Despite the strong electrostatic attraction between the oppositely charged members of an ion pair, these interactions contribute little to the stability of a native protein. This is because the free energy of an ion pair's charge–charge interactions usually fails to compensate for the loss of entropy of the side chains and the loss of solvation free energy when the charged groups form an ion pair. This accounts for the observation that ion pairs are poorly conserved among homologous proteins.

Chemical Cross-Links. Disulfide bonds (Fig. 4-6) within and between polypeptide chains form as a protein folds to its native conformation. Some polypeptides whose Cys residues have been derivatized or mutagenically replaced to prevent disulfide bond formation can still assume their fully active conformations, suggesting that disulfide bonds are not essential stabilizing forces. They may, however, be important for “locking in” a particular backbone folding pattern as the protein proceeds from its fully extended state to its mature form.

Disulfide bonds are rare in intracellular proteins because the cytoplasm is a reducing environment. Most disulfide bonds occur in proteins that are secreted from the cell into the more oxidizing extracellular environment. The relatively hostile extracellular world (e.g., uncontrolled temperature and pH) apparently requires the additional structural constraints conferred by disulfide bonds.

Metal ions may also function to internally cross-link proteins. For example, at least ten motifs collectively known as **zinc fingers** have been described in nucleic acid-binding proteins. These structures contain about 25–60 residues arranged around one or two Zn^{2+} ions that are tetrahedrally coordinated by the side chains of Cys, His, and occasionally Asp or Glu (Fig. 6-37). The Zn^{2+} ion allows relatively short stretches of polypeptide chain to fold into stable units that can interact with nucleic acids. Zinc fingers are too small to be stable in the absence of Zn^{2+} . Zinc is ideally suited to its structural role in intracellular proteins: Its filled d electron shell

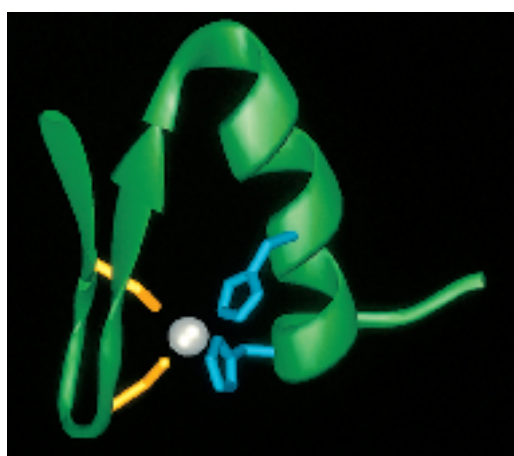


Figure 6-37 A zinc finger motif. This structure, from the DNA-binding protein Zif268, is known as a Cys_2His_2 zinc finger because the zinc atom (silver sphere) is coordinated by two Cys residues (yellow) and two His residues (cyan). [Based on an X-ray structure by Carl Pabo, MIT. PDBid 1ZAA.]

permits it to interact strongly with a variety of ligands (e.g., sulfur, nitrogen, or oxygen) from different amino acid residues. In addition, zinc has only one stable oxidation state (unlike, for example, copper and iron), so it does not undergo oxidation–reduction reactions in the cell.

B Protein Dynamics

The plethora of forces acting to stabilize proteins as well as the static way that their structures are usually portrayed may leave the false impression that proteins have fixed and rigid structures. In fact, *proteins are flexible and rapidly fluctuating molecules whose structural mobilities are functionally significant*. Groups ranging in size from individual side chains to entire domains or subunits may be displaced by up to several angstroms through random intramolecular movements or in response to a trigger such as the binding of a small molecule. Extended side chains, such as Lys, and the N- and C-termini of polypeptide chains are especially prone to wave around in solution because there are few forces holding them in place.

Theoretical calculations by Martin Karplus indicate that a protein's native structure probably consists of a large collection of rapidly interconverting conformations that have essentially equal stabilities (Fig. 6-38). Conformational flexibility, or **breathing**, with structural displacement of up to ~ 2 Å, allows small molecules to diffuse in and out of the interior of certain proteins.

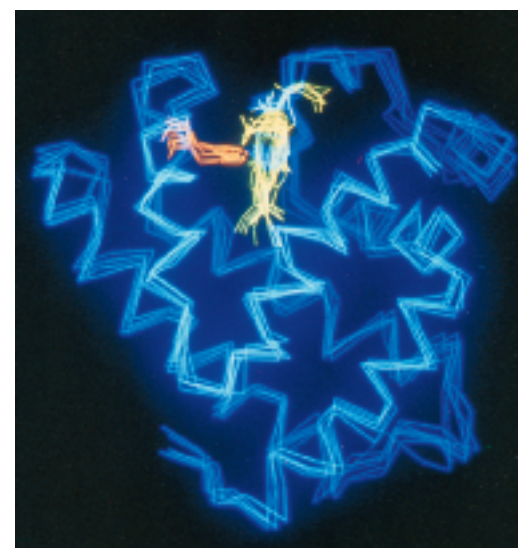
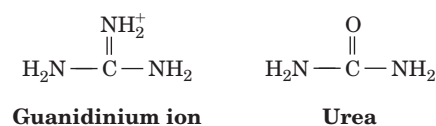


Figure 6-38 Molecular dynamics of myoglobin. Several “snapshots” of the protein calculated at intervals of 5×10^{-12} s are superimposed. The backbone is blue, the heme group is yellow, and the His side chain linking the heme to the protein is orange. [Courtesy of Martin Karplus, Harvard University.]

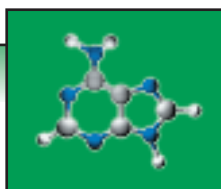
C Protein Denaturation and Renaturation

The low conformational stabilities of native proteins make them easily susceptible to denaturation by altering the balance of the weak nonbonding forces that maintain the native conformation. Proteins can be denatured by a variety of conditions and substances:

1. Heating causes a protein's conformationally sensitive properties, such as optical rotation (Section 4-2), viscosity, and UV absorption, to change abruptly over a narrow temperature range. Such a sharp transition indicates that the entire polypeptide unfolds or “melts” **cooperatively**, that is, nearly simultaneously. Most proteins have melting temperatures that are well below 100°C . Among the exceptions are the proteins of thermophilic bacteria (Box 6-3).
2. pH variations alter the ionization states of amino acid side chains, thereby changing protein charge distributions and hydrogen-bonding requirements.
3. Detergents associate with the nonpolar residues of a protein, thereby interfering with the hydrophobic interactions responsible for the protein's native structure.
4. The **chaotropic agents** guanidinium ion and urea,



in concentrations in the range 5 to 10 M, are the most commonly used protein denaturants. Chaotropic agents are ions or small organic molecules that increase the solubility of nonpolar substances in water. Their effectiveness as denaturants stems from their ability to disrupt hydrophobic interactions, although their mechanism of action is not well understood.



BOX 6 - 3

*Perspectives in Biochemistry**Thermostable Proteins*

Certain species of bacteria known as **hyperthermophiles** grow at temperatures near 100°C. They live in such places as hot springs and submarine hydrothermal vents, with the most extreme, the archeon *Pyrolobus fumarii*, able to grow at temperatures as high as 113°C. These organisms have many of the same metabolic pathways as do **mesophiles** (organisms that grow at “normal” temperatures). Yet most mesophilic proteins denature at temperatures where hyperthermophiles thrive. What is the structural basis for the thermostability of hyperthermophilic proteins?

The difference in the thermal stabilities of the corresponding (hyper)thermophilic and mesophilic proteins does not exceed $\sim 100 \text{ kJ} \cdot \text{mol}^{-1}$, the equivalent of a few noncovalent interactions. This is probably why comparisons of the X-ray structures of hyperthermophilic enzymes with their mesophilic counterparts have failed to reveal any striking differences between them. These proteins exhibit some variations in secondary structure but no more than would be expected for homologous proteins from distantly related mesophiles. However, several of these thermostable enzymes have a superabundance of salt bridges on their surfaces, many of which are arranged in extensive networks containing up to 18 side chains.

The idea that salt bridges can stabilize a protein structure appears to contradict the conclusion of Section 6-4A that ion pairs are, at best, marginally stable. The key to this apparent

paradox is that *the salt bridges in thermostable proteins form networks*. Thus, the gain in charge–charge free energy on associating a third charged group with an ion pair is comparable to that between the members of this ion pair, whereas the free energy lost on desolvating and immobilizing the third side chain is only about half that lost in bringing together the first two side chains. The same, of course, is true for the addition of a fourth, fifth, etc., side chain to a salt bridge network.

Not all thermostable proteins have such a high incidence of salt bridges. Structural comparisons suggest that these proteins are stabilized by a combination of small effects, the most important of which are an increased size of the protein's hydrophobic core, an increased size in the interface between its domains and/or subunits, and a more tightly packed core as evidenced by a reduced surface-to-volume ratio.

The fact that the proteins of hyperthermophiles and mesophiles are homologous and carry out much the same functions indicates that mesophilic proteins are by no means maximally stable. This, in turn, strongly suggests *that the marginal stability of most proteins under physiological conditions (averaging $\sim 0.4 \text{ kJ} \cdot \text{mol}^{-1}$ of amino acid residues) is an essential property that has arisen through evolutionary design*. Perhaps this marginal stability helps confer the structural flexibility that many proteins require to carry out their physiological functions.

Denatured Proteins Can Be Renatured. In 1957, the elegant experiments of Christian Anfinsen on **ribonuclease A (RNase A)** showed that proteins can be denatured reversibly. RNase A, a 124-residue single-chain protein, is completely unfolded and its four disulfide bonds reductively cleaved in an 8 M urea solution containing 2-mercaptoethanol. Dialyzing away the urea and reductant and exposing the resulting solution to O_2 at pH 8 (which oxidizes the SH groups to form disulfides) yields a protein that is virtually 100% enzymatically active and physically indistinguishable from native RNase A (Fig. 6-39). The protein must therefore **renature** spontaneously.

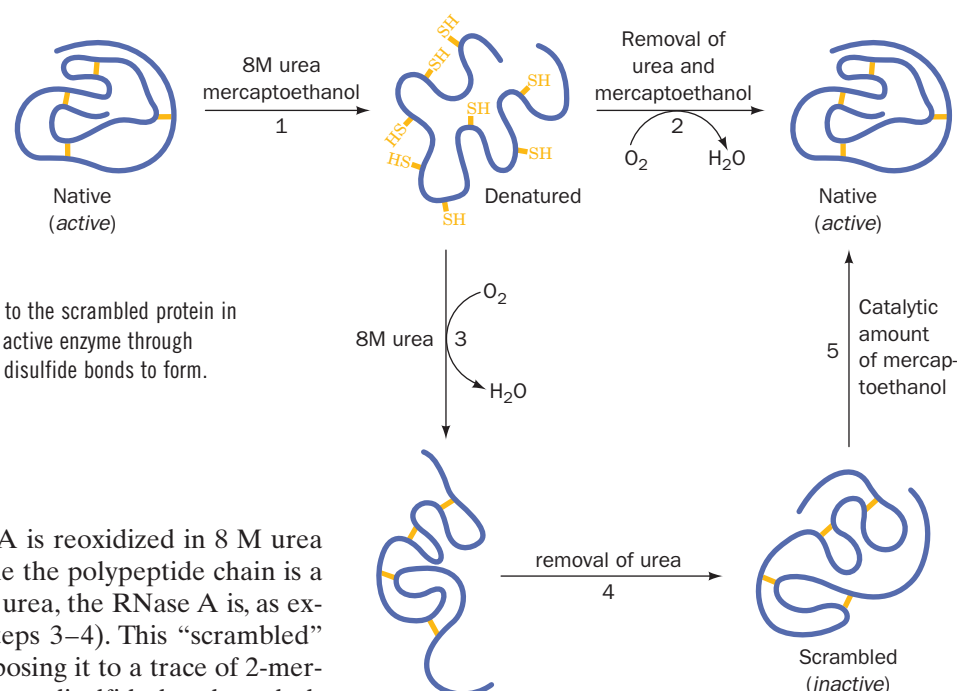
The renaturation of RNase A demands that its four disulfide bonds reform. The probability of one of the eight Cys residues randomly forming a disulfide bond with its proper mate among the other seven Cys residues is $1/7$; that of one of the remaining six Cys residues then randomly forming its proper disulfide bond is $1/5$; etc. The overall probability of RNase A re-forming its four native disulfide links at random is

$$\frac{1}{7} \times \frac{1}{5} \times \frac{1}{3} \times \frac{1}{1} = \frac{1}{105}$$

