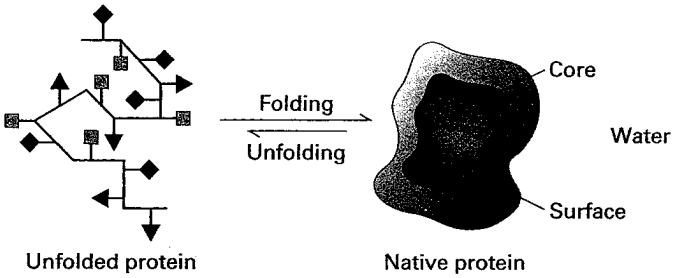
MEDIA CONNECTIONS



▲ **FIGURE 3-7 Oil drop model of protein folding.** The hydrophobic residues of a polypeptide chain tend to cluster together, somewhat like an oil drop, on the inside, or core, of a folded protein, driven away from the aqueous surroundings by the hydrophobic effect (Chapter 2). Charged and uncharged polar side chains appear on the protein's surface where they can form stabilizing interactions with surrounding water and ions.

structures, which are stabilized only by hydrogen bonds, tertiary structure is primarily stabilized by hydrophobic interactions between nonpolar side chains, together with hydrogen bonds between polar side chains and peptide bonds. These stabilizing forces compactly hold together elements of secondary structure—α helices, β strands, turns, and coils. Because the stabilizing interactions are weak, however, the tertiary structure of a protein is not rigidly fixed but undergoes continual, minute fluctuations, and some segments within the tertiary structure of a protein can be so very mobile they are considered to be disordered (that is, lacking well-defined, stable, three-dimensional structure). This variation in structure has important consequences for the function and regulation of proteins.

Chemical properties of amino acid side chains help define tertiary structure. Disulfide bonds between the side chains of cysteine residues in some proteins covalently link regions of proteins, thus restricting the mobility of proteins and increasing the stability of their tertiary structures. Amino acids with charged hydrophilic polar side chains tend to be on the outer surfaces of proteins; by interacting with water, they help to make proteins soluble in aqueous solutions and can form noncovalent interactions with other water-soluble molecules, including other proteins. In contrast, amino acids with hydrophobic nonpolar side chains are usually sequestered away from the water-facing surfaces of a protein, in many cases forming a water-insoluble central core (called the *oil drop model of globular proteins,* because of the relatively hydrophobic, or 'oily', core, Figure 3-7). Uncharged hydrophilic polar side chains are found on both the surface and inner core of proteins.

Proteins usually fall into one of three broad categories, based on their tertiary structure: fibrous proteins, globular proteins, and integral membrane proteins. *Fibrous proteins* are large, elongated, stiff molecules often composed of many tandem copies of a short sequence that forms a single repeating secondary structure (see the structure of collagen, the most abundant protein in mammals, in Chapter 19). Fibrous proteins, which often aggregate into large multiprotein fibers that do not readily dissolve in water, usually play a structural role or participate in cellular movements. *Globular proteins* are generally water-soluble, compactly folded structures, often but not exclusively spheroidal, that comprise a mixture of secondary structures (see the structure of myoglobin, below). *Integral membrane proteins* are embedded within the phospholipid bilayer of the membranes that serve as the walls of cells and organelles. The three broad categories of proteins noted here are not mutually exclusive—some proteins are made up of combinations of two or even all three of these categories.

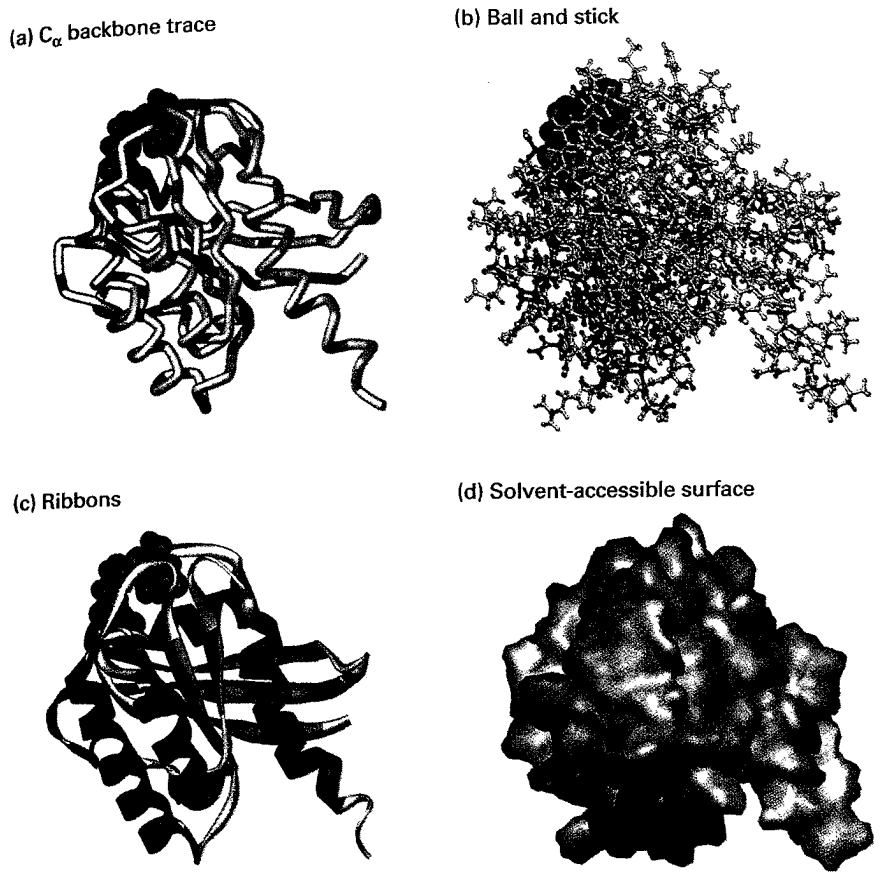
Different Ways of Depicting the Conformation of Proteins Convey Different Types of Information

