The outer shell of the rhinovirus, cause of the common cold, consists of 60 copies each of three proteins (colored red, blue, and green), which self-assemble to form a 300-Å-diameter capsule that encloses the viral genome. [Image generated by the Virus Particle Explorer (http://viperdb.scripps.edu); based on an X-ray structure by Verdaguer N., Blaas D., and Fita I.; Institut de Biologia de Barcelona, Spain. PDBid 1FPN.]

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 convincing 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 80,000 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 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

KEY CONCEPTS

• The planar character of the peptide group limits the conformational flexibility of the polypeptide chain.
• The α helix and the β sheet allow the polypeptide chain to adopt favorable φ and ψ angles and to form hydrogen bonds.
• Fibrous proteins contain long stretches of regular secondary structure, such as the coiled coils in α keratin and the triple helix in collagen.
• Not all polypeptide segments form regular secondary structure such as α helices or β sheets.

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 Planar Peptide Group Limits Polypeptide Conformations

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

This explanation is supported by the observations that a peptide group's C—N bond is 0.13 Å shorter than its N—Cα single bond and that its C=O bond is 0.02 Å 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 Cα atoms are on opposite sides of the peptide bond joining them (Fig. 6-2). The cis conformation, in which successive Cα atoms are on the same side of the peptide bond, is ~8 kJ 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 ~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 be described by the torsion angles (also called dihedral angles or rotation angles) around the Cα—N bond (ϕ) and the Cα—C bond (ψ) of each residue.
These angles, $\phi$ and $\psi$, are both defined as 180° when the polypeptide chain is in its fully extended conformation and increase clockwise when viewed from C$_\alpha$.

The conformational freedom and therefore the torsion angles of a polypeptide backbone are sterically constrained. Rotation around the C$_\alpha$–N and C$_\alpha$–C bonds to form certain combinations of $\phi$ and $\psi$ angles will cause the amide hydrogen, the carbonyl oxygen, or the substituents of C$_\alpha$ of adjacent residues to collide (e.g., Fig. 6-5). Certain conformations of 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 $\phi$ and $\psi$ can be calculated. Sterically forbidden conformations, such as the one shown in Fig. 6-5, have $\phi$ and $\psi$ values that would bring atoms closer than their 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 $\phi$ and $\psi$) represent forbidden conformations of a polypeptide chain. Only three small regions of the diagram are physically accessible to most residues. The observed $\phi$ and $\psi$ 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 $\phi$ values to angles of around $-60^\circ$, making it, not surprisingly, the most conformationally restricted amino acid residue.

2. Gly, the only residue without a $C_\beta$ atom, is much less sterically hindered than the other amino acid residues. Hence, its permissible range of $\phi$ and $\psi$ covers a larger area of the Ramachandran diagram. At Gly residues, polypeptide chains often assume conformations that are forbidden to other residues.

**B The Most Common Regular Secondary Structures Are the $\alpha$ Helix and the $\beta$ 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 $\alpha$ helix and the $\beta$ sheet are such elements; they are called regular secondary structures because they are composed of sequences of residues with repeating $\phi$ and $\psi$ values.

**The $\alpha$ Helix Is a Coil.** Only one polypeptide helix has both a favorable hydrogen bonding pattern and $\phi$ and $\psi$ values that fall within the fully allowed regions of the Ramachandran diagram: the $\alpha$ helix. Its discovery by Linus Pauling in 1951, through model building, ranks as one of the landmarks of structural biochemistry (Box 6-1).
Linus Pauling and Structural Biochemistry

Linus Pauling (1901–1994) 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 most 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 (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 molecules 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 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 whether 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 began this career by using the then recently invented technique of X-ray crystallography. 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 the 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 aboveground 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 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 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 nth residue points along the helix axis toward the peptide N–H group of the (n + 4)th residue. This results in a strong hydrogen bond that has the nearly optimum N···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.

**FIG. 6-7** 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.

How many amino acid residues are in this helix? How many intrachain hydrogen bonds?

**FIG. 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 (gold) project away from the helix. This α helix is a segment of sperm whale myoglobin. [Based on an X-ray structure by Ilme Schlichting, Max Planck Institut für Molekulare Physiologie, Dortmund, Germany. PDBid 1A6M (for the definition of “PDBid” see Section 6-2E).]
Sheets Are Formed from Extended Chains. 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 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 Å.

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 seven-stranded antiparallel β sheet is shown in Fig. 6-11.

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.
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 sheet’s interchain hydrogen bonds. The geometry of a particular 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 sheet can be quite complex. The connection between two antiparallel strands may be

**FIG. 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.

\[?\] How many residues are in this β sheet? How many interchain hydrogen bonds?

**FIG. 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 magenta spheres. This seven-stranded antiparallel β sheet, which is shown with its polypeptide strands approximately horizontal, is from the jack bean protein concanavalin A. [Based on an X-ray structure by Gerald Edelman, The Rockefeller University. PDBid 2CNA.]

See Kinemage Exercise 3-3.

**FIG. 6-12** X-Ray structure of bovine carboxypeptidase A. The polypeptide backbone is drawn in ribbon form with α helices depicted as cyan coils, the strands of the β sheet represented by green arrows pointing toward the C-terminus, and its remaining portions portrayed by orange worms. Side chains are not shown. The eight-stranded β sheet forms a saddle-shaped curved surface with a right-handed twist. [Based on an X-ray structure by William Lipscomb, Harvard University. PDBid 3CPA.]
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).

**Turns Connect Some Units of Secondary Structure.** Polypeptide segments with regular secondary structure such as α helices or the strands of β sheets are often 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; Fig. 6-13a) almost always occur at protein surfaces. They usually involve four successive amino acid residues 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. 6-14). 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 Cβ 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.

**C Fibrous Proteins Have Repeating Secondary Structures**

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 Is a Coiled Coil. Keratin is a mechanically durable and relatively 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.
Humans have over 50 keratin genes that are expressed in a tissue-specific manner. The X-ray diffraction pattern of keratin resembles that expected for an helix (hence the name keratin). However, keratin exhibits a 5.1-Å spacing rather than the 5.4-Å distance corresponding to the pitch of the helix. This discrepancy is the result of two keratin polypeptides, each of which forms an helix, twisting around each other to form a left-handed coil. The normal 5.4-Å repeat distance of each 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 helix itself follows a helical path.

The conformation of 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 \(\alpha\) helix has 3.6 residues per turn, \(\alpha\) keratin's \(a\) and \(d\) residues line up along one side of each \(\alpha\) helix (Fig. 6-15a). The hydrophobic strip along one helix associates with the hydrophobic strip on another helix. Because the 3.5-residue repeat in keratin is slightly smaller than the 3.6 residues per turn of a standard \(\alpha\) helix, the two keratin helices are inclined about 18° relative to one another, resulting in the coiled coil arrangement (Fig. 6-15b). Coiled coils also occur in numerous other proteins, some of which are globular rather than fibrous.

The higher order structure of \(\alpha\) 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. 6-16). 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 macrofibrils.

\(\alpha\) Keratin is rich in Cys residues, which form disulfide bonds that cross-link adjacent polypeptide chains. The \(\alpha\) 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. Keratin, such as that in feathers, exhibits a β-like pattern in its native state.

Collagen Is 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. Vertebrates have 46 genetically distinct polypeptide chains that are assembled into 28 collagen varieties found in different tissues in the same individual. One of the most common collagens, called Type I, consists of two $\alpha_1$ chains and one $\alpha_2$ chain. It has a molecular mass of $\sim 285$ kD, a width of $\sim 14$ Å, and a length of $\sim 3000$ 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. The disease scurvy results from the dietary deficiency of vitamin C (Box 6-2).