Clearly, the disulfide bonds do not randomly re-form under renaturing conditions, since, if they did, only 1% of the refolded protein would be cat-

Figure 6-39 Denaturation and renaturation of

RNase A. The polypeptide is represented by a blue line, with its disulfide bonds in yellow. After the protein has been denatured and its disulfide bonds cleaved (1), it will renature in the presence of O_2 when the denaturant (urea) and reductant (mercaptoethanol) are removed (2). If the disulfide bonds are allowed to re-form (3) before the urea is removed (4), the bonds form at random and the resulting protein is enzymatically inactive. Adding a small amount of mercaptoethanol to the scrambled protein in the absence of O_2 (5) catalyzes its conversion to the active enzyme through disulfide interchange reactions that allow the native disulfide bonds to form.



alytically active. Indeed, if the RNase A is reoxidized in 8 M urea so that its disulfide bonds re-form while the polypeptide chain is a random coil, then after removal of the urea, the RNase A is, as expected, only ~1% active (Fig. 6-39, Steps 3–4). This “scrambled” protein can be made fully active by exposing it to a trace of 2-mercaptoethanol, which breaks the improper disulfide bonds and allows the proper bonds to form. *Anfinsen’s work demonstrated that proteins can fold spontaneously into their native conformations under physiological conditions. This implies that a protein’s primary structure dictates its three-dimensional structure.*

5 Protein Folding

Studies of protein stability and renaturation suggest that protein folding is directed largely by the residues that occupy the interior of the folded protein. But *how* does a protein fold to its native conformation? One might guess that this process occurs through the protein’s random exploration of all the conformations available to it until it eventually stumbles onto the correct one. A simple calculation first made by Cyrus Levinthal, however, convincingly demonstrates that this cannot possibly be the case: Assume that an n -residue protein’s 2^n torsion angles, ϕ and ψ , each have three stable conformations. This yields $3^{2n} \approx 10^n$ possible conformations for the protein (a gross underestimate because we have completely neglected its side chains). Then, if the protein could explore a new conformation every 10^{-13} s (the rate at which single bonds reorient), the time t , in seconds, required for the protein to explore all the conformations available to it is

$$t = \frac{10^n}{10^{13}}$$

For a small protein of 100 residues, $t = 10^{87}$ s, which is immensely greater than the apparent age of the universe (20 billion years, or 6×10^{17} s). Clearly, proteins must fold more rapidly than this.

A Protein Folding Pathways

Experiments have shown that many proteins fold to their native conformations in less than a few seconds. This is because *proteins fold to their native conformations via directed pathways rather than stumbling on them through random conformational searches*. Thus, as a protein folds, its confor-

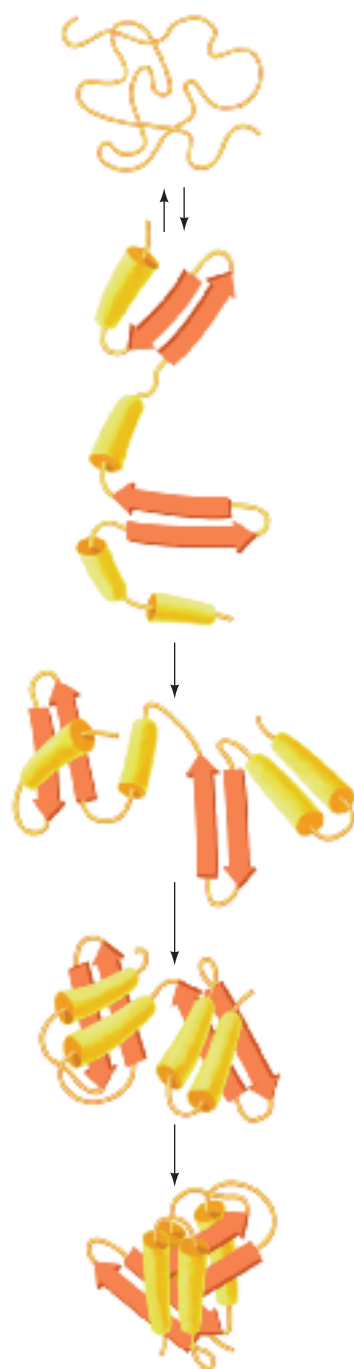


Figure 6-40 Hypothetical protein folding pathway. This example shows a linear pathway for folding a two-domain protein. [After Goldberg, M.E., *Trends Biochem. Sci.* **10**, 389 (1985).]

mational stability increases sharply (i.e., its free energy decreases sharply), which makes folding a one-way process. A hypothetical folding pathway is diagrammed in Fig. 6-40.

Experimental observations indicate that protein folding begins with the formation of local segments of secondary structure (α helices and β sheets). This early stage of protein folding is extremely rapid, with much of the native secondary structure in small proteins appearing within 5 ms of the initiation of folding. Since native proteins contain compact hydrophobic cores, it is likely that the driving force in protein folding is what has been termed a **hydrophobic collapse**. The collapsed state is known as a **molten globule**, a species that has much of the secondary structure of the native protein but little of its tertiary structure. Theoretical studies suggest that helices and sheets form in part because they are particularly compact ways of folding a polypeptide chain.

Over the next 5 to 1000 ms, the secondary structure becomes stabilized and tertiary structure begins to form. During this intermediate stage, the nativelike elements are thought to take the form of subdomains that are not yet properly docked to form domains. In the final stage of folding, which for small single-domain proteins occurs over the next few seconds, the protein undergoes a series of complex motions in which it attains its relatively stable internal side chain packing and hydrogen bonding while it expels the remaining water molecules from its hydrophobic core.

In multidomain and multisubunit proteins, the respective units then assemble in a similar manner, with a few slight conformational adjustments required to produce the protein's native tertiary or quaternary structure. Thus, *proteins appear to fold in a hierarchical manner, with small local elements of structure forming and then coalescing to yield larger elements, which coalesce with other such elements to form yet larger elements, etc.*

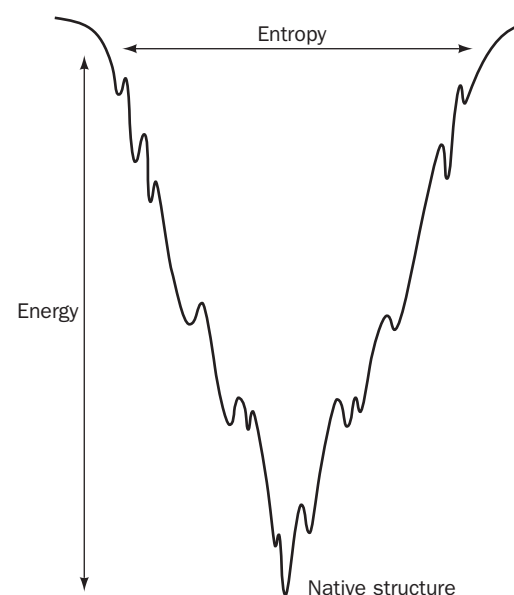
Folding, like denaturation, appears to be a cooperative process, with small elements of structure accelerating the formation of additional structures. A folding protein must proceed from a high-energy, high-entropy state to a low-energy, low-entropy state. This energy–entropy relationship, which is diagrammed in Fig. 6-41, is known as a **folding funnel**. An unfolded polypeptide has many possible conformations (high entropy). As it folds into an ever-decreasing number of possible conformations, its entropy and free energy decrease. The energy–entropy diagram is not a smooth valley but a jagged landscape. Minor clefts and gullies represent conformations that are temporarily trapped until, through random thermal activation, they overcome a slight “uphill” free energy barrier and can then proceed to a lower energy conformation. Evidently, *proteins have evolved to have efficient folding pathways as well as stable native conformations*.

Understanding the process of protein folding as well as the forces that stabilize folded proteins is essential for elucidating the rules that govern the relationship between a protein's amino acid sequence and its three-dimensional structure. Such information will prove useful in predicting the structures of the hundreds of thousands of proteins that are known only from their sequences (Box 6-4).

B Protein Disulfide Isomerase

Even under optimal experimental conditions, proteins often fold more slowly *in vitro* than they fold *in vivo*. One reason is that folding proteins often form disulfide bonds not present in the native proteins, which then slowly form native disulfide bonds through the process of disulfide interchange. **Protein disulfide isomerase (PDI)** catalyzes this process. Indeed, the observation that RNase A folds so much faster *in vivo* than *in vitro* led Anfinsen to discover this enzyme.

Figure 6-41 Energy–entropy diagram for protein folding. The width of the diagram represents entropy, and the depth, the energy. The unfolded polypeptide proceeds from a high-entropy, disordered state (*wide*) to a single low-entropy (*narrow*), low-energy native conformation. [After Onuchic, J.N., Wolynes, P.G., Luthey-Schulten, Z., and Socci, N.D., *Proc. Natl. Acad. Sci.* **92**, 3626 (1995).]



PDI binds to a wide variety of unfolded polypeptides via a hydrophobic patch on its surface. A Cys —SH group on reduced (SH-containing) PDI reacts with a disulfide group on the polypeptide to form a mixed disulfide and a Cys —SH group on the polypeptide (Fig. 6-42a). Another disulfide group on the polypeptide, brought into proximity by the spontaneous folding of the polypeptide, is attacked by this Cys —SH group. The newly liberated Cys —SH group then repeats this process with another disulfide bond, and so on, ultimately yielding the polypeptide containing only native disulfide bonds, along with regenerated PDI.

Oxidized (disulfide-containing) PDI also catalyzes the initial formation of a polypeptide's disulfide bonds by a similar mechanism (Fig. 6-42b). In this case, the reduced PDI reaction product must be reoxidized by cellular oxidizing agents in order to repeat the process.