The simplest way to represent three-dimensional protein structure is to trace the course of the backbone atoms, sometimes only the C_α atoms, with a solid line (called a C_α trace, Figure 3-8a); the most complex model shows every atom (Figure 3-8b). The former shows the overall fold of the polypeptide chain without consideration of the amino acid side chains; the latter, a ball-and-stick model, details the interactions between side-chain atoms, including those that stabilize the protein's conformation and interact with other molecules, as well as the atoms of the backbone. Even though both views are useful, the elements of secondary structure are not always easily discerned in them. Another type of representation uses common shorthand symbols for depicting secondary structure—for example, coiled ribbons or solid cylinders for α helices, flat ribbons or arrows for β strands, and flexible thin strands for β turns, coils, and loops (Figure 3-8c). In variations of ribbon diagrams, ball-and-stick or space-filling models of side chains can be attached to the backbone ribbon, while ribbon and cylinder models make the secondary structures of a protein easy to see.

However, none of these three ways of representing protein structure conveys much information about the protein surface, which is of interest because it is where other molecules usually bind to a protein. Computer analysis can identify the surface atoms that are in contact with the watery environment. On this water-accessible surface, regions having a common chemical character (hydrophobicity or hydrophilicity) and electrical character (basic or acidic) can be indicated by coloring (Figure 3-8d). Such models reveal the topography of the protein surface and the distribution of charge, both important features of binding sites, as well as clefts in the surface where small molecules bind. This view represents a protein as it is "seen" by another molecule.

Structural Motifs Are Regular Combinations of Secondary and Tertiary Structures

Particular combinations of secondary and tertiary structures, called **structural motifs** or **folds,** appear often as segments within many different proteins. Structural motifs

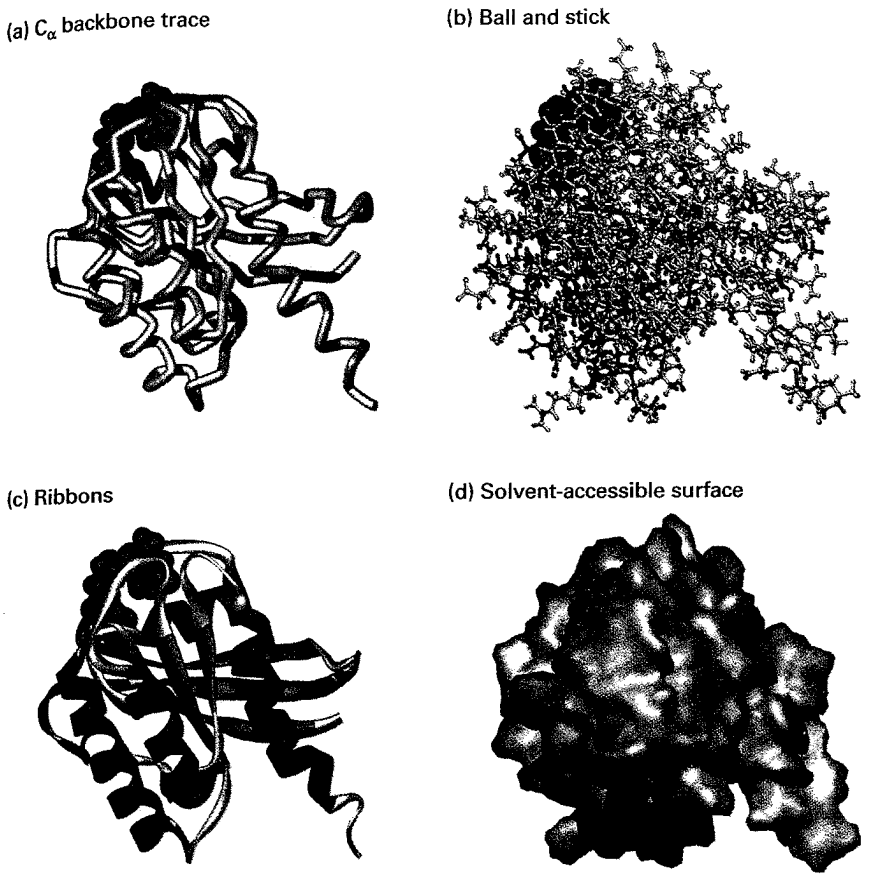


◀ **FIGURE 3-8 Four ways to visualize protein structure.** Ras, a monomeric guanine nucleotide–binding protein, is shown in all four panels, with guanosine diphosphate (GDP) always depicted in blue. (a) The C_α backbone trace demonstrates how the polypeptide is tightly packed into a small volume. (b) A ball-and-stick representation reveals the location of all atoms. (c) A ribbon representation emphasizes how β strands (light blue) and α helices (red) are organized in the protein. Note the turns and loops connecting pairs of helices and strands. (d) A model of the water-accessible surface reveals the numerous lumps, bumps, and crevices on the protein surface. Regions of positive charge are shaded purple; regions of negative charge are shaded red.

contribute to the global structure of the entire protein, and any particular structural motif often performs a common function in different proteins (e.g., binding to a particular small molecule or ion). The primary sequences responsible for any given structural motif may be very similar to one another. In other words, a common sequence motif can result in a common three-dimensional structural motif. However, it is possible for seemingly unrelated primary sequences to result in folding into a common structural motif. Conversely, it is possible that a commonly occurring sequence motif does not fold into a well-defined structural motif. Sometimes short sequence motifs that have an unusual abundance of a particular amino acid, e.g., proline or aspartate or glutamate, are called "domains"; however, these and other short contiguous segments are more appropriately called motifs than domains (which are defined below).