![Collagen triple helix](image)

The amino acid sequence of a typical collagen polypeptide consists of monotonously repeating triplets of sequence Gly-X-Y over a segment of $\sim 1000$ 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. 6-17).

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 level along the triple helix axis (Fig. 6-18). The peptide groups are oriented such that the N\(\text{H}\) of each Gly makes a strong hydrogen bond with the carbonyl oxygen of an X (Pro) residue on a neighboring chain (Fig. 6-18). The bulky and relatively inflexible Pro and Hyp residues confer rigidity on the entire assembly.

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 whose diets were devoid of fresh foods on long voyages. 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 (see below). 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. 6-18 locally distorts the already internally crowded collagen helix. This ruptures the hydrogen bond from the backbone N\(\text{H}\) 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.

**FIG. 6-18** 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 stick model of the middle portion of the triple helix oriented with its N-termini at the bottom. The C atoms of the three chains are colored orange, magenta, and gray. The N and O atoms on all chains are blue and red. Note how the replacement of Gly with the bulkier Ala (C atoms 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 stick form (C atoms green). 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.
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. 6-17). Successive levels of fiber bundles in high-quality ropes and cables, as well as in other proteins such as keratin (Fig. 6-16), 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. 6-19. 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 Most Proteins Include Nonrepetitive Structure

The majority of proteins are globular proteins that, unlike the fibrous proteins discussed above, 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. 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.

Sequence Affects 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 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).
**FIG. 6-19**  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.

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

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

Tertiary Structure

KEY CONCEPTS
- X-Ray crystallography and NMR spectroscopy are used to determine the positions of atoms in proteins.
- Nonpolar residues tend to occur in the protein interior and polar residues on the exterior.
- A protein's tertiary structure consists of secondary structural elements that combine to form motifs and domains.
- Over time, a protein's structure is more highly conserved than its sequence.
- Bioinformatics databases store macromolecular structure coordinates. Software makes it possible to visualize proteins and compare their structural features.

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 its side chains. This information is deposited in a database and is readily available via the Internet, 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 Most Protein Structures Are Determined by X-Ray Crystallography or Nuclear Magnetic Resonance

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. 6-20) 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 detector or, now infrequently, on photographic film (Fig. 6-21). The X-rays used in structural studies are produced by laboratory X-ray generators or, now commonly, by synchrotrons, particle accelerators that produce 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 computer graphics as one or more sets of contours, in which a contour represents a specific level of electron density in the same way that a contour on a topographic map indicates locations that have a particular altitude. A portion of an electron density map of a protein is shown in Fig. 6-22.

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, 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 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 6-23 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. 6-23). 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. 6-22). Yet side chains of comparable size and shape, such as those of Leu, Ile, Thr, and Val, cannot always be differentiated (hydrogen atoms, having only one electron, are visible only if the resolution limit is less than ~1.2 Å). Consequently, a protein structure cannot be elucidated from its electron density map alone, but knowing the primary structure of the protein permits the sequence of amino acid residues to be fitted to the electron density map. Mathematical refinement can then reduce the uncertainty in the crystal structure's atomic positions to as little as 0.1 Å.

Most Crystalline Proteins Maintain Their Native Conformations. Does the structure of a protein in a crystal accurately reflect the structure of the protein in solution, 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.
2. In cases where different crystal forms of a protein have been analyzed, or when a crystal structure has been compared to a solution structure (determined by NMR; see below), the molecules have virtually identical conformations. Evidently, crystal packing forces do not greatly perturb the structures of protein molecules.
3. Many enzymes are catalytically active in the crystalline state. Since the activity of an enzyme is very sensitive to the positions of the groups involved in binding and catalysis (Chapter 11), the crystalline enzymes must have conformations that closely resemble their solution conformations.
Protein Structures Can Be Determined by NMR. The basis of nuclear magnetic resonance (NMR) is the observation that an atomic nucleus, such as a proton (a hydrogen nucleus), resonates in an applied magnetic field in a way that is sensitive to its electronic environment and its interactions with nearby nuclei. The development of NMR techniques, since the mid-1980s, in large part by Kurt Wüthrich, has made it possible to determine the three-dimensional structures of small globular proteins in aqueous solution.

A protein's conventional (one-dimensional) NMR spectrum is crowded with overlapping peaks, since even a small protein has hundreds of protons. This problem is addressed by two-dimensional (2D) NMR spectroscopy, which yields additional peaks arising from the interactions of protons that are less than 5 Å apart. Correlation spectroscopy (COSY) provides interatomic distances between protons that are covalently connected through one or two other atoms, such as the H atoms attached to the N and C\textsubscript{\textalpha} of the same amino acid (corresponding to the \(\phi\) torsion angle). Nuclear Overhauser spectroscopy (NOESY) provides interatomic distances for protons that are close in space although they may be far apart in the protein sequence. An example of a NOESY spectrum is shown in Fig. 6-24.

Interatomic distance measurements, along with knowledge of the protein's sequence and 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 cannot imply a unique structure but rather are consistent with an ensemble of closely related structures. Consequently, an NMR structure of a protein (or another macromolecule) is often presented as a sample of structures that are consistent with the data (e.g., Fig. 6-25). The “tightness” of a bundle of such structures is indicative both

**Fig. 6-24** NOESY spectrum of a protein. The diagonal represents the conventional one-dimensional NMR spectrum presented as a contour plot. Note that it is too crowded with peaks to be directly interpretable (even a small protein has hundreds of protons). The cross (off-diagonal) peaks each arise from the interaction of two protons that are <5 Å apart in space (their one-dimensional NMR peaks are located where horizontal and vertical lines intersect the diagonal). The line to the left of the spectrum represents the extended polypeptide chain with its \(N\)- and \(C\)-termini labeled \(N\) and \(C\) and the positions of four protons labeled \(a\) to \(d\). The dashed arrows indicate the diagonal NMR peaks to which these protons give rise. Cross peaks, such as \(i\), \(j\), and \(k\), each located at the intersection of the corresponding horizontal and vertical lines, show that two protons are <5 Å apart. These distance relationships are schematically drawn as three looped structures of the polypeptide chain below the spectrum. 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).]

**Fig. 6-25** The NMR structure of a protein. The drawing represents 20 superimposed structures of a 64-residue polypeptide comprising the Src protein SH3 domain (Section 13-2B). The polypeptide backbone (its connected \(C\)-atoms) is white, and its Phe, Tyr, and Trp side chains are yellow, red, and blue, respectively. The polypeptide backbone folds into two 3-stranded antiparallel \(\beta\) sheets that form a sandwich. [Courtesy of Stuart Schreiber, Harvard University.]
of the accuracy with which the structure is known, which in the most favor-
able cases is roughly comparable to that of an X-ray crystal structure with a
resolution of 2 to 2.5 Å, and of the conformational fluctuations that the pro-
tein undergoes (Section 6-4A). Although present NMR methods are limited
to determining the structures of macromolecules with molecular masses no
greater than \( \sim 100 \text{ kD} \), recent advances in NMR technology suggest that this
limit may soon increase to \( \sim 1000 \text{ kD} \) or more.

NMR methods, besides validating the structures of proteins analyzed by
X-ray crystallography (or in some cases identifying protein residues that are
perturbed by crystallization), 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 (Sections 6-4 and 6-5).