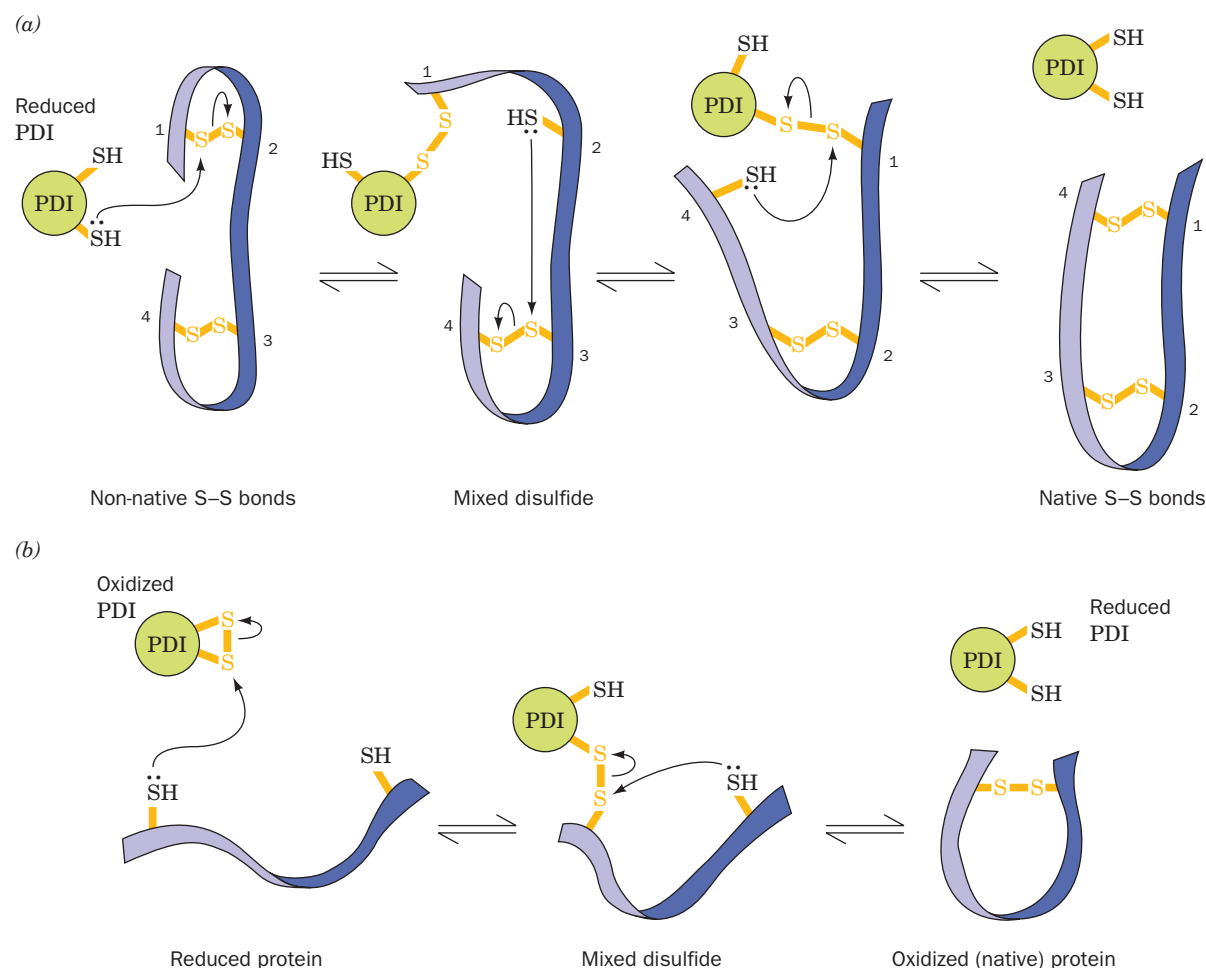

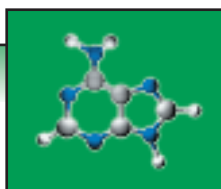


Figure 6-42 Mechanism of protein disulfide isomerase. (a) Reduced (SH-containing) PDI catalyzes the rearrangement of a polypeptide's non-native disulfide bonds via disulfide interchange reactions to yield native disulfide bonds. (b) Oxidized (disulfide-containing) PDI catalyzes the initial formation of

a polypeptide's disulfide bonds through the formation of a mixed disulfide. Reduced PDI can then react with a cellular oxidizing agent to regenerate oxidized PDI.  See the Animated Figures.



BOX 6 - 4

*Perspectives in Biochemistry**Protein Structure Prediction and Protein Design*

Hundreds of thousands of protein sequences are known either through direct protein sequencing (Section 5-3) or, more commonly, through genomic DNA sequencing (Section 3-4D). Yet the structures of only ~30,000 of these proteins have been determined by X-ray crystallography or NMR techniques. Consequently, there is a need to develop robust techniques for predicting a protein's structure from its amino acid sequence. This represents a formidable challenge but promises great rewards in terms of understanding protein function, identifying diseases related to abnormal protein sequences, and designing drugs to alter protein structure or function.

There are several major approaches to protein structure prediction. The simplest and most reliable approach, **homology modeling**, aligns the sequence of interest with the sequence of a homologous protein or domain of known structure—compensating for amino acid substitutions, insertions, and deletions—through modeling and energy minimization calculations. This method yields reliable models for proteins that have as little as 25% sequence identity with a protein of known structure, although, of course, the accuracy of the model increases with the degree of sequence identity. The emerging field of **structural genomics**, which seeks to determine the structures of all representative domains, is aimed at expanding this predictive technique. The identification of structural homology is likely to provide clues as to a protein's function even with imperfect structure prediction.

Distantly related proteins may be structurally similar even though they have diverged to such an extent that their sequences show no obvious resemblance. **Threading** is a computational technique that attempts to determine the unknown structure of a protein by ascertaining whether it is consistent with a known protein structure. It does so by placing (threading) the unknown protein's residues along the backbone of a known protein structure and then determining whether the amino acid side chains of the unknown protein are stable in that arrangement. This method is not yet reliable, although it has yielded encouraging results.

Empirical methods based on experimentally determined statistical information such as the α helix and β sheet propen-

sities deduced by Chou and Fasman (Table 6-1) have been moderately successful in predicting the secondary structures of proteins. Their main drawback is that neighboring residues in a polypeptide sometimes exert strong influence on a given residue's tendency to form a particular secondary structure.

Since the native structure of a protein ultimately depends on its amino acid sequence, it should be possible, in principle, to predict the structure of a protein based only on its chemical and physical properties (e.g., the hydrophobicity, size, hydrogen-bonding propensity, and charge of each of its amino acid residues). Such *ab initio* (from the beginning) methods are still only moderately successful in predicting the structures of small polypeptides. Computational molecular biologists periodically test their methods on sequences of proteins whose structures are undergoing conventional structural determination. In many cases, the theoretical results closely approximate the true structures (Fig. 1). This sort of modeling may be enough to provide clues to a protein's function, even if the exact positions of all its side chains are uncertain.

Protein design, the experimental inverse of protein structure prediction, has provided insights into protein folding and stability. Protein design may begin with a target structure such as a simple sandwich of β sheets or a bundle of four α helices. It attempts to construct an amino acid sequence that will form that structure. The designed polypeptide is then chemically or biologically synthesized, and its structure is determined. Fortunately, protein folding seems to be governed more by extended sequences of amino acids than by individual residues, which allows some room for error in designing polypeptides. Experimental results suggest that the greatest challenge of protein design may lie not in getting the polypeptide to fold to the desired conformation but in preventing it from folding into other unwanted conformations. In this respect, science lags far behind nature.

The first wholly successful *de novo* (from the beginning) protein design, accomplished by Stephen Mayo, was for a 28-residue $\beta\beta\alpha$ motif that has a backbone conformation designed to resemble a zinc finger (Fig. 6-37) but that

C Molecular Chaperones

Proteins begin to fold as they are being synthesized, so the renaturation of a denatured protein *in vitro* may not entirely mimic the folding of a protein *in vivo*. In addition, proteins fold *in vivo* in the presence of extremely high concentrations of other proteins with which they can potentially interact. **Molecular chaperones** are essential proteins that bind to unfolded and partially folded polypeptide chains to prevent the improper association of ex-

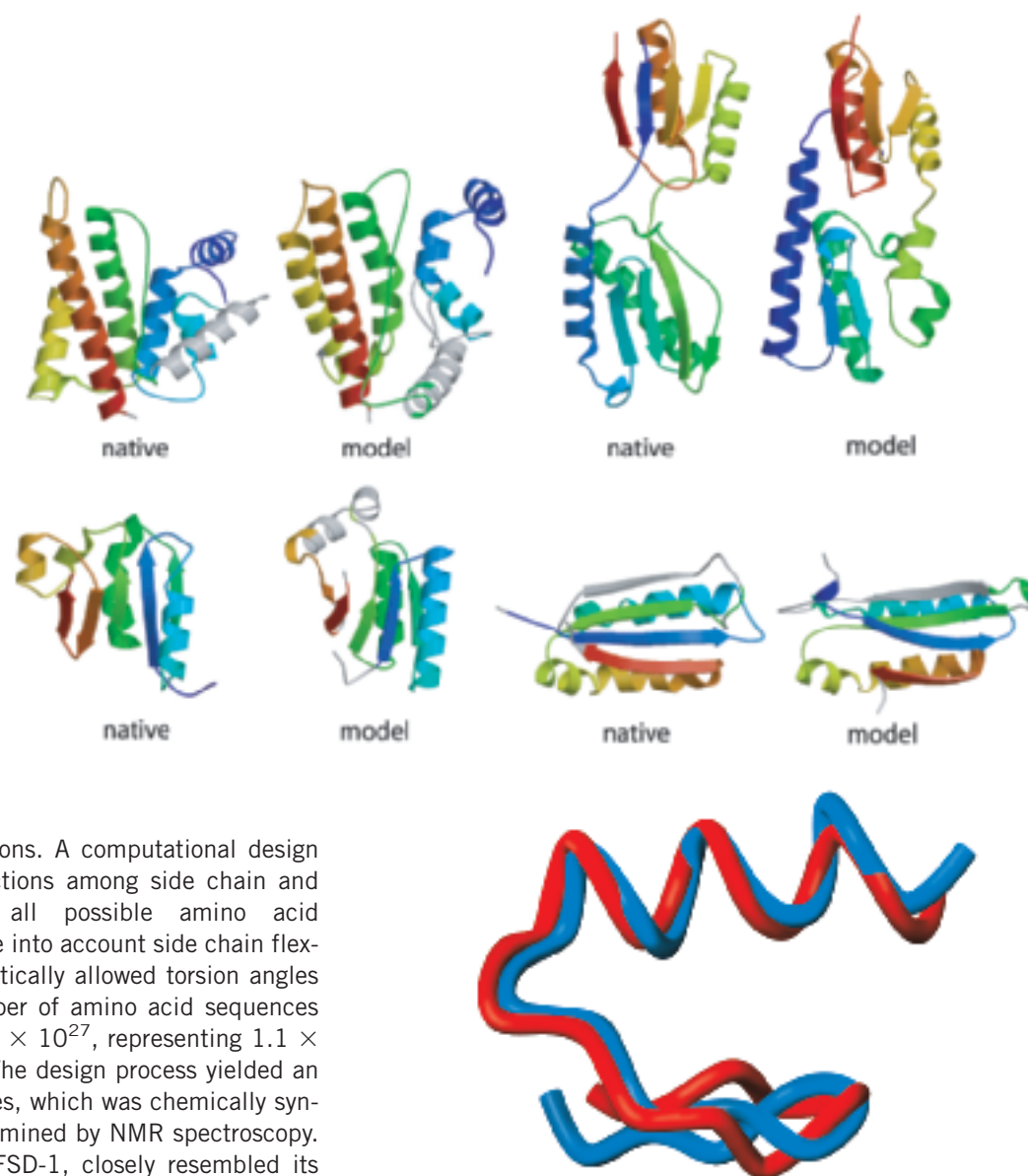


Figure 1 Comparison of experimentally determined (native) and predicted (model) folds of polypeptides. The polypeptides are colored in rainbow order from N-terminus (*indigo*) to C-terminus (*red*). [Courtesy of David Baker, University of Washington.]

contains no stabilizing metal ions. A computational design process considered the interactions among side chain and backbone atoms, screened all possible amino acid sequences, and, in order to take into account side chain flexibility, tested all sets of energetically allowed torsion angles for each side chain. The number of amino acid sequences to be tested was limited to 1.9×10^{27} , representing 1.1×10^{62} possible conformations! The design process yielded an optimal sequence of 28 residues, which was chemically synthesized and its structure determined by NMR spectroscopy. The designed protein, called FSD-1, closely resembled its predicted structure, and its backbone conformation was nearly superimposable on that of a known zinc finger motif (Fig. 2). Although FSD is relatively small, it folds into a unique stable structure, thereby demonstrating the power of protein design techniques.

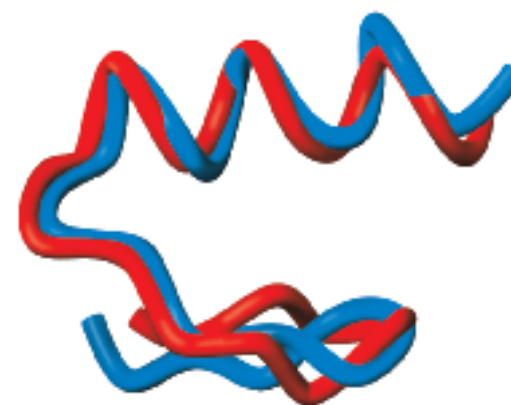


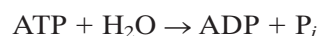
Figure 2 Structure of a designed protein. The structure of the 28-residue FSD-1 was designed to resemble that of a zinc finger motif (from the protein Zif268). The polypeptide backbones of FSD-1 (*blue*) and Zif268 (*red*) are nearly superimposable. [Courtesy of Stephen Mayo, California Institute of Technology.]

posed hydrophobic segments that might lead to non-native folding as well as polypeptide aggregation and precipitation. This is especially important for multidomain and multisubunit proteins, whose components must fold fully before they can properly associate with each other. Molecular chaperones also induce misfolded proteins to refold to their native conformations.