Many proteins, including fibrous proteins and DNA-regulating proteins called transcription factors (Chapter 7), assemble into dimers or trimers by using an α helix–based **coiled coil**, or heptad-repeat, structural motif. In this structural motif, α helices from two, three, or even four separate polypeptide chains coil about one another—resulting in a coil of coils, hence the name (Figure 3-9a). The individual helices bind tightly to one another because each helix has a strip of aliphatic (hydrophobic, but not aromatic) side chains (leucine valine, etc.) running along one side of the helix that interacts with a similar strip in the adjacent helix, thus sequestering the hydrophobic groups away from water and stabilizing the assembly of multiple independent helices. These hydrophobic strips are generated along only one side of the helix because the primary sequences of the helices exhibit a motif of repeating segments of seven amino acids (heptads) in which the sides chains of the first and fourth residues are aliphatic and the other side chains are often hydrophilic (Figure 3-9a). Because hydrophilic side chains extend from one side of the helix and hydrophobic side chains extend from the opposite side, the overall helical structure is **amphipathic.** Because leucine frequently appears in the fourth positions and the hydrophobic side chains merge together like the teeth of a zipper, these structural motifs are also called **leucine zippers.**

Many other structural motifs employ α helices. A common calcium-binding motif called the **EF hand** uses two short helices connected by a loop (Figure 3-9b). This structural motif found in more than 100 proteins is used for sensing the calcium levels in cells. The binding of a Ca^{2+} ion to oxygen atoms in conserved residues in the loop depends on the concentration of Ca^{2+} and often induces a conformational change in the protein, altering its activity. Thus, calcium concentrations can directly control proteins' structures and functions. Somewhat different helix-turn-helix and **basic helix-loop-helix** (bHLH) structural motifs are used for protein binding to DNA and consequently the regulation of gene activity. Yet another motif commonly found in proteins that bind RNA or DNA is the **zinc finger,** which contains three secondary structures—an α helix and two β strands

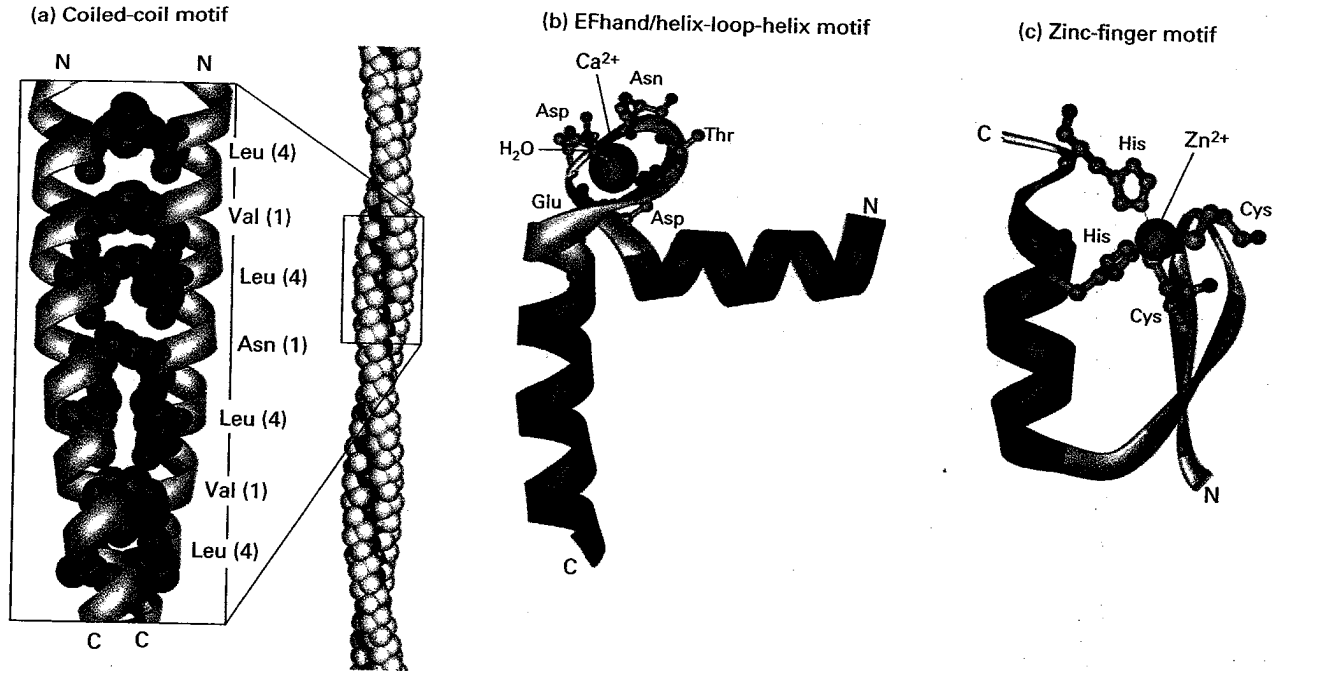


◀ **FIGURE 3-8 Four ways to visualize protein structure.** Ras, a monomeric guanine nucleotide–binding protein, is shown in all four panels, with guanosine diphosphate (GDP) always depicted in blue. (a) The C_α backbone trace demonstrates how the polypeptide is tightly packed into a small volume. (b) A ball-and-stick representation reveals the location of all atoms. (c) A ribbon representation emphasizes how β strands (light blue) and α helices (red) are organized in the protein. Note the turns and loops connecting pairs of helices and strands. (d) A model of the water-accessible surface reveals the numerous lumps, bumps, and crevices on the protein surface. Regions of positive charge are shaded purple; regions of negative charge are shaded red.