Proteins Can Be Depicted in Different Ways. The huge number of atoms
in proteins makes it difficult to visualize them using the same sorts of mod-
els employed for small organic molecules. Ball-and-stick representations show-
ing 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 in-
ternal details of the protein. Accordingly, computer-generated or artistic ren-
ditions (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 \( \text{C}_\alpha \) atoms or by representing helices as helical ribbons or cylin-
ders and \( \beta \) sheets as sets of flat arrows pointing from the N- to the C-termini.

B Side Chain Location Varies with Polarity

In the years since Kendrew solved the structure of myoglobin, nearly 80,000
protein structures have been reported. No two are exactly alike, but they ex-
hibit remarkable consistencies. The primary structures of globular proteins
generally lack the repeating sequences that support the regular conformations
seen in fibrous proteins. However, \textit{the amino acid side chains in globular pro-
teins 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 hy-
drophobic 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 lo-
cated 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 hydro-
gen bonded to other groups; in a sense, the formation of a hydrogen
   bond "neutralizes" their polarity. This is also the case with the polypep-
tide backbone.

These general principles of side chain distribution are evident in individ-
ual elements of secondary structure (Fig. 6-26) as well as in whole 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. Turns and loops joining second-
ary structural elements usually occur at the protein surface.

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
FIG. 6-26  Side chain locations in an α helix and a β sheet. In these space-filling models, the main chain is gray, nonpolar side chains are gold, 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. [Based on X-ray structures by Ilme Schlichting, Max Planck Institut für Molekulare Physiologie, Dortmund, Germany, and Gerald Edelman, The Rockefeller University. PDBids 1A6M and 2CNA.]

FIG. 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, 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.
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 Tertiary Structures Contain Combinations of Secondary Structure

Globular proteins—each with a unique tertiary structure—are built from combinations of secondary structural elements. The proportions of α helices and β sheets and the order in which they are connected provide an informative way of classifying and analyzing protein structure.

#### Certain Combinations of Secondary Structure Form Motifs.

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 \( β\alphaβ \) motif, in which an α helix connects two parallel strands of a β sheet (Fig. 6-28a).
2. Another common supersecondary structure, the β hairpin motif, consists of antiparallel strands connected by relatively tight reverse turns (Fig. 6-28b).
3. In an \( α\alpha \) 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. 6-28c). Similar associations stabilize the coiled coil conformation of α keratin and troponymosin (Fig. 6-15b), although their helices are parallel rather than antiparallel.
4. In the Greek key motif (Fig. 6-28d; 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.

#### Most Proteins Can Be Classified as \( α \), \( β \), or \( α/β \).

The major types of secondary structural elements occur in globular proteins in varying proportions and combinations. Some proteins, such as E. coli cytochrome \( b_{562} \) (Fig. 6-29a), consist only of α helices spanned by short connecting links and are therefore

---

**FIG. 6-28** Schematic diagrams of supersecondary structures. (a) A \( β\alphaβ \) motif, (b) a β hairpin motif, (c) an \( α\alpha \) motif, and (d) a Greek key motif, showing how it is constructed from a folded-over β hairpin. The polypeptide backbones are drawn as ribbons, with β strands shown as flat arrows pointing from N- to C-terminus, and α helices represented by cylinders.
classified as \(\alpha\) proteins. Others, such as immunoglobulins, contain the \textit{immunoglobulin fold} (Fig. 6-29\(b\)), and are called \(\beta\) proteins because they have a large proportion of \(\beta\) sheets and are devoid of \(\alpha\) helices. Most proteins, however, including lactate dehydrogenase (Fig. 6-29\(c\)) and carboxypeptidase A (Fig. 6-12), are known as \(\alpha/\beta\) proteins because they largely consist of mixtures of both types of secondary structure (proteins, on average, contain \(\sim31\%\) \(\alpha\) helix and \(\sim28\%\) \(\beta\) sheet).

The \(\alpha\), \(\beta\), and \(\alpha/\beta\) classes of proteins can be further subdivided by their topology, that is, according to how their secondary structural elements are connected. For example, extended \(\beta\) sheets often roll up to form \(\beta\) \textit{barrels}. Three different types of \(8\)-stranded \(\beta\) barrels, each with a different topology, are
Two of these (Fig. 6-30a, b) are all-\( \beta \)-structures containing multiple \( \beta \) hairpin motifs. The third, known as an \( \alpha/\beta \) barrel (Fig. 6-30c), can be considered as a set of overlapping \( \beta\alpha\beta \) motifs (and is a member of the \( \alpha/\beta \) class of proteins).

**Large Polypeptides Form Domains.** Polypeptide chains containing more than \( \sim 200 \) residues usually fold into two or more globular clusters known as domains, which give these proteins a bi- or multilobal appearance. Each subunit...
of the enzyme glyceraldehyde-3-phosphate dehydrogenase, for example, has two distinct domains (Fig. 6-31). Most domains consist of 40 to 200 amino acid residues and have an average diameter of \( \sim 25 \) Å. 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.

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.

Domains often have a specific function such as the binding of a small molecule. In Fig. 6-31, for example, the dinucleotide NAD\(^+\) (nicotinamide adenine dinucleotide; Fig. 11-4) binds to the N-terminal domain of glyceraldehyde-3-phosphate dehydrogenase. Michael Rossmann has shown that a \( \beta\alpha\beta\) unit, in which the \( \beta \) strands form a parallel sheet with \( \alpha \) helical connections, often acts as a nucleotide-binding site. Two of these \( \beta\alpha\beta\) units combine to form a domain known as a dinucleotide-binding fold, or Rossmann fold. Glyceraldehyde-3-phosphate dehydrogenase’s N-terminal domain contains such a fold, as does lactate dehydrogenase (Fig. 6-29c). In some multidomain proteins, binding sites 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 Structure Is Conserved More than Sequence**

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. Although it is estimated that there are as many as 1400 different protein domain families, approximately 200 different folding patterns account for about half of all known protein structures. As described in Section 5-4B, the domain is the fundamental unit of protein evolution. Apparently, the most common protein domains are 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 (Table 5-6) 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 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.

**E Structural Bioinformatics Provides Tools for Storing, Visualizing, and Comparing Protein Structural Information**

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 listed in Table 6-2 and described in Bioinformatics Project 3.

**The Protein Data Bank Is the Repository for Structural Information.** The atomic coordinates of nearly 80,000 macromolecular structures, including proteins, nucleic acids, and carbohydrates, 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.

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 includes the macromolecule's source (the organism from which it was obtained), the author(s) who determined the structure, key journal references, information about how the structure was determined, and 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 records (lines), each of which provides the three-dimensional (x, y, z) coordinates in angstroms of one atom in the structure. Each atom is identified by a serial number, an atom name (for example, C and O for an amino acid residue’s carbonyl C and O atoms, CA and CB for Cα and Cβ atoms), the name of the residue, and a letter to identify the chain to which it belongs (for structures that have more than one chain). For NMR-based structures, the PDB file contains a full set of records for each member of the ensemble of structures (the most representative member of such a coordinate set can be obtained from another database; see Table 6-2).

A particular PDB file may be located according to its PDBid or, if this is unknown, by searching with a variety of criteria including a protein’s name, its source, or the author(s). 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 (NDB) archives the atomic coordinates of structures that contain nucleic acids, using roughly the same format as PDB files.