Many molecular chaperones were first described as **heat shock proteins (Hsp)** because their rate of synthesis is increased at elevated temperatures.

Presumably, the additional chaperones are required to recover heat-denatured proteins or to prevent misfolding under conditions of environmental stress.

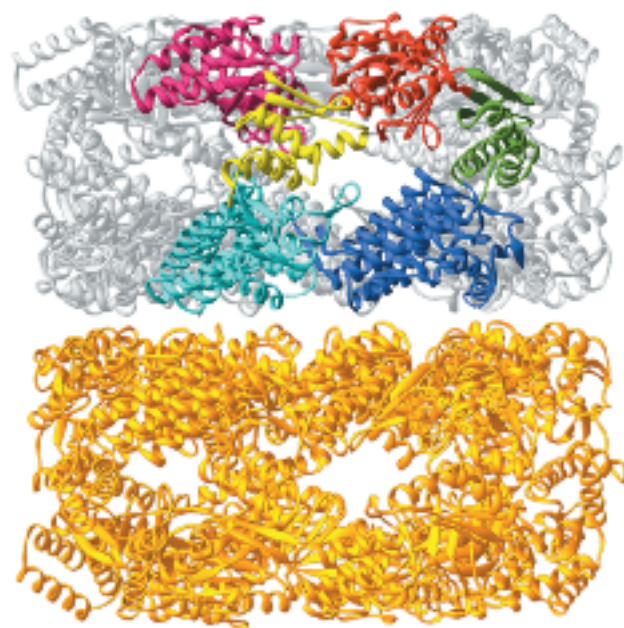
There are several classes of molecular chaperones in both prokaryotes and eukaryotes, including: (1) the **Hsp70** family of 70-kD proteins, which function as monomers; (2) the **chaperonins**, which are large multisubunit proteins; and (3) the **Hsp90** proteins, which are mainly involved with the folding of proteins involved with signal transduction such as **steroid receptors** (Section 27-3B). All of these molecular chaperones operate by binding to an unfolded or aggregated polypeptide's solvent-exposed hydrophobic surface and subsequently releasing it, often repeatedly, in a manner that facilitates its proper folding. Many molecular chaperones are **ATPases**, that is, enzymes that catalyze the hydrolysis of ATP (adenosine triphosphate) to ADP (adenosine diphosphate) and P_i (inorganic phosphate):



The favorable free energy change of ATP hydrolysis drives the chaperone's bind-and-release reaction cycle.

Hsp70 proteins are highly conserved 70-kD monomeric proteins in both prokaryotes and eukaryotes. An Hsp70 chaperone, which functions in association with the **cochaperone** protein **Hsp40**, appears to bind to a newly synthesized polypeptide as it emerges from the ribosome. The binding and release of small hydrophobic regions on the new polypeptide may prevent its premature folding. Other chaperones apparently complete the job begun by the Hsp70 proteins. The Hsp70 proteins also function to unfold proteins in preparation for their transport through membranes (Section 9-4D) and to subsequently refold them.

(a)



(b)

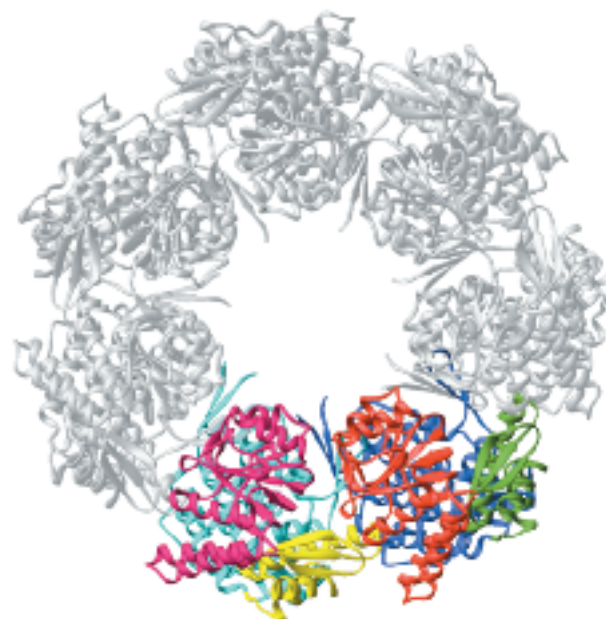


Figure 6-43 X-ray structure of GroEL. (a) Side view perpendicular to the 7-fold axis in which the seven identical subunits of the lower ring are gold and those of the upper ring are silver, with the exception of the two subunits nearest the viewer, whose equatorial, intermediate, and apical domains are colored blue, green, and red on the right subunit and cyan, yellow, and magenta on the left subunit. The two rings of the complex are held together

through side chain interactions that are not seen in this drawing. (b) Top view along the 7-fold axis in which only the upper ring is shown for the sake of clarity. Note the large central channel that appears to run the length of the protein. [Based on an X-ray structure by Axel Brünger, Arthur Horwich, and Paul Sigler, Yale University. PDBid 1OEL.]

The GroEL/ES Chaperones Form a Barrel Structure. The chaperonins consist of two types of proteins, named **Hsp60** and **Hsp10**. Those in *E. coli*, the best-characterized chaperonins, are called **GroEL** and **GroES**. They are essential for the survival of *E. coli* under all conditions tested. Fourteen identical 549-residue GroEL subunits are arranged in two stacked rings of seven subunits each (Fig. 6-43) to form a complex with D_7 symmetry (with perpendicular 7-fold and 2-fold axes of symmetry). The X-ray structure of GroEL, determined by Arthur Horwich and Paul Sigler, reveals that it forms a porous thick-walled hollow cylinder with an inner diameter of ~ 45 Å. The central channel forms two chambers in which partially folded proteins fold to their native structures. A constriction in the center prevents a folding protein from passing between the two GroEL rings.

The 97-residue GroES subunits form a domelike heptameric ring with C_7 symmetry (Fig. 6-44) in which the inner surface of the GroES dome is lined with hydrophilic residues. The X-ray structure of a GroEL–GroES–(ADP)₇ complex, also determined by Horwich and Sigler, indicates that GroES closes over one GroEL ring like a lid on a pot to form a bullet-shaped structure with C_7 symmetry (Fig. 6-45). The GroEL ring that contacts the GroES heptamer is called the *cis* ring; the opposing GroEL ring is known as the *trans* ring.

ATP Binding and Hydrolysis Coordinate the Conformational Changes in GroEL/ES. Each GroEL subunit has a binding pocket for ATP. A conformational change activates GroEL's ATPase activity by completely enclosing the ATP with protein (Fig. 6-45c) while shifting a catalytically essential Asp side chain into a productive position. In the structure shown in Fig. 6-45, the *cis* ring has hydrolyzed its seven molecules of ATP to ADP and has undergone

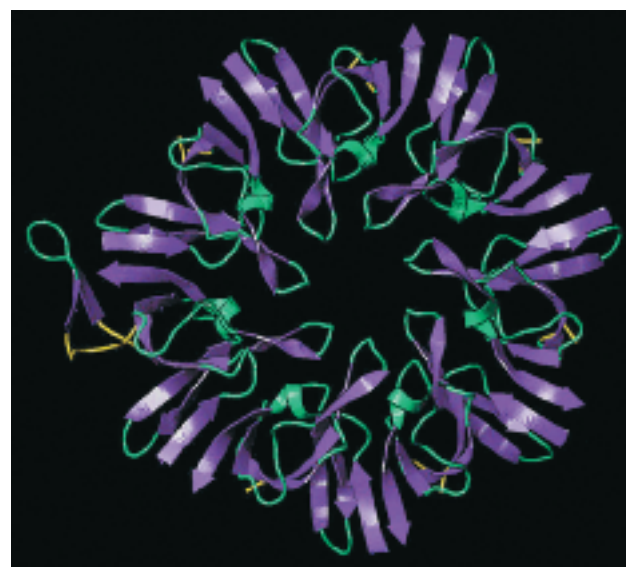
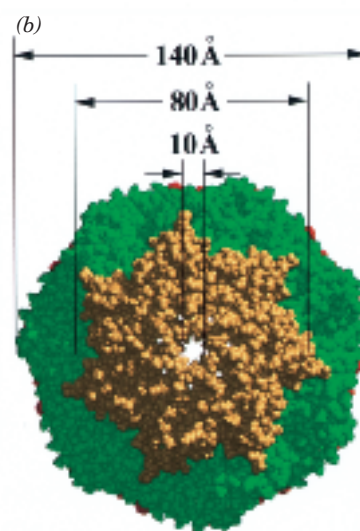
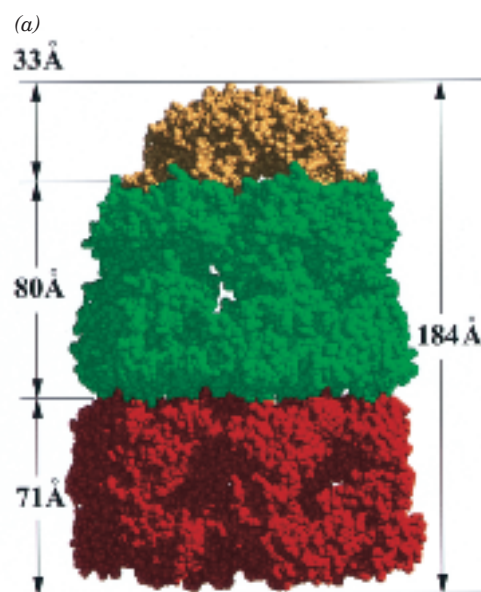


Figure 6-44 X-ray structure of GroES as viewed along its 7-fold axis. A mobile loop in only one of the seven identical subunits (*left*) is visible in this structure. The polypeptide segments that flank the mobile loop are yellow. [Courtesy of Johann Dieneshofer, University of Texas Southwest Medical Center, Dallas, Texas.]



(c)

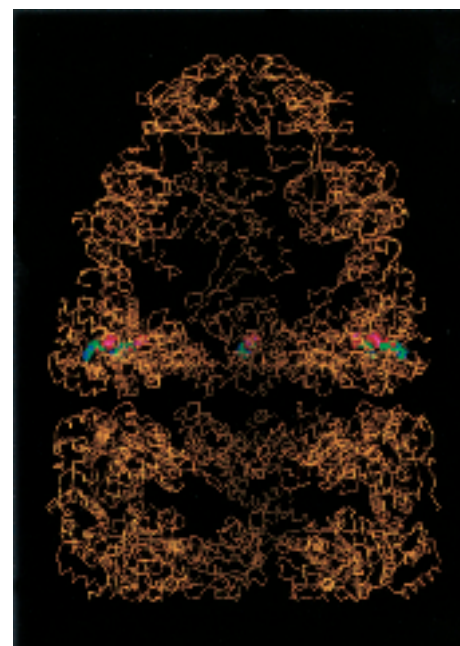


Figure 6-45 X-Ray structure of the GroEL–GroES–(ADP)₇ complex.