contribute to the global structure of the entire protein, and any particular structural motif often performs a common function in different proteins (e.g., binding to a particular small molecule or ion). The primary sequences responsible for any given structural motif may be very similar to one another. In other words, a common sequence motif can result in a common three-dimensional structural motif. However, it is possible for seemingly unrelated primary sequences to result in folding into a common structural motif. Conversely, it is possible that a commonly occurring sequence motif does not fold into a well-defined structural motif. Sometimes short sequence motifs that have an unusual abundance of a particular amino acid, e.g., proline or aspartate or glutamate, are called "domains"; however, these and other short contiguous segments are more appropriately called motifs than domains (which are defined below).

Many proteins, including fibrous proteins and DNA-regulating proteins called transcription factors (Chapter 7), assemble into dimers or trimers by using an α helix–based **coiled coil,** or heptad-repeat, structural motif. In this structural motif, α helices from two, three, or even four separate polypeptide chains coil about one another—resulting in a coil of coils, hence the name (Figure 3-9a). The individual helices bind tightly to one another because each helix has a strip of aliphatic (hydrophobic, but not aromatic) side chains (leucine valine, etc.) running along one side of the helix that interacts with a similar strip in the adjacent helix, thus sequestering the hydrophobic groups away from water and stabilizing the assembly of multiple independent helices. These hydrophobic strips are generated along only one side of the helix because the primary sequences of the helices exhibit a motif of repeating segments of seven amino acids (heptads) in which the sides chains of the first and fourth residues are aliphatic and the other side chains are often hydrophilic (Figure 3-9a). Because hydrophilic side chains extend from one side of the helix and hydrophobic side chains extend from the opposite side, the overall helical structure is **amphipathic.** Because leucine frequently appears in the fourth positions and the hydrophobic side chains merge together like the teeth of a zipper, these structural motifs are also called **leucine zippers.**

Many other structural motifs employ α helices. A common calcium-binding motif called the **EF hand** uses two short helices connected by a loop (Figure 3-9b). This structural motif found in more than 100 proteins is used for sensing the calcium levels in cells. The binding of a Ca^{2+} ion to oxygen atoms in conserved residues in the loop depends on the concentration of Ca^{2+} and often induces a conformational change in the protein, altering its activity. Thus, calcium concentrations can directly control proteins' structures and functions. Somewhat different helix-turn-helix and **basic helix-loop-helix** (bHLH) structural motifs are used for protein binding to DNA and consequently the regulation of gene activity. Yet another motif commonly found in proteins that bind RNA or DNA is the **zinc finger,** which contains three secondary structures—an α helix and two β strands



▲ **FIGURE 3-9 Motifs of protein secondary structure.** (a) The parallel two-stranded coiled-coil motif (*Left*) is characterized by two α helices wound around each other. Helix packing is stabilized by interactions between hydrophobic side chains (red and blue) present at regular intervals along each strand, and found along the seam of the intertwined helices. Each α helix exhibits a characteristic heptad repeat sequence with a hydrophobic residue often, but not always, at positions 1 and 4, as indicated. The coiled-coil nature of this structural motif is more apparent in long coiled coils (*Right* drawn at different scale). (b) An EF hand a type of helix-loop-helix motif, consists of two helices connected by a short loop in a specific conformation common to many proteins, including many calcium-binding and DNA-binding regulatory proteins. In calcium-binding proteins such as calmodulin, oxygen atoms from five residues in the acidic glutamate- and aspartate-rich loop and one water molecule form ionic bonds with a Ca^{2+} ion. (c) The zinc-finger motif is present in many DNA-binding proteins that help regulate transcription. A Zn^{2+} ion is held between a pair of β strands (blue) and a single α helix (red) by a pair of cysteine residues and a pair of histidine residues. The two invariant cysteine residues are usually at positions 3 and 6, and the two invariant histidine residues are at positions 20 and 24 in this 25-residue motif. [See A. Lewit-Bentley and S. Rety, 2000, EF-hand calcium-binding proteins, *Curr. Opin. Struc. Biol.* **10**:637–643; S. A. Wolfe, L. Nekludova, and C. O. Pabo, 2000, DNA recognition by Cys2His2 zinc finger proteins, *Ann. Rev. Biophys. Biomol. Struc.* **29**:183–212.]