**Molecular Graphics Programs Interactively Show Macromolecules in Three Dimensions.** The most informative way to examine a macromolecular structure is through the use of molecular graphics programs that 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. Most molecular graphics programs use PDB files as input. The programs described here can be downloaded from

### TABLE 6-2 Structural Bioinformatics Internet Addresses

<table>
<thead>
<tr>
<th>Structural Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Data Bank (PDB): <a href="http://www.rcsb.org/">http://www.rcsb.org/</a></td>
</tr>
<tr>
<td>Nucleic Acid Database: <a href="http://ndbserver.rutgers.edu/">http://ndbserver.rutgers.edu/</a></td>
</tr>
<tr>
<td>Most Representative NMR Structure in an Ensemble: <a href="http://www.ebi.ac.uk/msd-srv/olderado">http://www.ebi.ac.uk/msd-srv/olderado</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Molecular Graphics Programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>FirstGlance: <a href="http://molvis.sdc.edu/fgij/index.htm">http://molvis.sdc.edu/fgij/index.htm</a></td>
</tr>
<tr>
<td>Jmol: <a href="http://jmol.sourceforge.net/">http://jmol.sourceforge.net/</a></td>
</tr>
<tr>
<td>KiNG: <a href="http://kinemage.biochem.duke.edu/software/king.php">http://kinemage.biochem.duke.edu/software/king.php</a></td>
</tr>
<tr>
<td>Proteopedia: <a href="http://www.proteopedia.org/">http://www.proteopedia.org/</a></td>
</tr>
<tr>
<td>Swiss-Pdb Viewer (DeepView): <a href="http://spdbv.vital-it.ch/">http://spdbv.vital-it.ch/</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structural Classification Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE (Combinatorial Extension of optimal pathway): <a href="http://cl.sdsc.edu/">http://cl.sdsc.edu/</a></td>
</tr>
<tr>
<td>Pfam (protein families): <a href="http://pfam.sanger.ac.uk/">http://pfam.sanger.ac.uk/</a> or <a href="http://pfam.janelia.org/">http://pfam.janelia.org/</a></td>
</tr>
<tr>
<td>SCOP (Structural Classification Of Proteins): http:// scop.mrc-lmb.cam.ac.uk/scop/</td>
</tr>
</tbody>
</table>
the Internet addresses listed in Table 6-2, some of which also provide instructions for the program's use.

**Jmol**, which functions as both a Web browser-based applet or as a standalone program, allows the user to display user-selected macromolecules in a variety of colors and formats (e.g., wire frame, ball-and-stick, backbone, space-filling, and cartoons). The Interactive Exercises on the website that accompanies this textbook (http://www.wiley.com/college/voet/) all use Jmol. FirstGlance uses Jmol to display macromolecules via a user-friendly interface. **KiNG**, which also has Web browser-based and standalone versions, displays the so-called Kinemages on this textbook's accompanying website. KiNG provides a generally more author-directed user environment than does Jmol. Macromolecules can be displayed directly from their corresponding PDB page using either Jmol or KiNG. The **Swiss-PDB Viewer** (also called Deep View), in addition to displaying molecular structures, 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. **Proteopedia** is a 3D interactive encyclopedia of proteins and other macromolecules that resembles Wikipedia in that it is user edited. It uses mainly Jmol as a viewer.

**Structure Comparisons Reveal Evolutionary Relationships.** Most proteins are structurally related to other proteins, since evolution tends to conserve the structures of proteins rather than their sequences. The computational tools described below facilitate the classification and comparison of protein structures. These programs can be accessed directly via their Internet addresses and in some cases 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 **C**lass, **A**rchitecture, **T**opology, and **H**omologous 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 levels of gross secondary structure (Mainly α, Mainly β, α/β, and Few Secondary Structures); (2) Architecture, the gross arrangement of secondary structure; (3) Topology, which depends on both the overall shape of the protein domain and the connectivity of its secondary structural elements; and (4) Homologous Superfamily, which identifies the protein as a member of a group that shares a common ancestor.

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. CE can optimally align and display two user-selected structures.

3. **Pfam** (for **P**rotein **fam**ilies) is a database of nearly 11,000 multiple sequence alignments of protein domains (called Pfam families). Using Pfam, one can analyze a protein for Pfam matches (74% of proteins have at least one match in Pfam), determine the domain organization of a protein based on its sequence or its structure, examine the phylogenetic tree of a Pfam family, and view the occurrence of a protein's domains across different species.

4. **SCOP** (Structural Classification Of Proteins) classifies protein structures based mainly on manually generated topological considerations according to a six-level hierarchy: Class (e.g., all-α, all-β, α/β), Fold (based on the arrangement 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 the molecular graphics program **Cn3D**. VAST also reports a precomputed list of proteins that are similar to the query protein in sequence (“sequence neighbors”).

### 3 Quaternary Structure and Symmetry

**KEY CONCEPT**

- Some proteins contain multiple subunits, usually arranged symmetrically.

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 7-1D 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

**FIG. 6-33 Quaternary structure of hemoglobin.** In this space-filling model, the $\alpha_1$, $\alpha_2$, $\beta_1$, and $\beta_2$ 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, Cambridge, U.K. PDBid 2DHJ.]
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.

In the simplest type of rotational symmetry, cyclic symmetry, protomers are related by a single axis of rotation (Fig. 6-34a). Objects with 2-, 3-, 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 2-fold rotation axis at right angles (Fig. 6-34b). An oligomer with \( D_n \) symmetry consists of \( 2n \) protomers. \( D_5 \) 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 (see page 127) are built on these geometric plans.

**CHECKPOINT**

- List the advantages of multiple subunits in proteins.
- Why can’t proteins have mirror symmetry?

**4 Protein Stability**

**KEY CONCEPTS**

- Protein stability depends primarily on hydrophobic effects and secondarily on electrostatic interactions.
- A protein that has been denatured may undergo renaturation.
- Protein structures are flexible and may include unfolded regions.

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 protein structure is the result of a delicate balance among powerful countervailing forces.

**A Proteins Are Stabilized by Several Forces**

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

**The Hydrophobic Effect Has the Greatest Influence on Protein Stability.** 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 6-3). 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 Contribute to Protein Stability.** 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 only act over 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 (\( -2 \text{ to } 8 \text{ kJ} \cdot \text{mol}^{-1} \) 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 6-3 Hydropathy Scale for Amino Acid Side Chains**

<table>
<thead>
<tr>
<th>Side Chain</th>
<th>Hydropathy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ile</td>
<td>4.5</td>
</tr>
<tr>
<td>Val</td>
<td>4.2</td>
</tr>
<tr>
<td>Leu</td>
<td>3.8</td>
</tr>
<tr>
<td>Phe</td>
<td>2.8</td>
</tr>
<tr>
<td>Cys</td>
<td>2.5</td>
</tr>
<tr>
<td>Met</td>
<td>1.9</td>
</tr>
<tr>
<td>Ala</td>
<td>1.8</td>
</tr>
<tr>
<td>Gly</td>
<td>–0.4</td>
</tr>
<tr>
<td>Thr</td>
<td>–0.7</td>
</tr>
<tr>
<td>Ser</td>
<td>–0.8</td>
</tr>
<tr>
<td>Trp</td>
<td>–0.9</td>
</tr>
<tr>
<td>Tyr</td>
<td>–1.3</td>
</tr>
<tr>
<td>Pro</td>
<td>–1.6</td>
</tr>
<tr>
<td>His</td>
<td>–3.2</td>
</tr>
<tr>
<td>Glu</td>
<td>–3.5</td>
</tr>
<tr>
<td>Gln</td>
<td>–3.5</td>
</tr>
<tr>
<td>Asp</td>
<td>–3.5</td>
</tr>
<tr>
<td>Asn</td>
<td>–3.5</td>
</tr>
<tr>
<td>Lys</td>
<td>–3.9</td>
</tr>
<tr>
<td>Arg</td>
<td>–4.5</td>
</tr>
</tbody>
</table>