(a) A space-filling drawing as viewed perpendicularly to the complex's 7-fold axis with the GroES ring gold, the *cis* ring of GroEL green, and the *trans* ring of GroEL red. The dimensions of the complex are indicated. Note the different conformations of the two GroEL rings. (b) As in Part (a) but viewed along the

7-fold axis. (c) The C_α backbone of the complex viewed as in Part (a) but which is cut away along the plane containing the complex's 7-fold axis. The ADPs bound to the *cis* ring of GroEL are shown in space-filling form. Note the much larger size of the cavity formed by the *cis* ring and GroES in comparison to that of the *trans* ring. [Courtesy of Paul Sigler, Yale University. PDBid 1AON.]

conformational changes relative to the trans ring that widen and elongate the cis cavity in a way that more than doubles its volume (from 85,000 Å³ to 175,000 Å³). The enlarged cavity is able to enclose a partially folded substrate protein of at least 70 kD. *All seven subunits of the GroEL ring act in concert; that is, they are mechanically linked such that they can only change their conformations simultaneously.*

The cis and trans GroEL rings undergo conformational changes in a reciprocating fashion, with events in one ring influencing events in the other ring. The entire GroEL/ES chaperonin complex functions as follows (Fig. 6-46):

1. We begin with one GroEL ring binding 7 ATP and an improperly folded substrate protein, which associates with hydrophobic patches on the GroEL apical domains (labeled with an A in Fig. 6-46). The GroEL ring then binds a GroES cap to become the cis ring. GroES binding induces a conformational change in the cis ring that moves the hydrophobic patches to an interior position in GroEL, thereby depriving the substrate protein of its binding sites. This releases the substrate protein into the now enlarged and closed cavity, where the substrate protein commences folding. The cavity, which is now lined only with hydrophilic groups, provides the substrate protein with an

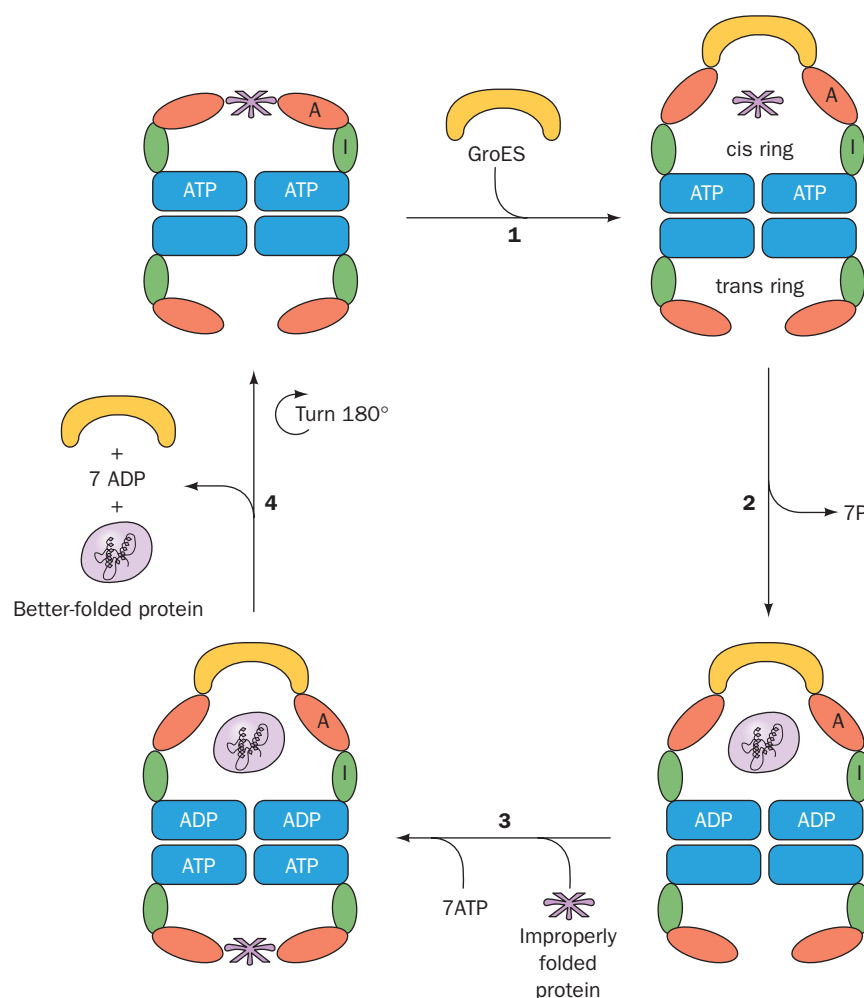


Figure 6-46 Reaction cycle of the GroEL/ES chaperonin. See the text for an explanation.

- isolated microenvironment that prevents it from nonspecifically aggregating with other misfolded proteins.
2. Within ~ 13 s (the time the substrate protein has to fold), the cis ring catalyzes the hydrolysis of its 7 bound ATPs to ADP + P_i and the P_i is released. The absence of ATP's γ phosphate group weakens the interactions that bind GroES to GroEL.
 3. A second molecule of substrate protein binds to the trans ring followed by 7 ATP. Conformational linkages between the cis and trans rings prevent the binding of both substrate protein and ATP to the trans ring until the ATP in the cis ring has been hydrolyzed.
 4. The binding of substrate protein and ATP to the trans ring conformationally induces the cis ring to release its bound GroES, 7 ADP, and the presumably now better folded substrate protein. This leaves ATP and substrate protein bound only to the trans ring of GroEL, which now becomes the cis ring as it binds GroES.

Steps 1 through 4 are then repeated. The GroEL/ES system expends 7 ATPs per folding cycle. If the released substrate protein has not achieved its native state, it may subsequently rebind to GroEL (a substrate protein that has achieved its native fold lacks exposed hydrophobic groups and hence cannot rebind to GroEL). It requires an average of 24 folding cycles for a protein to attain its native state, which necessitates the hydrolysis of 168 ATPs (which appears to be a profligate use of ATP but constitutes only a small fraction of the thousands of ATPs that must be hydrolyzed to synthesize a typical polypeptide and its component amino acids). Because protein folding occurs alternately in the two GroEL rings, the proper functioning of the chaperonin requires both GroEL rings, even though their two cavities are unconnected.

In addition to forming a protective cage around a folding protein, the chaperonin apparently promotes the refolding of an improperly folded protein by an **iterative annealing** process. In this model, a misfolded protein binds to the hydrophobic patches on two or more of the seven GroEL subunits. The binding of ATP and GroES then triggers the conformational changes that mask these hydrophobic patches, a process that stretches and thereby partially unfolds the bound protein before it is released. This rescues the protein from a local energy minimum in which it had become trapped (Fig. 6-41) and thereby permits it to continue its conformational journey down the folding funnel toward its native state (the state of lowest free energy).

What Kinds of Proteins Are Folded by Chaperonins? *In vivo*, the GroEL/ES system interacts with only a subset of *E. coli* proteins. To identify these proteins, Ulrich Hartl supplied *E. coli* cells with [^{35}S]methionine (^{35}S is a radioactive isotope of S) for the 15 s it takes to synthesize an average-length polypeptide and then added an excess of unlabeled methionine (a **pulse-chase** experiment) for varying lengths of time. The cells were then lysed, the GroEL–GroES–substrate complexes were immunoprecipitated with anti-GroEL antibodies, and the GroEL-bound proteins were separated by electrophoresis. Of the ~ 2500 cytosolic proteins that could be detected, only ~ 300 proteins were reproducibly observed to be associated with GroEL. These proteins were isolated and 52 of them were unequivocally identified via the mass spectrometry of their tryptic fragments (see Section 5-3D).

Nearly all of the GroEL substrate proteins that were identified are enzymes that participate in a wide variety of metabolic functions or in tran-

scription or translation. Their molecular masses are mostly in the range 20 to 60 kD. Analysis of these proteins revealed that they tend to contain two or more $\alpha\beta$ domains that mainly consist of open β sheets. Such a protein is expected to fold only slowly to its native state because the formation of its hydrophobic β sheets requires the assembly of a large number of specific long-range interactions in their proper orientations. Moreover, such a protein may easily misfold or become kinetically trapped due to the improper packing of its helices and sheets in one domain or, more likely, between domains.

Most of the GroEL-associated proteins dissociated from the chaperonin within a few minutes, after having achieved their native folds. However, ~ 100 proteins remained partially associated with GroEL, even after 2 hours. Evidently, these proteins repeatedly return to GroEL for conformational maintenance, which strongly suggests that they are structurally labile and/or prone to aggregation.

D Diseases Caused by Protein Misfolding

Most proteins in the body maintain their native conformations or, if they become partially denatured, are either renatured through the auspices of molecular chaperones or are proteolytically degraded (Section 20-1). However, at least 18 different—and usually fatal—human diseases are associated with the extracellular deposition of normally soluble proteins in certain tissues in the form of insoluble fibrous aggregates known as **amyloid** (this term means starchlike; it was originally thought that this material resembled starch). These diseases include **Alzheimer's disease**, the **transmissible spongiform encephalopathies (TSEs)**, and the **amyloidoses**. The deposition of amyloid interferes with normal cellular function, resulting in cell death and eventual organ failure.

Alzheimer's Disease. Alzheimer's disease, a neurodegenerative condition that strikes mainly the elderly, causes devastating mental deterioration and eventual death (it affects $\sim 10\%$ of those over 65 and $\sim 50\%$ of those over 85). It is characterized by brain tissue containing abundant amyloid **plaques** (deposits) surrounded by dead and dying neurons (Fig. 6-47). The amyloid plaques consist mainly of fibrils of a 40- to 42-residue protein named **amyloid- β protein ($A\beta$)**. $A\beta$ is a fragment of a 770-residue membrane protein called the **$A\beta$ precursor protein (β PP)**, whose normal function is un-

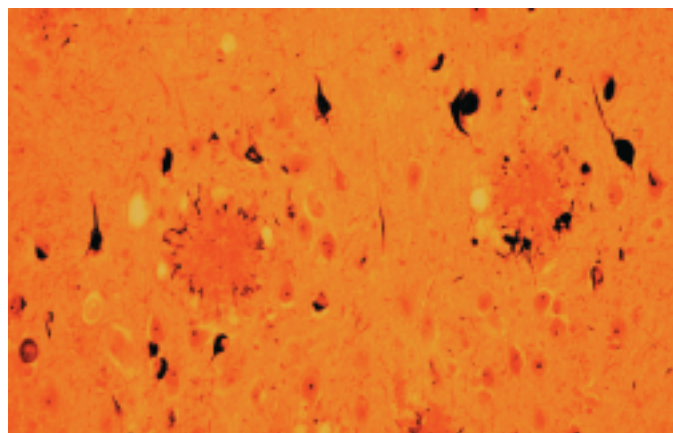


Figure 6-47 Photomicrograph of brain tissue from an individual with Alzheimer's disease. The two circular objects are plaques that consist of amyloid deposits of $A\beta$ protein surrounded by a halo of neurites (axons and dendrites) from dead and dying neurons. [Courtesy of Dennis Selkoe and Marcia Podlisny, Harvard University Medical School.]

known. A β is excised from β PP in a multistep process through the actions of two proteolytic enzymes dubbed **β -** and **γ -secretases**.

It had been hotly debated whether A β causes Alzheimer's disease or is merely a product of its neurodegenerative processes. This argument was largely put to rest by the observation that microinjecting 200 pg of fibrillar but not soluble A β (the amount of A β in an A β plaque) into the cerebral cortexes of aged but not young monkeys causes marked neuronal loss and other microscopic changes characteristic of Alzheimer's disease as far as 1.5 mm from the injection site. Evidently, *the A β fibrils are neurotoxic even before their deposition in amyloid plaques*.

The age dependence of Alzheimer's disease suggests that A β deposition is an ongoing process. Indeed, there are several rare variants of the β PP gene with mutations in their A β regions that result in the onset of Alzheimer's disease as early as the fourth decade of life. These mutations affect the proteolytic processing of β PP in a way that increases the rate of A β production. A similar phenomenon is seen in individuals with **Down's syndrome** [a condition characterized by mental retardation and a distinctive physical appearance caused by the trisomy (3 copies per cell) of chromosome 21 rather than the normal two copies], who invariably develop Alzheimer's disease by their 40th year. This is because the gene encoding β PP is located on chromosome 21 and hence individuals with Down's syndrome produce β PP and presumably A β at an accelerated rate. Consequently, a promising strategy for halting the progression of Alzheimer's disease is to develop drugs that inhibit the action of the β - and/or γ -secretases so as to decrease the rate of A β production.

Prion Diseases. Certain infectious diseases that affect the mammalian central nervous system were originally thought to be caused by "slow viruses" because they take months, years, or even decades to develop. Among them are **scrapie** (a neurological disorder of sheep and goats), **bovine spongiform encephalopathy (BSE or mad cow disease)**, and **kuru** (a degenerative brain disease in humans that was transmitted by ritual cannibalism among the Fore people of Papua New Guinea; *kuru* means "trembling"). There is also a sporadic (spontaneously arising) human disease with similar symptoms, **Creutzfeldt-Jakob disease (CJD)**, which strikes one person per million per year and which may be identical to kuru. In all of these invariably fatal diseases, neurons develop large vacuoles that give brain tissue a spongelike microscopic appearance. Hence the diseases are collectively known as **transmissible spongiform encephalopathies (TSEs)**.