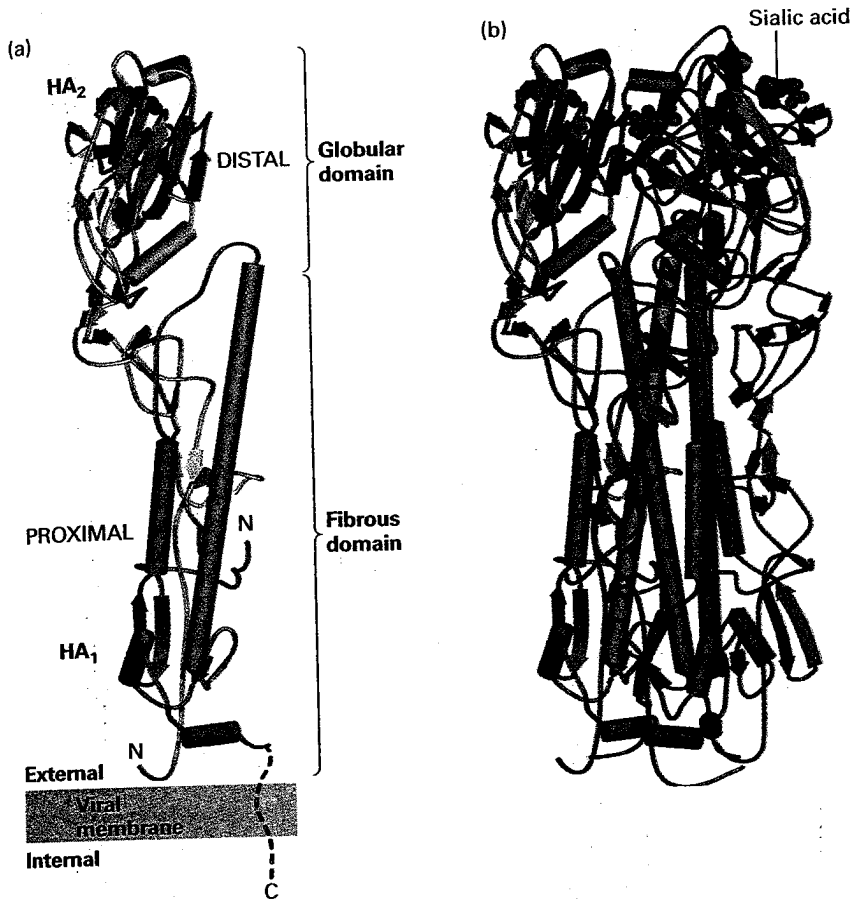
with an antiparallel orientation—that form a fingerlike bundle held together by a zinc ion (Figure 3-9c).

We will encounter numerous additional motifs in later discussions of other proteins in this and other chapters. The presence of the same structural motif in different proteins with similar functions clearly indicates that these useful combinations of secondary structures have been conserved in evolution.

Structural and Functional Domains Are Modules of Tertiary Structure

Distinct regions of protein tertiary structure are often referred to as **domains.** There are three main classes of protein domains: functional, structural, and topological. A *functional domain* is a region of a protein that exhibits a particular activity characteristic of the protein, usually even when isolated from the rest of the protein. For instance, a particular region of a protein may be responsible for its catalytic activity (e.g., a kinase domain that covalently adds a phosphate group to another molecule) or binding ability (e.g., a DNA-binding domain or a membrane-binding domain). Functional domains are often identified experimentally by whittling down a protein to its smallest active fragment with the aid of **proteases,** enzymes that cleave one or more peptide bonds in a target polypeptide. Alternatively, the DNA encoding a protein can be modified so that when the modified DNA is used to generate a protein, only a particular region, or domain, of the full-length protein is made. Thus it is possible to determine if specific portions of a protein are responsible for particular activities exhibited by the protein. Indeed, functional domains are often also associated with corresponding structural domains.

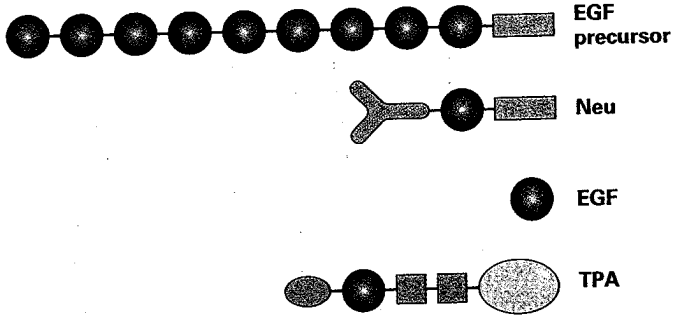
A *structural domain* is a region ≈40 or more amino acids in length, arranged in a stable, distinct secondary or tertiary structure, that often can fold into its characteristic structure independently of the rest of the protein. As a consequence, distinct structural domains can be linked together—sometimes by short or long spacers—to form a large, multidomain protein. Each of the subunits in hemagglutinin, for example, contains



◀ **FIGURE 3-10 Tertiary and quaternary levels of structure.** The protein pictured here, hemagglutinin (HA), is found on the surface of influenza virus. This long, multimeric molecule has three identical subunits, each composed of two polypeptide chains, HA_1 and HA_2. (a) Tertiary structure of each HA subunit comprises the folding of its helices and strands into a compact structure that is 13.5 nm long and divided into two domains. The membrane-distal domain (silver) is folded into a globular conformation. The membrane-proximal domain (gold) has a fibrous, stemlike conformation owing to the alignment of two long α helices (cylinders) of HA_2 with β strands in HA_1. Short turns and longer loops, often at the surface of the molecule, connect the helices and strands in each chain. (b) Quaternary structure of HA is stabilized by lateral interactions between the long helices (cylinders) in the fibrous domains of the three subunits (gold, blue, and green), forming a triple-stranded coiled-coil stalk. Each of the distal globular domains in HA binds sialic acid (red) on the surface of target cells. Like many membrane proteins, HA contains several covalently linked carbohydrate chains (not shown).

a globular domain and a fibrous domain (Figure 3-10a). Like structural motifs (composed of secondary structures), structural domains (composed of secondary and tertiary structures) are incorporated as modules into different proteins. The modular approach to protein architecture is particularly easy to recognize in large proteins, which tend to be mosaics of different domains that confer distinct activities and thus can perform different functions simultaneously. Structural domains frequently are also functional domains in that they can have an activity independent of the rest of the protein. In Chapter 6 we consider the mechanism by which the gene segments that correspond to domains became shuffled in the course of evolution, resulting in their appearance in many proteins.