**FIG. 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).]
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.

Disulfide Bonds Cross-Link Extracellular Proteins. 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 Stabilize Some Small Domains. 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 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 Proteins Can Undergo 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,

\[
\begin{align*}
\text{Guanidinium ion} & \quad \text{Urea} \\
\text{NH}_2^+ & \quad \text{O} \\
\text{H}_2\text{N} - \text{C} - \text{NH}_2 & \quad \text{H}_2\text{N} - \text{C} - \text{NH}_2
\end{align*}
\]

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

---

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 archaeabacterium *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 ~100 kJ · 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 of 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 ~0.4 kJ · mol⁻¹ of amino acid residues) is an essential property that has arisen through natural selection. Perhaps this marginal stability helps confer the structural flexibility that many proteins require to carry out their physiological functions.
effectiveness as denaturants stems from their ability to disrupt hydrophobic interactions, although their mechanism of action is not well understood.

Many 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-mercaptopethanol. 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. 6-38). The protein must therefore renature spontaneously.

The renaturation of RNase A demands that its four disulfide bonds re-form. 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. Thus 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 catalytically 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 \( \sim 1\% \) active (Fig. 6-38, 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.

C Proteins Are Dynamic

The static way that protein 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. 6-39). Conformational flexibility, or breathing, with structural displacement of up to \( \sim 2 \) Å, allows small molecules to diffuse in and out of the interior of certain proteins. In some cases, a protein’s conformational flexibility includes two stable alternatives in dynamic equilibrium. A change in cellular conditions, such as pH or oxidation state, or the presence of a binding partner can tip the balance toward one conformation or the other.

Some Proteins Contain Unfolded Regions. An entire protein or a long polypeptide segment (>30 residues) may lack defined structure in its native state. Such intrinsically disordered proteins are characterized by sequences rich in certain polar and charged amino acids (Gln, Ser, Pro, Glu, Lys, Gly, and Ala) and lacking in bulky hydrophobic groups (Val, Leu, Ile, Met, Phe, Trp, and Tyr). Sequence analysis suggests that approximately 33% of eukaryotic proteins may contain long disordered segments, whereas only a few percent of prokaryotic proteins do.

Most intrinsically disordered proteins adopt a specific secondary or tertiary structure when they bind to some other molecule. For example, the transcription factor known as CREB (cyclic AMP response element–binding protein) is disordered when free in solution but folds to an ordered confor-

**FIG. 6-39** Molecular dynamics of myoglobin. Several “snapshots” of the protein calculated at intervals of \( 5 \times 10^{-12} \) 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]
**FIG. 6-40** Conformational change in a CREB domain. The backbone of the kinase-inducible domain (KID) of CREB is drawn as a pink worm. It is unstructured when free in solution (left) but becomes two perpendicular helices when it forms a complex with the KID-binding domain of another protein (shown with its solvent-accessible surface in gray, right). The KID residues Ser 133, which has a covalently attached phosphoryl group, and Leu 141 are drawn in ball-and-stick form with C green, O red, and P yellow. [Courtesy of Peter Wright, Scripps Research Institute, La Jolla, California. PDBid 1KDX.]

**CHECKPOINT**

- Describe the hydropathic index plot for a fibrous protein such as collagen or keratin.
- Describe the forces that stabilize proteins, and rank their relative importance.
- Summarize the results of Anfinsen’s experiment with RNase A.
- Why would it be advantageous for a protein or a segment of a protein to lack defined secondary or tertiary structure?

5 Protein Folding

**KEY CONCEPTS**

- A folding protein follows a pathway from high energy and high entropy to low energy and low entropy.
- Protein disulfide isomerase catalyzes disulfide bond formation.
- A variety of molecular chaperones assist protein folding via an ATP-dependent bind-and-release mechanism.
- Amyloid diseases result from protein misfolding.
- The misfolded proteins form fibrils containing extensive \( \beta \) structure.

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, \( \phi \) and \( \psi \), each have three stable conformations. This yields \( 3^{2n} = 10^n \) 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

\[
    t = \frac{10^{10}}{10^{13}}
\]

For a small protein of 100 residues, \( t = 10^{87} \) s, which is immensely greater than the apparent age of the universe (~13.7 billion years = \( 4.3 \times 10^{17} \) s). Clearly, proteins must fold more rapidly than this.

A Proteins Follow 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 conformational 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. 6-41.

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 native-like 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 rearrangements 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. 6-42, 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

---

**Fig. 6-41** 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).]

**Fig. 6-42** 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).]
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 millions of proteins that are known only from their sequences (Box 6-4).

**Protein Disulfide Isomerase Acts during Protein Folding.** 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, and 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.

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. 6-43a). Another disulfide group

![FIG. 6-43](image)

**FIG. 6-43  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

Around 7 million polypeptide sequences are known, yet the structures of only ~80,000 proteins have been determined. 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 X-ray 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 propensities 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). The most successful of these ab initio (from the beginning) methods is the Rosetta program formulated by David Baker. To satisfy the program’s computational needs, a volunteer network of ~100,000 computers, known as Rosetta@home, provides the 500,000 or so hours of processing time required to generate a structure. Two examples of successful protein structure prediction by Rosetta appear above, marked (a) and (b). The predicted model (gray) for each of these bacterial proteins is superimposed on the experimentally determined X-ray structure, colored in rainbow order from N-terminus (blue) to C-terminus (red) with core side chains drawn as sticks.

Protein design, the experimental inverse of protein structure prediction, has provided insights into protein folding and stability. Protein design attempts to construct an amino acid sequence that will form a structure such as a sandwich of β sheets or a bundle of α helices. The designed polypeptide is then chemically or biologically synthesized, and its structure is determined. 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 (beginning anew) protein design, accomplished by Stephen Mayo, was for a 28-residue ββα motif that has a backbone conformation designed to resemble a zinc finger (Fig. 6-37) but that 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 $1.9 \times 10^{27}$, representing $1.1 \times 10^{25}$ 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 (blue) was nearly superimposable on that of a known zinc finger motif (red). Although FSD-1 is relatively small, it folds into a unique stable structure, thereby demonstrating the power of protein design techniques.

[Figures courtesy of Gautam Dantas, Washington University School of Medicine; PDBids 1WHZ and 2HH6; and Stephen Mayo, California Institute of Technology]
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. 6-43b). In this case, the reduced PDI reaction product must be reoxidized by cellular oxidizing agents in order to repeat the process.

**B Molecular Chaperones Assist Protein Folding**

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 exposed hydrophobic segments that might lead to nonnative 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.

**Most Chaperones Require ATP.** There are several classes of molecular chaperones in both prokaryotes and eukaryotes, including the following:

1. The Hsp70 family of proteins are highly conserved 70-kD proteins in both prokaryotes and eukaryotes. In association with the co-chaperone protein Hsp40, they facilitate the folding of newly synthesized proteins and reverse the denaturation and aggregation of proteins. Hsp70 proteins also function to unfold proteins in preparation for their transport through membranes (Section 9-4D) and to subsequently refold them.

2. **Trigger factor** is a ribosome-associated chaperone in prokaryotes that prevents the aggregation of polypeptides as they emerge from the ribosome (Section 27-5A). Trigger factor and Hsp70 are the first chaperones a newly made prokaryotic protein encounters. Subsequently, many partially folded proteins are handed off to other chaperones to complete the folding process. Eukaryotes lack trigger factor but contain other small chaperones that have similar functions.