Unlike other infectious diseases, *the TSEs are not caused by a virus or microorganism*. Indeed, extensive investigations have failed to show that they are associated with any nucleic acid. Instead, as Stanley Pruisner demonstrated for scrapie, the infectious agent is a protein called a **prion** (for *proteinaceous infectious particle* that lacks nucleic acid) and hence TSEs are alternatively called **prion diseases**. The scrapie prion, which is named **PrP** (for *Prion Protein*), consists of 208 mostly hydrophobic residues. This hydrophobicity causes partially proteolyzed PrP to aggregate as clusters of rodlike particles that closely resemble the amyloid fibrils seen on electron microscopic examination of prion-infected brain tissue (Fig. 6-48). These fibrils presumably form the amyloid plaques that appear to be directly responsible for the neuronal degeneration in TSEs.

How are prion diseases transmitted? PrP is the product of a normal cellular gene that has no known function (mice that mutagenically fail to ex-

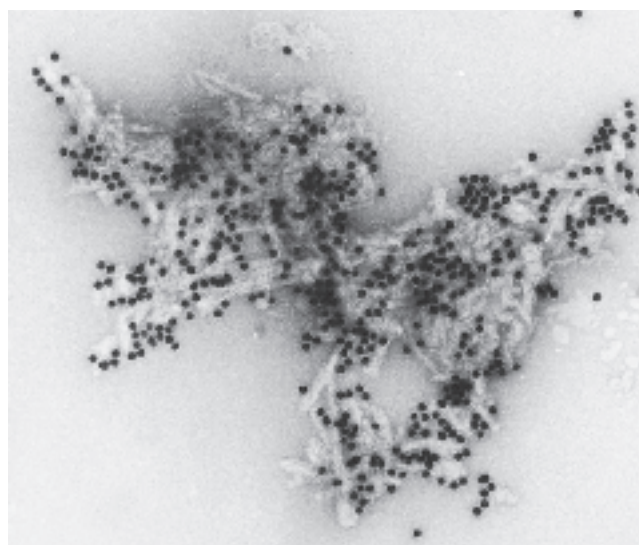


Figure 6-48 Electron micrograph of a cluster of partially proteolyzed prion rods. The black dots are colloidal gold beads that are coupled to anti-PrP antibodies adhering to the PrP. [Courtesy of Stanley Pruisner, University of California at San Francisco Medical Center.]

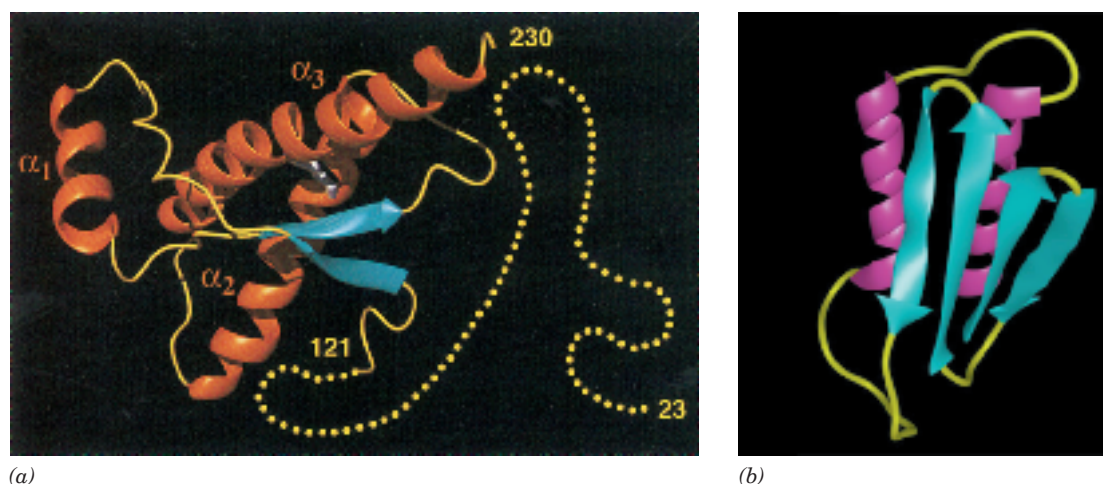


Figure 6-49 Prion protein conformations. (a) The NMR structure of human prion protein (PrP^{C}). Its flexibly disordered N-terminal “tail” (residues 23–121) is represented by yellow dots (the protein’s N-terminal 23 residues have been posttranslationally excised). (b) A plausible model for the structure

of PrP^{Sc} . [Part a courtesy of Kurt Wüthrich, Eidgenössische Technische Hochschule, Zurich, Switzerland. Part b courtesy of Fred Cohen University of California at San Francisco.]

press PrP appear to be normal and have apparently normal progeny). Infection of cells by prions somehow alters the PrP protein (the transcription of the PrP gene itself is not altered). Indeed, various methods have demonstrated that the scrapie form of PrP (PrP^{Sc}) is identical to normal cellular PrP (PrP^{C}) in sequence and chemical structure but that they differ in their secondary and/or tertiary structures. This suggests that PrP^{Sc} induces PrP^{C} to adopt the conformation of PrP^{Sc} , that is, PrP^{Sc} formation is **auto-catalytic** (the initially formed PrP^{Sc} induces the formation of additional PrP^{Sc} from PrP^{C} , etc). This accounts for the observation that mice that do not express the gene encoding PrP cannot be infected with scrapie (which can be transmitted to normal mice by the intracerebral inoculation of PrP^{Sc}).

Human PrP^{C} consists of a disordered (and hence unseen) 99-residue N-terminal “tail” and a 110-residue C-terminal globular domain containing three α helices and a short two-stranded antiparallel β sheet (Fig. 6-49a). Unfortunately, the insolubility of PrP^{Sc} has precluded its structural determination, but spectroscopic methods indicate that it has a lower α helix content and a higher β sheet content. This suggests that the protein has refolded (Fig. 6-49b). The high β sheet content of PrP^{Sc} presumably facilitates the aggregation of PrP^{Sc} as amyloid fibrils (see below).

Prion diseases can be transmitted by the consumption of nerve tissue from infected individuals, as was first seen in the case of kuru. This has become particularly evident in the case of BSE, which was unknown before 1985 but reached epidemic proportions among cattle in the U.K. in 1993. This is because the process for preparing meat-and-bone meal, which was routinely fed to cattle, was changed in the 1970s from one that fully inactivated prions to one that fails to do so. The BSE epidemic rapidly abated after 1993 due to the banning, in 1988, of the feeding of ruminants with ruminant-derived products other than milk, together with the slaughter of a large number of cattle at risk for having BSE. However, it is now clear that BSE was transmitted to humans who ate meat from BSE-infected cattle: Some 130 cases of so-called **new variant CJD (nvCJD)** have been reported to date, almost entirely in the U.K., many of which occurred in teenagers and young adults. Yet before 1994, CJD under the age of 40 was

extremely rare (its average age of onset is ~ 64). It should be noted that the transmission of BSE from cattle to humans was unexpected: Scrapie-infected sheep have long been consumed worldwide and yet the incidence of CJD in mainly meat-eating countries such as the U.K. (in which sheep are particularly abundant) was no greater than that in largely vegetarian countries such as India.

Amyloidoses. Many amyloidogenic proteins are mutant forms of normally occurring proteins. These include **lysozyme** (an enzyme that hydrolyzes bacterial cell walls; Section 11-4) in the disease **familial visceral amyloidosis**, **transthyretin** (a blood plasma protein that functions as a carrier for water-insoluble hormones including the thyroid hormone **thyroxine**; Fig. 4-16) in **familial amyloid polyneuropathy**, and **fibrinogen** (the precursor of **fibrin**, which forms blood clots; Box 11-4) in **hereditary renal amyloidosis**. Most such diseases do not present (become symptomatic) until the third to seventh decades of life and typically progress over 5 to 15 years, ending in death. In addition, three dominantly inherited neurodegenerative diseases have been traced to mutations in the gene encoding PrP: **familial CJD**, **Gerstmann-Sträussler-Scheinker syndrome**, and **fatal familial insomnia**. These mutations presumably increase the rate that PrP^C converts to PrP^{Sc}.

Amyloid Fibrils Are β Sheet Structures. The amyloid fibers that characterize the amyloidoses, Alzheimer's disease, and the TSEs are built from proteins that exhibit no structural or functional similarities in their native states. In contrast, the appearance of their fibrillar forms is strikingly similar. Furthermore, the ability to form amyloid fibrils is not unique to the small set of proteins associated with specific diseases. Under the appropriate conditions, almost any protein can be induced to aggregate. Thus, *the ability to form amyloid may be an intrinsic property of all polypeptide chains*.

Spectroscopic analysis of amyloid fibrils indicates that they are rich in β structure, with individual β strands oriented perpendicular to the fiber axis (Fig. 6-50). Even myoglobin, a globular protein consisting almost entirely of α helices (Fig. 6-38) can be coaxed into a fibrillar form by a pH of 9 and a temperature of 65°C. Results such as these raise the question of how a given amino acid sequence can adopt two very different but well-ordered states. The native structure of myoglobin is defined in large part by interactions between side chains; in the amyloid form, main chain interactions may dominate (side chain interactions are less important in stabilizing β sheets).

A variety of experiments indicate that amyloidogenic mutant proteins are significantly less stable than their wild-type counterparts (e.g., they have significantly lower melting tempera-

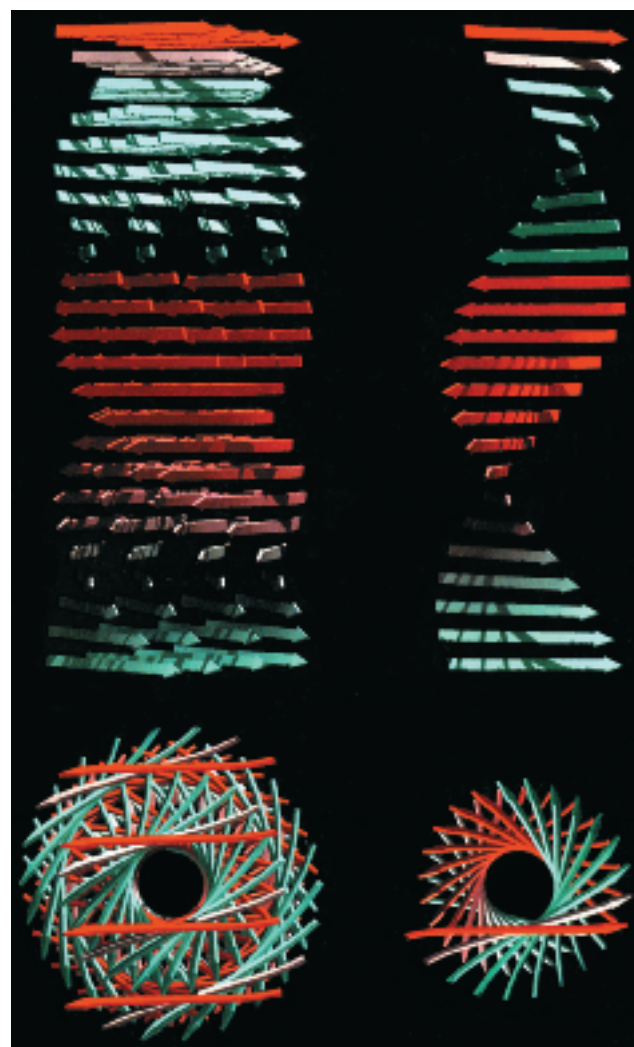


Figure 6-50 A model, based on X-ray fiber diffraction measurements, of an amyloid fibril. (a) The model is viewed normal to the fibril axis (*above*) and along the fibril axis (*below*). The arrowheads indicate the path but not necessarily the direction of the β strands. (b) An isolated β sheet, which is shown for clarity. The loop regions connecting the β strands have unknown structure. [Courtesy of Colin Blake, Oxford University, Oxford, U.K. and Louise Serpell, University of Cambridge, U.K.]