The epidermal growth factor (EGF) domain is a structural domain present in several proteins (Figure 3-11). EGF is a small, soluble peptide hormone that binds to cells in the embryo and in skin and connective tissue in adults, causing them to divide. It is generated by proteolytic (breaking of peptide bond) cleavage between repeated EGF domains in the EGF precursor protein, which is anchored in the cell membrane by a membrane-spanning domain. EGF domains with sequences similar to, but not identical with, those in the EGF peptide hormone are present in other proteins and can be liberated by proteolysis. These proteins include tissue plasminogen activator (TPA), a protease that is used to dissolve blood clots in heart attack victims; Neu protein, which takes part in embryonic differentiation; and Notch protein, a receptor protein in the plasma membrane that functions in developmentally important signaling (Chapter 16). Besides the EGF domain, these proteins have other domains in common with other proteins. For example, TPA possesses a trypsin domain, a functional domain in some proteases. It is estimated that there are about 1000 different types of structural domains in all proteins. Some of these are not very common, whereas others are found in many different proteins. Indeed, by some estimates only nine major types of domains account for as much as a third of all



▲ **FIGURE 3-11 Modular nature of protein domains.** Epidermal growth factor (EGF) is generated by proteolytic cleavage of a precursor protein containing multiple EGF domains (green) and a membrane-spanning domain (blue). The EGF domain is also present in the Neu protein and in tissue plasminogen activator (TPA). These proteins also contain other widely distributed domains, indicated by shape and color. [Adapted from I. D. Campbell and P. Bork, 1993, *Curr. Opin. Struc. Biol.* **3**:385.]

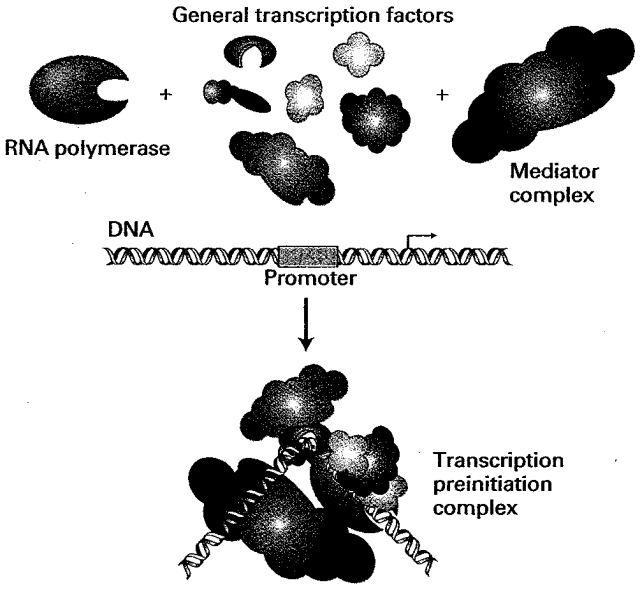
the domains in all proteins. Structural domains can be recognized in proteins whose structures have been determined by x-ray crystallography or nuclear magnetic resonance (NMR) analysis or in images captured by electron microscopy.

Regions of proteins that are defined by their distinctive spatial relationships to the rest of the protein are *topological domains*. For example, some proteins associated with cell-surface membranes can have a portion extending inward into the cytoplasm (cytoplasmic domain), a portion embedded within the phospholipid bilayer membrane (membrane-spanning domain), and a portion extending outward into the extracellular space (extracellular domain). Each of these can comprise one or more structural motifs and structural and functional domains.

Proteins Associate into Multimeric Structures and Macromolecular Assemblies

Multimeric proteins consist of two or more polypeptide chains or subunits. A fourth level of structural organization, **quaternary structure**, describes the number (stoichiometry) and relative positions of the subunits in multimeric proteins. Hemagglutinin, for example, is a trimer of three identical subunits (homotrimer) held together by noncovalent bonds (Figure 3-10b). Other multimeric proteins can be composed of various numbers of identical (homomeric) or different (heteromeric) subunits (see the discussion of hemoglobin, below). Often, the individual monomeric subunits of a multimeric protein cannot function normally unless they are assembled into the multimeric protein. In some cases, assembly into a multimeric protein (oligomerization) permits proteins that act sequentially in a pathway to increase their efficiency of operation owing to their juxtaposition in space.

The highest level in the hierarchy of protein structure is the association of proteins into macromolecular assemblies. Typically, such structures are very large, in some cases exceeding 1 MDa in mass, approaching 30–300 nm in size, and containing tens to hundreds of polypeptide chains, and sometimes other biopolymers such as nucleic acids. The capsid that encases the nucleic acids of the viral genome is an example of a macromolecular assembly with a structural function. The bundles of cytoskeletal filaments that support and give shape to the plasma membrane are another example. Other macromolecular assemblies act as molecular machines, carrying out the most complex cellular processes by integrating individual functions into one coordinated process. For example, a transcriptional machine is responsible for synthesizing messenger RNA (mRNA) using a DNA template. This transcriptional machine, the operational details of which are discussed in Chapter 4, consists of RNA polymerase, itself a multimeric protein, and at least 50 additional components including general transcription factors, promoter-binding proteins, helicase, and other protein complexes (Figure 3-12). Ribosomes, also discussed in Chapter 4, are complex multiprotein and multi-nucleic acid machines that synthesize proteins.