3. The **chaperonins** form large, multisubunit, cagelike assemblies in both prokaryotes and eukaryotes. They bind improperly folded proteins and induce them to refold inside an internal cavity (see below).

4. The **Hsp90** proteins are eukaryotic proteins that mainly facilitate the late stages of folding of proteins involved in cellular signaling (Chapter 13). Hsp90 proteins are among the most abundant proteins in eukaryotes, accounting for up to 6% of cellular protein under stressful conditions that destabilize proteins.

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. Most molecular chaperones are ATPases, that is, enzymes that catalyze the
hydrolysis of ATP (adenosine triphosphate) to ADP (adenosine diphosphate) and P\(_i\) (inorganic phosphate):

\[
\text{ATP} + \text{H}_2\text{O} \rightarrow \text{ADP} + \text{P}_i
\]

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

**The GroEL/ES Chaperonin Forms Closed Chambers in Which Proteins Fold.** The chaperonins in *E. coli* consist of two types of subunits named GroEL and GroES. The X-ray structure of a GroEL–GroES–(ADP)_7 complex (Fig. 6-44), determined by Arthur Horwich and Paul Sigler, reveals fourteen identical 549-residue GroEL subunits arranged in two stacked rings of seven subunits each. This complex is capped at one end by a domelike heptameric ring of 97-residue GroES subunits to form a bullet-shaped complex with \(C_7\) symmetry. The two GroEL rings each enclose a central chamber with a diameter of \( \sim 45 \) Å in which partially folded proteins fold to their native conformations. A barrier in the center of the complex (Fig. 6-44c) prevents a folding protein from passing between the two GroEL chambers. 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 Drive the Conformational Changes in GroEL/ES.** Each GroEL subunit has a binding pocket for ATP that catalyzes the hydrolysis of its bound ATP to ADP and \(\text{P}_i\). When the cis ring subunits hydrolyze their bound ATP molecules and release the product \(\text{P}_i\), the protein undergoes a conformational change that widens and elongates the cis inner cavity so as to more than double its volume from 85,000 Å\(^3\) to 175,000 Å\(^3\). (In the structure shown in Fig. 6-44, the cis ring has already hydrolyzed its seven molecules of ATP to ADP.) The expanded cavity can enclose a partially folded substrate protein of at least 70 kD. All seven subunits of the GroEL ring act

---

**FIG. 6-44** X-Ray structure of the GroEL–GroES–(ADP)_7 complex. (a) A space-filling drawing as viewed perpendicularly to the complex’s sevenfold axis with the GroES ring orange, the cis ring of GroEL green, and the trans ring of GroEL red with one subunit of each ring shaded more brightly. The dimensions of the complex are indicated. Note the different conformations of the two GroEL rings. The ADPs, whose binding sites are in the base of each cis ring GroEL subunit, are not seen because they are surrounded by protein. (b) As in Part a but viewed along the sevenfold axis. (c) As in Part a but with the two GroEL subunits closest to the viewer in both the cis and trans rings removed to expose the interior of the complex. The level of fog increases with the distance from the viewer. Note the much larger size of the cavity formed by the cis ring and GroES in comparison to that of the trans ring. [Based on an X-ray structure by Paul Sigler, Yale University. PDBid 1AON.]
in concert; that is, they are mechanically linked such that they 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. 6-45):

1. One GroEL ring that has bound 7 ATP also binds an improperly folded substrate protein, which associates with hydrophobic patches that line the inner wall of the GroEL chamber. The GroES cap then binds to the GroEL ring like a lid on a pot, inducing a conformational change in the resulting cis ring that buries the hydrophobic patches, thereby depriving the substrate protein of its binding sites. This releases the substrate protein into the now enlarged and closed cavity, where it commences folding. The cavity, which is now lined only with hydrophilic groups, provides the substrate protein with an isolated microenvironment that prevents it from nonspecifically aggregating with other misfolded proteins. Moreover, the conformational change that buries GroEL's hydrophobic patches stretches and thereby partially unfolds the improperly folded substrate protein before it is released. This rescues the substrate protein from a local energy minimum in which it had become trapped (Fig. 6-42), thereby permitting it to continue its conformational journey down the folding funnel toward its native state (the state of lowest free energy).

2. Within, ~10 s (the time the substrate protein has to fold), the cis ring catalyzes the hydrolysis of its 7 bound ATPs to ADP + P, and the P_i

3. A second molecule of improperly folded substrate protein binds to the trans ring, followed by 7 ATP.

4. The cis ring releases GroES cap, ADP, and a better-folded substrate protein.

5. GroES cap binds to the GroEL ring. This induces a conformational change in the now cis ring, which releases the improperly folded protein into the enlarged cavity where it commences to refold.

6. The trans ring with its bound ATP now becomes the cis ring.

7. Within ~10 s the cis ring catalyzes the hydrolysis of its 7 bound ATPs and releases the resulting P_i, which weakens the interactions binding GroES to GroEL.

FIG. 6-45 Reaction cycle of the GroEL/ES chaperonin. The protein complex is colored as in Fig. 6-44. See the text for an explanation.
is released. The absence of ATP’s γ phosphate group weakens the interactions that bind GroES to GroEL.

3. A second molecule of improperly folded 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). Typically, only ~5% of substrate proteins fold to their native state in each reaction cycle. Thus, to fold half the substrate protein present would require \( \log(1 - 0.5) = \log(1 - 0.05) \approx 14 \) reaction cycles and hence \( 7 \times 14 = 98 \) 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.

Experiments indicate that the GroEL/ES system interacts with only a subset of E. coli proteins, most with molecular masses in the range 20 to 60 kD. These proteins tend to contain two or more α/β domains that mainly consist of open β sheets. Such proteins are expected to fold only slowly to their native state because the formation of hydrophobic sheets requires a large number of specific long-range interactions. Proteins dissociate from GroEL/ES after folding, but some frequently revisit the chaperonin, apparently because they are structurally labile or prone to aggregate and must return to GroEL for periodic maintenance.

Eukaryotic cells contain the chaperonin TRiC, with double rings of eight nonidentical subunits, each of which resembles a GroEL subunit. However, the TRiC proteins contain an additional segment that acts as a built-in lid, so the complex encloses a polypeptide chain and mediates protein folding without the assistance of a GroES-like cochaperone. Like its bacterial counterpart, TRiC operates in an ATP-dependent fashion. Around 10% of eukaryotic proteins transiently interact with TRiC.

### Some Diseases Are 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 21-1). However, at least 35 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 (Table 6-4). The aggregates are known as amyloids, a term that means starchlike because it was originally thought that the material resembled starch.

The diseases known as amyloidoses are a set of relatively rare inherited diseases in which mutant forms of normally occurring proteins [e.g., lysozyme, an enzyme that hydrolyzes bacterial cell walls (Section 11-4), and fibrinogen, a blood plasma protein that is the precursor of fibrin, which forms...
blood clots (Box 11-4)] accumulate in a variety of tissues as amyloids. The symptoms of amyloidoses usually do not become apparent until the third to seventh decade of life and typically progress over 5 to 15 years, ending in death.