(a)

(b)

tures). This suggests that the partially unfolded, aggregation-prone forms are in dynamic equilibrium with the native conformation, even under conditions in which the native state is thermodynamically stable [keep in mind that the equilibrium ratio of unfolded (U) to native (N) protein molecules in the reaction $N \rightleftharpoons U$ is governed by Eq. 1-17: $[U]/[N] = e^{-\Delta G^{\circ'}/RT}$, where $\Delta G^{\circ'}$ is the standard free energy of unfolding, so that as $\Delta G^{\circ'}$ decreases, the equilibrium proportion of U increases]. It is therefore likely that fibrillogenesis is initiated by the association of the β domains of two or more partially unfolded amyloidogenic proteins to form a more extensive β sheet. This would provide a template or nucleus for the recruitment of additional polypeptide chains to form the growing fibril. However, the several decades that most amyloid diseases require to become symptomatic suggests that the spontaneous generation of an amyloid nucleus is a rare event, that is, it has a high free energy of activation (activation barriers and their relationship to reaction rates are discussed in Section 12-1B).

Once it has formed, an amyloid fibril is virtually indestructible under physiological conditions, possibly due to the large number of hydrogen bonds that must be broken in order to separate the individual polypeptide strands. It seems likely that protein folding pathways have evolved not only to allow polypeptides to assume stable native structures but also to avoid forming interchain hydrogen bonds that would lead to fibril formation. The factors that trigger amyloid formation remain obscure, even when mutations (in the case of hereditary amyloidoses) or infection (in the case of TSEs) appear to be the cause. One possibility is that the amyloid diseases result in part from malfunctions in the mechanisms that govern protein folding or the disposal of misfolded proteins.

6 Structural Bioinformatics

The data obtained by X-ray crystallography, NMR spectroscopy, and certain other techniques take the form of three-dimensional coordinates describing the spatial positions of atoms in molecules. This kind of information can be easily stored, displayed, and compared, much like sequence information obtained by nucleotide or protein sequencing methods (see Sections 3-4 and 5-3). **Bioinformatics** is the rapidly growing discipline that deals with the burgeoning amount of information related to molecular sequences and structures. **Structural bioinformatics** is a branch of bioinformatics that is concerned with how macromolecular structures are displayed and compared. Some of the databases and analytical tools that are used in structural bioinformatics are described here.

The Protein Data Bank. The atomic coordinates of nearly all known macromolecular structures are archived in the **Protein Data Bank (PDB)**. Indeed, most scientific journals that publish macromolecular structures require that authors deposit their structure's coordinates in the PDB. The PDB contains the coordinates of nearly 30,000 macromolecular structures (proteins, nucleic acids, and carbohydrates as determined by X-ray and other diffraction-based techniques, NMR, and theoretical modeling) and is growing exponentially. The PDB's Web address, from which these coordinates are publicly available, is listed in Table 6-3.

Each independently determined structure in the PDB is assigned a unique four-character identifier (its PDBid). For example, the PDBid for the structure of sperm whale myoglobin is 1MBO. A coordinate file begins with information that identifies the macromolecule, its source (the organ-

Table 6-3 Structural Bioinformatics Internet Addresses**Structural Databases**

Protein Data Bank (PDB): <http://www.rcsb.org/pdb/>
 Nucleic Acid Databank: <http://ndbserver.rutgers.edu/>
 Molecular Modeling Database (MMDB): <http://www.ncbi.nlm.nih.gov/Structure/index.shtml>
 Most Representative NMR Structure in an Ensemble: <http://pqs.ebi.ac.uk/pqs-nmr.html>
 PQS Protein Quaternary Structure Query Form at the EBI: <http://pqs.ebi.ac.uk/>

Molecular Graphics Programs/Plug-Ins

Chime: <http://mdli.com/products/framework/chemscape/>
 Cn3D: <http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml>
 Mage: <http://kinemage.biochem.duke.edu/>
 Protein Explorer: <http://www.umass.edu/microbio/chime/explorer/index.htm>
 RasMol: <http://www.bernstein-plus-sons.com/software/rasmol/> and <http://www.umass.edu/microbio/rasmol/>
 Swiss-PDB Viewer (Deep View): <http://us.expasy.org/spdbv/>

Structural Classification Algorithms

CATH (Class, Architecture, Topology, and Homologous superfamily): <http://www.biochem.ucl.ac.uk/bsm/cath/>
 CE (Combinatorial Extension of optimal pathway): <http://cl.sdsc.edu/>
 FSSP (Fold classification based on Structure-Structure alignment of Proteins): <http://www2.ebi.ac.uk/dali/fssp/>
 SCOP (Structural Classification Of Proteins): <http://scop.mrc-lmb.cam.ac.uk/scop/>
 VAST (Vector Alignment Search Tool): <http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml>

ism from which it was obtained), the author(s) who determined the structure, and key journal references. The file continues with a synopsis of how the structure was determined together with indicators of its accuracy. The sequences of the structure's various chains are then listed together with the descriptions and formulas of its so-called HET (for heterogen) groups, which are molecular entities that are not among the "standard" amino acid or nucleotide residues (for example, organic molecules, nonstandard residues such as Hyp, metal ions, and bound water molecules). The positions of the structure's secondary structural elements and its disulfide bonds are then provided.

The bulk of a PDB file consists of a series of lines, each of which provides the coordinates of one ATOM (for a "standard" residue) or HETATM (for a heterogen) in the structure. Each ATOM or HETATM is identified by a serial number and an atom name (for example, C and O for an amino acid residue's carbonyl C and O atoms, CA and CB for C_α and C_β atoms, N1 for atom N1 of a nucleic acid base, C4* for atom C4' of a ribose or deoxyribose residue). The record also includes the name of the residue (for example, PHE for phenylalanine, G for guanosine, and HEM for a heme group) and a letter to identify the chain to which it belongs (for structures that have more than one chain). The record then continues with the atom's three-dimensional (*x*, *y*, *z*) coordinates in angstroms. For NMR-based structures, the PDB file contains a full set of ATOM and HETATM records for each member of the ensemble of structures that were calculated in solving the structure (the most representative member of such a coordinate set can be obtained from <http://pqs.ebi.ac.uk/pqs-nmr.html>). PDB files usually end with CONECT (connectivity) records, which denote the nonstandard connectivities between atoms such as disulfide bonds and hydrogen bonds.

A particular PDB file may be located according to its PDBid or, if this is unknown, by searching for a protein's name, its source, the author(s), or

the experimental technique used to determine the structure. Selecting a particular macromolecule in the PDB initially displays a summary page with options for viewing the structure (either statically or interactively), for viewing or downloading the coordinate file, and for classifying or analyzing the structure in terms of its geometric properties and sequence.

The Nucleic Acid Database. The **Nucleic Acid Database (NDB)** archives the atomic coordinates of structures that contain nucleic acids. Its coordinate files have substantially the same format as PDB files. In fact, the same information is included in the PDB, but the NDB's organization and search algorithms are specialized for dealing with nucleic acids. This is useful because many nucleic acids of known structure are identified only by their sequences—rather than by names, as proteins are—and consequently could easily be overlooked in a search of the PDB.

Viewing Macromolecular Structures in Three Dimensions. The most informative way to examine a macromolecular structure is through the use of molecular graphics programs. Such programs permit the user to interactively rotate a macromolecule and thereby perceive its three-dimensional structure. This impression may be further enhanced by simultaneously viewing the macromolecule in stereo. Nearly all molecular graphics programs use PDB files as input. The programs described here can be downloaded from the Web addresses listed in Table 6-3, some of which also provide instructions for the program's use.

RasMol, a widely used molecular graphics program written by Roger Sayle, is publicly available for use on a variety of computer platforms (Windows, MacOS, and UNIX). Its Web browser-based counterpart (plugin) is named **Chime**. RasMol and Chime allow the user to simultaneously display different user-selected portions of a macromolecule in a variety of colors and formats (e.g., wire frame, ball and stick, backbone, space-filling, and cartoons). The Interactive Exercises on the CD-ROM that accompanies this textbook all use Chime. Moreover, the PDB provides the facility for viewing user-selected structures over the Web using several types of viewers including the Chime-based program **Protein Explorer** by Eric Martz. Another molecular graphics program, **Mage**, which was written by David Richardson, displays so-called **Kinemages** on this textbook's accompanying CD-ROM. Mage provides a generally more author-directed user environment than does RasMol or Chime.

The **Swiss-Pdb Viewer** (also called **Deep View**), in addition to displaying molecular structure, provides tools for basic model building, homology modeling, energy minimization, and multiple-sequence alignment. One advantage of the Swiss-Pdb Viewer is that it allows users to easily superimpose two or more models or parts of models. It too is publicly available and works on all major computer platforms.

Structural Classification and Comparison. Most proteins are structurally related to other proteins. Indeed, *evolution tends to conserve the structures of proteins rather than their sequences* (see Section 6-2D). The computational tools described below facilitate the classification and comparison of protein structures. These programs can be accessed directly via their Web addresses or through the PDB. Studies using these programs yield functional insights, reveal distant evolutionary relationships that are not apparent from sequence comparisons, generate libraries of unique folds for structure prediction, and provide indications as to why certain types of structures are preferred over others.

1. **CATH** (for *Class, Architecture, Topology, and Homologous superfamily*), as its name suggests, categorizes proteins in a four-level structural hierarchy. (1) “Class,” the highest level, places the selected protein in one of four categories of gross secondary structure: Mainly α , Mainly β , α/β (having both α helices and β sheets), and Few Secondary Structures. (2) “Architecture” is the description of the gross arrangement of secondary structure independent of topology. (3) “Topology” is indicative of both the overall shape and connectivity of the protein’s secondary structures. (4) “Homologous superfamily” is those proteins of known structure that are homologous (share a common ancestor) to the selected protein. A static or an interactive (Chime/RasMol or VRML) drawing of each of these proteins can be displayed.
2. **CE** (for *Combinatorial Extension of the optimal path*) finds all proteins in the PDB that can be structurally aligned with the query structure to within user-specified geometric criteria. The amino acid sequences of any or all of these proteins can be aligned on the basis of this structural alignment rather than sequence alignment. CE can likewise optimally align and display two user-selected structures.
3. **FSSP** (*Fold classification based on Structure–Structure alignment of Proteins*) lists the protein structures in the PDB that, at least in part, structurally resemble the query protein based on continuously updated all-against-all comparisons of the protein structures in the PDB. These structural comparisons are made by a program called **Dali** based on the distances between the various atoms in each domain of a protein.
4. **SCOP** (*Structural Classification Of Proteins*) classifies protein structures based mainly on manually generated topological considerations according to a six-level hierarchy: Class [all- α , all- β , α/β (having α helices and β strands that are largely interspersed), $\alpha + \beta$ (having α helices and β strands that are largely segregated), and multidomain (having domains of different classes)], Fold (groups that have similar arrangements of secondary structural elements), Superfamily (indicative of distant evolutionary relationships based on structural criteria and functional features), Family (indicative of near evolutionary relationships based on sequence as well as on structure), Protein, and Species. SCOP permits the user to navigate through its treelike hierarchical organization and lists the known members of any particular branch.
5. **VAST** (*Vector Alignment Search Tool*), a component of the National Center for Biotechnology Information (NCBI) Entrez system, reports a precomputed list of proteins of known structure that structurally resemble the query protein (“structure neighbors”). The VAST system uses the **Molecular Modeling Database (MMDB)**, an NCBI-generated database that is derived from PDB coordinates but in which molecules are represented by connectivity graphs rather than sets of atomic coordinates. VAST displays the superposition of the query protein in its structural alignment with up to five other proteins using **Cn3D** (a molecular graphics program that displays MMDB files and that is publicly available for a variety of computer platforms) or with only one other protein using Mage. VAST also reports a precomputed list of proteins that are similar to the query protein in sequence (“sequence neighbors”) and provides links from a selected protein to several bibliographic databases including MedLine.