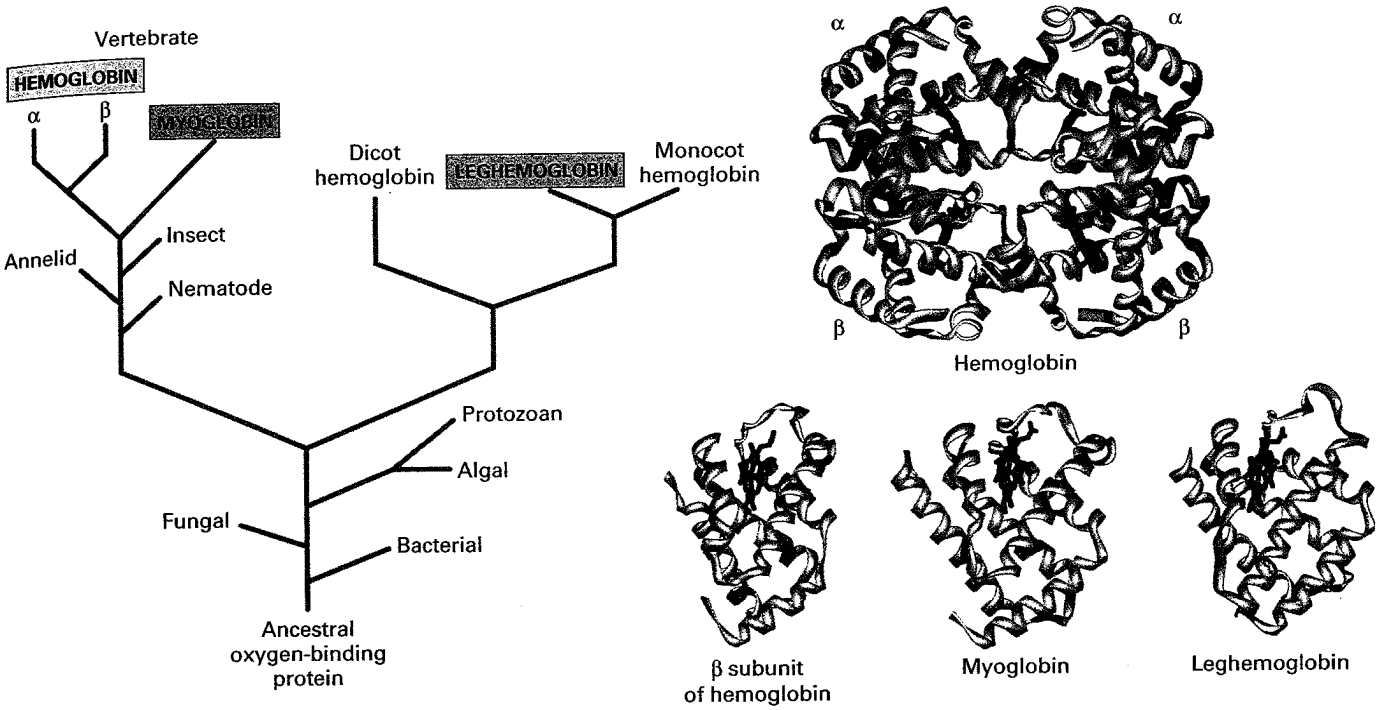
▲ **FIGURE 3-12 A macromolecular machine: the transcription-initiation complex.** The core RNA polymerase, general transcription factors, a mediator complex containing about 20 subunits, and other protein complexes not depicted here assemble at a promoter in DNA. The polymerase carries out transcription of DNA; the associated proteins are required for initial binding of polymerase to a specific promoter. The multiple components function together as a machine.

Members of Protein Families Have a Common Evolutionary Ancestor

Studies of myoglobin and hemoglobin, the oxygen-carrying proteins in muscle and red blood cells, respectively, provided early evidence that a protein's function derives from its three-dimensional structure, which in turn is specified by amino acid sequence. X-ray crystallographic analysis showed that the three-dimensional structures of myoglobin (a monomer) and the α and β subunits of hemoglobin (a $\alpha_2\beta_2$ tetramer) are remarkably similar. Sequencing of myoglobin and the hemoglobin subunits revealed that many identical or chemically similar residues are found in identical positions throughout the primary structures of both proteins. A mutation in the gene encoding the β chain that results in the substitution of a valine for a glutamic acid disturbs the folding and function of hemoglobin and causes sickle-cell anemia.

Similar comparisons between other proteins conclusively confirmed the relation between the amino acid sequence, three-dimensional structure, and function of proteins. Use of sequence comparisons to deduce protein function has expanded substantially in recent years as the genomes of more and more organisms have been sequenced.

The molecular revolution in biology during the last decades of the twentieth century also created a new scheme of biological classification based on similarities and differences in the amino acid sequences of proteins. Proteins that have a common ancestor are referred to as **homologs**. The main evidence for **homology** among proteins, and hence for their common ancestry, is similarity in their sequences or



▲ **FIGURE 3-13 Evolution of the globin protein family.** *Left:* A primitive monomeric oxygen-binding globin is thought to be the ancestor of modern-day blood hemoglobins, muscle myoglobins, and plant leghemoglobins. Sequence comparisons have revealed that evolution of the globin proteins parallels the evolution of animals and plants. Major junctions occurred with the divergence of plant globins from animal globins and of myoglobin from hemoglobin. Later gene duplication gave rise to the α and β subunits of hemoglobin. *Right:* Hemoglobin is a tetramer of two α and two β subunits. The structural similarity of these subunits with leghemoglobin and myoglobin, both of which are monomers, is evident. A heme molecule (red) noncovalently associated with each globin polypeptide is directly responsible for oxygen-binding in these proteins. [(*Left*) Adapted from R. C. Hardison, 1996, *Proc. Nat'l Acad. Sci. USA* **93**:5675.]

structures. We can therefore describe homologous proteins as belonging to a "family" and can trace their lineage from comparisons of their sequences. The folded three-dimensional structures of homologous proteins are similar even if parts of their primary structure show little evidence of homology. Initially, proteins with relatively high sequence similarities (>50 percent exact matches, or "identities") and related functions or structures were defined as an evolutionarily related *family,* while a *superfamily* encompassed two or more families in which the interfamily sequences matched less well (≈30–40 percent identities) than within one family. It is generally thought that proteins with 30 percent sequence identity are likely to have similar three-dimensional structures; however, proteins with far less sequence matching can have very similar structures. Recently, revised definitions of *family* and *superfamily* have been proposed, in which a family comprises proteins with a clear evolutionary relationship (>30 percent identity or additional structural and functional information showing common descent but <30 percent identity), while a superfamily comprises proteins with only a probable common evolutionary origin (e.g., lower percent sequence identities). Often investigators consider proteins to constitute a common superfamily (have a common evolutionary origin) when they contain one or more common motifs or domains.