**Amyloid-β Protein Accumulates in Alzheimer’s Disease.** Alzheimer’s disease, a neurodegenerative condition that strikes mainly the elderly, causes devastating mental deterioration and eventual death (it affects ~10% of those over 65 and ~50% of those over 85). It is characterized by brain tissue containing abundant amyloid plaques (deposits) surrounded by dead and dying neurons (Fig. 6-46). The amyloid plaques consist mainly of fibrils of a 40- to 42-residue protein named amyloid-β protein (Aβ). Aβ is a fragment of a 770-residue membrane protein called the Aβ precursor protein (BPP), whose normal function is unknown. Aβ is excised from BPP in a multistep process through the actions of two proteolytic enzymes dubbed β- and γ-secretases. The neurotoxic effects of Aβ begin even before significant amyloid deposits appear (see below).

The age dependence of Alzheimer’s disease suggests that Aβ deposition is an ongoing process. Indeed, several rare mutations in the BPP gene that increase the rate of Aβ production result in the onset of Alzheimer’s disease as early as the fourth decade of life. A similar phenomenon occurs 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. These individuals invariably develop Alzheimer’s disease by their 40th year because the gene encoding BPP is located on chromosome 21 and hence individuals with Down’s syndrome produce BPP and presumably Aβ 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 Aβ production.

**Prion Diseases Are Infectious.** Certain 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 spongillike 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 Prusiner demonstrated for scrapie, the infectious agent is a protein called a prion (for proteaceous 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. 6-47). These fibrils presumably form the amyloid plaques that accompany the neuronal degeneration in TSEs.

How are prion diseases transmitted? PrP is the product of a normal cellular gene that has no known function (genetically engineered mice that fail to express PrP appear to be normal). Infection of cells by prions somehow...
alters the PrP protein. Various methods have demonstrated that the scrapie form of PrP (PrP<sup>Sc</sup>) is identical to normal cellular PrP (PrP<sup>C</sup>) in sequence but differs in secondary and/or tertiary structure. This suggests that PrP<sup>Sc</sup> induces PrP<sup>C</sup> to adopt the conformation of PrP<sup>Sc</sup>, that is, a small amount of PrP<sup>Sc</sup> triggers the formation of additional PrP<sup>Sc</sup> from PrP<sup>C</sup>, which triggers more PrP<sup>Sc</sup> to form, and so on. This accounts for the observation that mice that do not express the gene encoding PrP cannot be infected with scrapie.

Human PrP<sup>C</sup> consists of a disordered (and hence unseen) 98-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. 6-48a). Unfortunately, the insolubility of PrP<sup>Sc</sup> has precluded its structural determination, but spectroscopic methods indicate that it has a lower α helix content and a higher β sheet content than PrP<sup>C</sup>. This suggests that the protein has refolded (Fig. 6-48b). The high β sheet content of PrP<sup>Sc</sup> presumably facilitates the aggregation of PrP<sup>Sc</sup> as amyloid fibrils (see below).

Prion diseases can be transmitted by the consumption of nerve tissue from infected individuals, as illustrated by the incidence of BSE. This disease was unknown before 1985 but reached epidemic proportions among cattle in the U.K. in 1993. The rise in BSE reflects the practice, beginning in the 1970s, of feeding cattle preparations of meat and bone meal that were derived from other animals by a method that failed to inactivate prions. The BSE epidemic abated due to the banning of such feeding in 1988, together with the slaughter of a large number of animals at risk for having BSE. However, it is now clear that BSE was transmitted to humans who ate meat from BSE-infected cattle: Some 200 cases of so-called new variant CJD 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. 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.

![Prion protein conformations. (a) The NMR structure of human prion protein (PrP<sup>C</sup>). The protein, missing its first 23 residues, is drawn in ribbon form with helices red, β sheets green, and other segments orange. Its disulfide bond is shown in yellow. (b) A plausible model for the structure of PrP<sup>Sc</sup> represented as in Part a. Note the formation of structure in what was the flexibly disordered N-terminal region. [Courtesy of Fred Cohen, University of California at San Francisco. Part a based on an NMR structure by Kurt Wüthrich, Eidgenössische Technische Hochschule, Zurich, Switzerland. PDBid 1QLX.](image-url)
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. Spectroscopic analysis of amyloid fibrils indicates that they are rich in β structure, with individual β strands oriented perpendicular to the fiber axis (Fig. 6-49). 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.

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 temperatures). This suggests that the partially unfolded, aggregation-prone forms are in 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 ⇌ U is governed by Eq. 1-17: $K_{eq} = [U]/[N] = e^{\Delta G^{\circ}/RT}$, 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 fibril formation 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. Since most amyloid diseases require several decades to become symptomatic, the development of an amyloid nucleus must be a rare event. Once an amyloid fiber begins to grow, however, its development is more rapid.

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. After it has formed, an amyloid fibril is virtually indestructible under physiological conditions, possibly due to the large number of main-chain hydrogen bonds that must be broken in order to separate individual polypeptide strands (side chain interactions are less important in stabilizing β sheets). 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.

Are fibrillar deposits directly responsible for the neurodegeneration seen in many amyloid diseases? A growing body of evidence suggests that cellular damage begins when the misfolded proteins first aggregate but are still soluble. For example, in mouse models of Alzheimer’s disease, cognitive impairment is evident before amyloid plaques develop. Other experiments show that the most infectious prion preparations contain just 14–28 PrPSc molecules, that is, a nucleus for a fibril, not the fibril itself. Even a modest number of misfolded protein molecules could be toxic if they prevented the cell’s chaperones from assisting other more critical proteins to fold. The appearance of extracellular—and sometimes intracellular—amyloid fibrils may simply represent the accumulation of protein that has overwhelmed the cellular mechanisms that govern protein folding or the disposal of misfolded proteins.
Summary

1 Secondary Structure
• Four levels of structural complexity are used to describe the three-dimensional shapes of proteins.
• The conformational flexibility of the peptide group is described by its $\phi$ and $\psi$ torsion angles.
• The $\alpha$ helix is a regular secondary structure in which hydrogen bonds form between backbone groups four residues apart. In the $\beta$ sheet, hydrogen bonds form between the backbones of separate polypeptide segments.
• Fibrous proteins are characterized by a single type of secondary structure: $\alpha$ keratin is a left-handed coil of two $\alpha$ helices, and collagen is a left-handed triple helix with three residues per turn.

2 Tertiary Structure
• The structures of proteins have been determined mainly by X-ray crystallography and NMR spectroscopy.
• The nonpolar side chains of a globular protein tend to occupy the protein’s interior; the polar side chains tend to define its surface.
• Protein structures can be classified on the basis of motifs, secondary structure content, topology, or domain architecture. Structural elements are more likely to be evolutionarily conserved than are amino acid sequences.
• Structural bioinformatics is concerned with the storage, visualization, analysis, and comparison of macromolecular structures.

3 Quaternary Structure and Symmetry
• The individual subunits of multisubunit proteins are usually symmetrically arranged.

4 Protein Stability
• 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.
• Studies of protein denaturation and renaturation indicate that the primary structure of a protein determines its three-dimensional structure.

5 Protein Folding
• Proteins fold to their native conformations via directed pathways in which small elements of structure coalesce into larger structures.
• Molecular chaperones facilitate protein folding in vivo by repeatedly binding and releasing a polypeptide in an ATP-dependent manner and providing it with an isolated microenvironment in which to fold.
• Diseases caused by protein misfolding include the amyloidoses, Alzheimer’s disease, and the transmissible spongiform encephalopathies (TSEs).