SUMMARY

1. Four levels of structural complexity are used to describe the three-dimensional shapes of proteins.
2. The conformational flexibility of the peptide group is described by its ϕ and ψ torsion angles.
3. The α helix is a regular secondary structure in which hydrogen bonds form between backbone groups four residues apart. In the β sheet, hydrogen bonds form between the backbones of separate polypeptide segments.
4. Fibrous proteins are characterized by a single type of secondary structure: α keratin is a left-handed coil of two α helices, and collagen is a left-handed triple helix with three residues per turn.
5. Nonrepetitive structures include variations in regular secondary structures, turns, and loops.
6. The structures of proteins can be determined by X-ray crystallography and NMR spectroscopy.
7. The nonpolar side chains of a globular protein tend to occupy the protein's interior, whereas the polar side chains tend to define its surface.
8. Protein structures, which often contain common supersecondary structures (motifs), can be grouped into families according to their folding patterns. Structural elements are more likely to be evolutionarily conserved than are amino acid sequences.
9. The individual subunits of multisubunit proteins are usually symmetrically arranged.
10. Native protein structures are only slightly more stable than their denatured forms. The hydrophobic effect is the primary determinant of protein stability. Hydrogen bonding and ion pairing contribute relatively little to a protein's stability.
11. Studies of protein denaturation and renaturation indicate that the primary structure of a protein determines its three-dimensional structure.
12. Proteins have conformational flexibility that results in small molecular motions.
13. Proteins fold to their native conformations via directed pathways in which small elements of structure coalesce into larger structures.
14. Protein disulfide isomerases and molecular chaperones facilitate protein folding *in vivo*.
15. Diseases caused by protein misfolding include Alzheimer's disease, the transmissible spongiform encephalopathies (TSEs), and the amyloidoses.
16. The field of structural bioinformatics is concerned with the storage, visualization, analysis, and comparison of macromolecular structures.

REFERENCES

General

- Branden, C. and Tooze, J., *Introduction to Protein Structure* (2nd ed.), Garland Publishing (1999). [A well-illustrated book with chapters introducing amino acids and protein structure, plus chapters on specific proteins categorized by their structure and function.]
- Lesk, A.M., *Introduction to Protein Architecture*, Oxford University Press (2001).
- Milner-White, E.J., The partial charge of the nitrogen atom in peptide bonds, *Protein Science* **6**, 2477–2482 (1997). [Discusses the origin of the peptide N atom's partial negative charge.]
- Tanford, C. and Reynolds, J., *Nature's Robots*, Oxford University Press (2001). [A history of proteins.]

Fibrous Proteins

- Baum, J. and Brodsky, B., Folding of peptide models of collagen and misfolding in disease, *Curr. Opin. Struct. Biol.* **9**, 122–128 (1999).
- Kramer, R.Z., Bella, J., Mayville, P., Brodsky, B., and Berman, H.M., Sequence dependent conformational variations of collagen triple-helical structure, *Nature Struct. Biol.* **6**, 454–457 (1999).

Macromolecular Structure Determination

- McPherson, A., *Macromolecular Crystallography*, Wiley (2002).

McRee, D.E., *Practical Protein Crystallography* (2nd ed.), Elsevier Science (2002).

Rhodes, G., *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models* (2nd ed.), Academic Press (2000). [Includes overviews, methods, and discussions of model quality.]

Wider, G. and Wüthrich, K., NMR spectroscopy of large molecules and multimolecular assemblies in solution, *Curr. Opin. Struct. Biol.* **9**, 594–601 (1999).

Quaternary Structure

- Goodsell, D.S. and Olson, J., Structural symmetry and protein function, *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153 (2000).
- Sheinerman, F.B., Norel, R., and Honig, B., Electrostatic aspects of protein–protein interactions, *Curr. Opin. Struct. Biol.* **10**, 153–159 (2000).

Protein Stability

- Fersht, A., *Structure and Mechanism in Protein Science*, Chapter 11, Freeman (1999).
- Jaenicke, R. and Böhm, G., The stability of proteins in extreme environments, *Curr. Opin. Struct. Biol.* **8**, 738–748 (1998).
- Jones, S. and Thornton, J.M., Principles of protein–protein interactions, *Proc. Natl. Acad. Sci.* **93**, 13–20 (1996).

Protein Folding

- Baker, D. and Sali, A., Protein structure prediction and structural genomics, *Science* **294**, 93–96 (2001). [Summarizes the state of the art of protein structure prediction methods and how they can be applied.]
- Baldwin, R.L. and Rose, G.D., Is protein folding hierarchic? I. Local structure and peptide folding; and II. Folding intermediates and transition states, *Trends Biochem. Sci.* **24**, 26–33; and 77–83 (1999).
- Dobson, C.M. and Karplus, M., The fundamentals of protein folding: bringing together theory and experiment, *Curr. Opin. Struct. Biol.* **9**, 92–101 (1999).
- Hartl, F.U. and Hayer-Hartl, M., Molecular chaperones in the cytosol: from nascent chain to folded protein, *Nature* **295**, 1852–1858 (2002). [Discusses the importance of proper folding for newly synthesized proteins and reviews the major chaperone systems.]
- Horwich, A.R. (Ed.), Protein folding in the cell, *Adv. Prot. Chem.* **59** (2002).
- Rye, H.S., Roseman, A.M., Chen, S., Furtak, K., Fenton, W.A., Saibil, H.R., and Horwich, A.L., GroEL-GroES cycling: ATP and nonnative polypeptide direct alternation of folding-active rings, *Cell* **97**, 325–338 (1999).

Protein Misfolding Diseases

- Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L., Zurdo, J., Taddei, N., Ramponi, G., Dobson, C.M., and Stefani, M., Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases, *Nature* **416**, 507–511 (2002). [Provides evidence that a variety of misfolded proteins can form fibrous aggregates that can potentially damage cells.]
- Caughey, B., Interactions between prion protein isoforms: the kiss of death? *Trends Biochem. Sci.* **26**, 235–242 (2001).
- Jackson, G.S. and Clarke, A.R., Mammalian prion proteins, *Curr. Opin. Struct. Biol.* **10**, 69–74 (2000).
- Pruisner, S.B., Scott, M.R., DeArmond, S.J., and Cohen, F.E., Prion protein biology, *Cell* **93**, 337–348 (1998).

Structural Bioinformatics

- Bourne, P.E. and Weissig, H. (Eds.), *Structural Bioinformatics*, Wiley-Liss (2003).
- Hadley, C. and Jones, J.T., A systematic comparison of protein structure classifications: SCOP, CATH, and FSSP, *Structure* **7**, 1099–1112 (1999).
- Orengo, C.A., Todd, A.E., and Thornton, J.M., From protein structure to function, *Curr. Opin. Struct. Biol.* **9**, 374–382 (1999). [Describes how functional information can be derived by examining families of structurally related proteins.]

KEY TERMS

secondary structure	topology	contour map	hydropathy
tertiary structure	fibrous protein	supersecondary structure	ion pair (salt bridge)
quaternary structure	globular protein	(motif)	zinc finger
peptide group	coiled coil	$\beta\alpha\beta$ motif	breathing
trans conformation	denatured	β hairpin	cooperativity
cis conformation	native	$\alpha\alpha$ motif	chaotropic agent
backbone	β bulge	β barrel	renaturation
torsion (dihedral) angle	helix capping	dinucleotide-binding	hydrophobic collapse
ϕ	reverse turn	(Rossmann) fold	molten globule
ψ	(β bend)	domain	molecular chaperone
Ramachandran diagram	Ω loop	oligomer	heat shock protein
α helix	X-ray crystallography	protomer	ATPase
pitch	NMR	rotational symmetry	amyloid
antiparallel β sheet	diffraction pattern	cyclic symmetry	prion
parallel β sheet	electron density	dihedral symmetry	structural bioinformatics

STUDY EXERCISES

1. Explain why the conformational freedom of peptide bonds is limited.
2. What distinguishes regular and irregular secondary structures?
3. Describe the hydrogen-bonding pattern of an α helix.
4. Why are β sheets pleated?
5. What properties do fibrous proteins confer on substances such as hair, horns, bones, and tendons?
6. Why do turns and loops most often occur on the protein surface?
7. Which side chains usually occur in a protein's interior? On its surface?
8. Give some reasons why the number of possible protein structures is much less than the number of amino acid sequences.
9. List the advantages of multiple subunits in proteins.
10. Why can't proteins have mirror symmetry?
11. Describe the forces that stabilize proteins.
12. Describe the energy and entropy changes that occur during protein folding.

13. How does protein renaturation *in vitro* differ from protein folding *in vivo*?
14. What are amyloid fibrils, what is their origin, and why are they harmful?
15. What is structural bioinformatics?

PROBLEMS

1. Draw a *cis* peptide bond and identify the groups that experience steric interference.
2. Helices can be described by the notation n_m , where n is the number of residues per helical turn and m is the number of atoms, including H, in the ring that is closed by the hydrogen bond. (a) What is this notation for the α helix? (b) Is the 3_{10} helix steeper or shallower than the α helix?
3. Calculate the length in angstroms of a 100-residue segment of the α keratin coiled coil.
4. Hydrophobic residues usually appear at the first and fourth positions in the seven-residue repeats of polypeptides that form coiled coils. (a) Why do polar or charged residues usually appear in the remaining five positions? (b) Why is the sequence Ile-Gln-Glu-Val-Glu-Arg-Asp more likely than the sequence Trp-Gln-Glu-Tyr-Glu-Arg-Asp to appear in a coiled coil?
5. Globular proteins are typically constructed from several layers of secondary structure, with a hydrophobic core and a hydrophilic surface. Is this true for a fibrous protein such as α keratin?
6. The digestive tract of the larvae of clothes moths is a strongly reducing environment. Why is this beneficial to the larvae?
7. Describe the primary, secondary, tertiary, and quaternary structures of collagen.
8. Explain why gelatin, which is mostly collagen, is nutritionally inferior to other types of protein.
9. Is it possible for a native protein to be entirely irregular, that is, without α helices, β sheets, or other repetitive secondary structure?
10. (a) Is Trp or Gln more likely to be on a protein's surface? (b) Is Ser or Val less likely to be in the protein's interior? (c) Is Leu or Ile less likely to be found in the middle of an α helix? (d) Is Cys or Ser more likely to be in a β sheet?
11. What types of rotational symmetry are possible for a protein with (a) four or (b) six identical subunits?
12. You are performing site-directed mutagenesis to test predictions about which residues are essential for a protein's function. Which of each pair of amino acid substitutions listed below would you expect to disrupt protein structure the most? Explain.
 - (a) Val replaced by Ala or Phe.
 - (b) Lys replaced by Asp or Arg.
 - (c) Gln replaced by Glu or Asn.
 - (d) Pro replaced by His or Gly.
13. Laboratory techniques for randomly linking together amino acids typically generate an insoluble polypeptide, yet a naturally occurring polypeptide of the same length is usually soluble. Explain.
14. Given enough time, can all denatured proteins spontaneously renature?
15. Describe the intra- and intermolecular bonds/interactions that are broken or retained when collagen is heated to produce gelatin.
16. Under physiological conditions, polylysine assumes a random coil conformation. Under what conditions might it form an α helix?
17. It is often stated that proteins are quite large compared to the molecules they bind. However, what constitutes a large number depends on your point of view. Calculate the ratio of the volume of a hemoglobin molecule (65 kD) to that of the four O_2 molecules that it binds and the ratio of the volume of a typical office ($4 \times 4 \times 3$ m) to that of the typical (70-kg) office worker that occupies it. Assume that the molecular volumes of hemoglobin and O_2 are in equal proportions to their molecular masses and that the office worker has a density of 1.0 g/cm^3 . Compare these ratios. Is this the result you expected?
18. Which of the following polypeptides is most likely to form an α helix? Which is least likely to form a β strand?
 - (a) CRAGNRKIVLETY
 - (b) SEDNFGAPKSILW
 - (c) QKASVEMAVRNSG
 [Problem by Bruce Wightman, Muhlenberg College.]
19. The X-ray crystallographic analysis of a protein often fails to reveal the positions of the first few and/or the last few residues of a polypeptide chain. Explain.