Key Terms
secondary structure 127
tertiary structure 127
quaternary structure 128
peptide group 129
trans conformation 129
cis conformation 129
backbone 129
torsion (dihedral) angle 129
$\phi$ 130
$\psi$ 130
Ramachandran diagram 130
$\alpha$ helix 131
regular secondary structure 131
pitch 133
antiparallel $\beta$ sheet 134
parallel $\beta$ sheet 134
topology 135
reverse turn ($\beta$ bend) 136
fibrous protein 136
globular protein 136
coiled coil 137
random coil 140
denaturation 140
native structure 140
$\beta$ bulge 140
helix cap 140
X-ray crystallography 142
diffraction pattern 142
electron density 142
contour map 143
NMR 145
supersecondary structure 148
$\beta\beta$ motif 148
$\beta$ hairpin 148
$\alpha\alpha$ motif 148
$\beta$ barrel 149
$\alpha/\beta$ barrel 150
domain 150
dinucleotide-binding (Rossmann) fold 150
structural bioinformatics 152
oligomer 155
protomer 155
rotational symmetry 156
cyclic symmetry 156
dihedral symmetry 156
hydropathy 157
ion pair (salt bridge) 158
zinc finger 158
cooperativity 159
renaturation 160
breathing 161
intrinsically disordered protein 161
hydrophobic collapse 163
molten globule 163
homology modeling 165
threading 165
ab initio 165
molecular chaperone 166
heat shock protein 166
ATPase 166
amyloid 169
Alzheimer’s disease 170
Creutzfeldt-Jakob disease 170
transmissible spongiform encephalopathies (TSEs) 170
prion 170

Problems
1. Draw a cis peptide bond and identify the groups that experience steric interference.
2. How many peptide bonds are shown in the structure drawn in Fig. 6-7?
3. 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 $\alpha$ helix? (b) Is the $3_1$ helix steeper or shallower than the $\alpha$ helix?
4. Why would you be unlikely to see an $\alpha$ helix containing only the following amino acids: Arg, Lys, Met, Phe, Trp, Tyr, Val?
5. Calculate the length in angstroms of a 100-residue segment of the α keratin coiled coil.

6. 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?

7. The digestive tract of the larvae of clothes moths is a strongly reducing environment. Why is this beneficial to the larvae?

8. Collagen IV, which occurs in basement membranes, contains a sulfilimine bond (colored red in the structure below) that cross-links two collagen triple helices. Identify the parent amino acid residues that participate in this linkage.

\[
\begin{align*}
    &| \quad \text{C} = \text{O} \\
    &| \quad \text{OH} \\
    &| \quad \text{CH} - \text{CH}_2 - \text{CH}_2 - \text{CH} - \text{CH}_2 - \text{N} \equiv \text{S} - \text{CH}_2 - \text{CH}_2 - \text{CH}_3 \\
    &| \quad \text{NH} \\
    &| \quad \text{CH}_3 \\
    &| \quad \text{NH}
\end{align*}
\]

9. Describe the primary, secondary, tertiary, and quaternary structures of collagen.

10. Explain why gelatin, which is mostly collagen, is nutritionally inferior to other types of protein.

11. 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?

12. Is it possible for a native protein to be entirely irregular, that is, without α helices, β sheets, or other repetitive secondary structure?

13. Which of the following polypeptides is most likely to form an α helix?
   (a) CRAGNRKIVLETY
   (b) SEDNFGAPSIILW
   (c) QKASVEMAVRNSG

14. Which of the peptides in Problem 13 is least likely to form a β strand?

15. 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.

16. (a) Is Trp or Gln more likely to be on a protein’s surface? (b) Is Ser or Val less likely to be in a protein’s interior? (c) Is Leu or Ile less likely to be found in a middle of an α helix? (d) Is Cys or Ser more likely to be in a β sheet?

17. What types of rotational symmetry are possible for a protein with (a) four or (b) six identical subunits?

18. Bacterial glutamate synthetase consists of 12 identical subunits arranged in two stacked rings of six subunits. How would you describe this protein’s symmetry?

19. 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.

20. 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.

21. Given enough time, will all denatured proteins spontaneously renature?

22. Describe the intra- and intermolecular bonds or interactions that are broken or retained when collagen is heated to produce gelatin.

23. Under physiological conditions, polylysine assumes a random coil conformation. Under what conditions might it form an α helix?

24. Would intrinsically disordered polypeptide segments contain relatively more hydrophilic or hydrophobic residues? Explain.

25. 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 O₂ molecules that it binds and the ratio of the volume of a typical office (4 x 4 x 3 m) to that of the typical (70-kg) office worker that occupies it. Assume that the molecular volumes of hemoglobin and O₂ are in equal proportions to their molecular masses and that the office worker has a density of 1.0 g/cm³. Compare these ratios. Is this the result you expected?

26. In prokaryotes, the error rate in protein synthesis may be as high as 5 x 10⁻⁴ per codon. What fraction of polypeptides containing (a) 500 residues or (b) 2000 residues would you expect to contain at least one amino acid substitution?

27. Not all heat shock proteins are chaperones; some are proteins that facilitate the degradation rather than the refolding of other proteins. Explain why the rate of protein degradation would increase during heat shock.

28. Protein denaturation can be triggered by a variety of environmental insults, including high temperature, covalent modification, and oxidation. Explain why researchers have observed a correlation between the level of heat shock proteins and the ratio of oxidized to reduced glutathione (see Section 4-3B) in cells subjected to oxidative stress.

29. Researchers introduced prions into normal mice and mice that were genetically predisposed to develop a disease resembling Alzheimer’s. Explain why the Alzheimer’s-prone mice displayed symptoms of the prion disease much sooner than did the normal mice.

30. The genetically engineered proteins that accumulate in bacterial inclusion bodies (Fig. 5-2) form amyloid structures. Such proteins are often difficult to recover in functional form from the bacteria. Explain.

### BIOINFORMATICS

**Project 3: Visualizing Three-Dimensional Protein Structures**

1. **Obtaining Structural Information.** Compare different secondary structure predictions for a given protein sequence, then inspect its X-ray crystallographic structure.

2. **Exploring the Protein Data Bank.** Learn how to locate and download specific protein structure files, sequences, and images. Explore additional educational resources such as Molecule of the Month and links to additional structural biology resources.

3. **Using Jmol and PyMol.** Examine a protein structure file and use molecular modeling programs to visualize the protein and highlight selected features.

4. **Protein Families.** Identify homologous proteins in other structural databases.
CASE STUDIES

Case 4 The Structure of Insulin
Focus concept: The primary structure of insulin is examined, and the sequences of various animal insulins are compared.
Prerequisites: Chapters 4, 5, and 6
• Amino acid structure
• Protein architecture
• Basic immunology

Case 5 Characterization of Subtilisin from the Antarctic Psychrophile Bacillus TA41
Focus concept: The structural features involved in protein adaptation to cold temperatures are explored.
Prerequisite: Chapter 6
• Protein architecture
• Principles of protein folding

References

General
Fibrillar Proteins
Macromolecular Structure Determination
Protein Stability

Case 6 A Collection of Collagen Cases
Focus concept: Factors important in the stability of collagen are examined.
Prerequisites: Chapters 4, 5, and 6
• Amino acid structures and properties
• Primary and secondary structure
• Basic collagen structure

MORE TO EXPLORE
A number of human diseases are caused by point mutations that alter a single amino acid in a polypeptide. (a) Explain how such a small change in a collagen subunit can destabilize the entire collagen structure. (b) Investigate how an amino acid substitution in an enzyme can render the enzyme inactive. (c) Investigate how an amino acid substitution can affect a protein’s folding efficiency without affecting its ultimate structure or function.

Protein Folding

Protein Misfolding Diseases