The kinship among homologous proteins is most easily visualized by a tree diagram based on sequence analyses. For example, the amino acid sequences of globins, the proteins of hemoglobin and myoglobin and their relatives from bacteria, plants, and animals, suggest that they evolved from an ancestral monomeric, oxygen-binding protein (Figure 3-13). With the passage of time, the gene for this ancestral protein slowly changed, initially diverging into lineages leading to animal and plant globins. Subsequent changes gave rise to myoglobin, the monomeric oxygen-storing protein in muscle, and to the α and β subunits of the tetrameric hemoglobin molecule ($\alpha_2\beta_2$) of the circulatory system.

KEY CONCEPTS OF SECTION 3.1

Hierarchical Structure of Proteins

- A protein is a linear polymer of amino acids linked together by peptide bonds. Various, mostly noncovalent, interactions between amino acids in the linear sequence stabilize a protein's specific folded three-dimensional structure, or conformation.
- The α helix, β strand and sheet, and β turn are the most prevalent elements of protein secondary structure. Secondary structures are stabilized by hydrogen bonds between atoms of the peptide backbone.
- Protein tertiary structure results from hydrophobic interactions between nonpolar side groups and hydrogen bonds between polar side groups and the polypeptide backbone.

These interactions stabilize folding of the secondary structure into a compact overall arrangement.

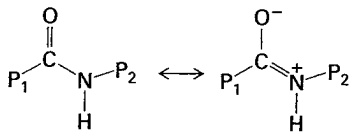
- Certain combinations of secondary structures give rise to different motifs, which are found in a variety of proteins and are often associated with specific functions (see Figure 3-9).
- Proteins often contain distinct domains, independently folded regions of secondary or tertiary structure with characteristic structural, functional, and topological properties (see Figure 3-10).
- The incorporation of domains as modules in different proteins in the course of evolution has generated diversity in protein structure and function.
- The number and organization of individual polypeptide subunits in multimeric proteins define their quaternary structure.
- Cells contain large macromolecular assemblies in which all the necessary participants in complex cellular processes (e.g., DNA, RNA, and protein synthesis; photosynthesis; signal transduction) are integrated to form molecular machines.
- Homologous proteins, which have similar sequences, structures, and functions, evolved from a common ancestor. They can be classified into families and superfamilies.

3.2 Protein Folding

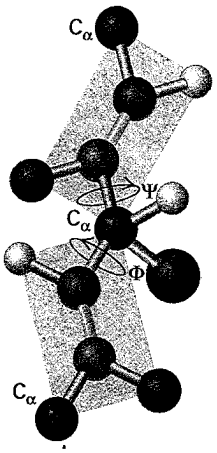
As noted above, when it comes to biological structures such as proteins, "form follows function" and "form is function." Thus it is essential that when a polypeptide is synthesized with its particular primary structure (sequence), it folds into the proper three-dimension conformation with the appropriate secondary, tertiary, and possibly quaternary structure. A polypeptide chain is synthesized by a complex process called **translation** in which the assembly of amino acids in a particular sequence is dictated by **messenger RNA (mRNA)** and performed by a large protein–nucleic acid complex called a **ribosome.** The intricacies of translation are considered in Chapter 4. Here, we describe the key determinants of the proper folding of a *nascent* (newly formed or forming) polypeptide chain.

Planar Peptide Bonds Limit the Shapes into Which Proteins Can Fold

A critical structural feature of polypeptides that limits how the chain can fold is the planar peptide bond. Figure 3-3 illustrates the amide group in peptide bonds in a polypeptide chain. Because the peptide bond itself behaves partially like a double bond,



the carbonyl carbon and amide nitrogen and those atoms directly bonded to them must all lie in a fixed plane (Figure 3-14); there is no rotation possible about the peptide bond itself. As a consequence, the only flexibility in a polypeptide chain backbone, allowing it to adopt varying conformations (twists and turns to fold into different three-dimensional shapes), is rotation of the fixed planes of peptide bonds with respect to one another about two bonds—the C_α–amino nitrogen bond (rotational angle called ϕ) and the C_α–carbonyl carbon bond (rotational angle called ψ).



▲ **FIGURE 3-14 Rotation between planar peptide groups in proteins.** Rotation about the C_α–amino nitrogen bond (the ϕ angle) and the C_α–carbonyl carbon bond (the ψ angle) permits polypeptide backbones, in principle, to adopt a very large number of potential conformations. However steric restraints due to the structure of the polypeptide backbone and the properties of the amino acid side chains dramatically restrict the potential conformations that can be adopted by any given protein.

Yet a further constraint on the potential conformations that a polypeptide backbone chain can adopt is that only a limited number of ϕ and ψ angles are possible, because for most ϕ and ψ angles the backbone or side chain atoms would come too close to one another and thus the associated conformation would be highly unstable or even physically impossible to achieve.

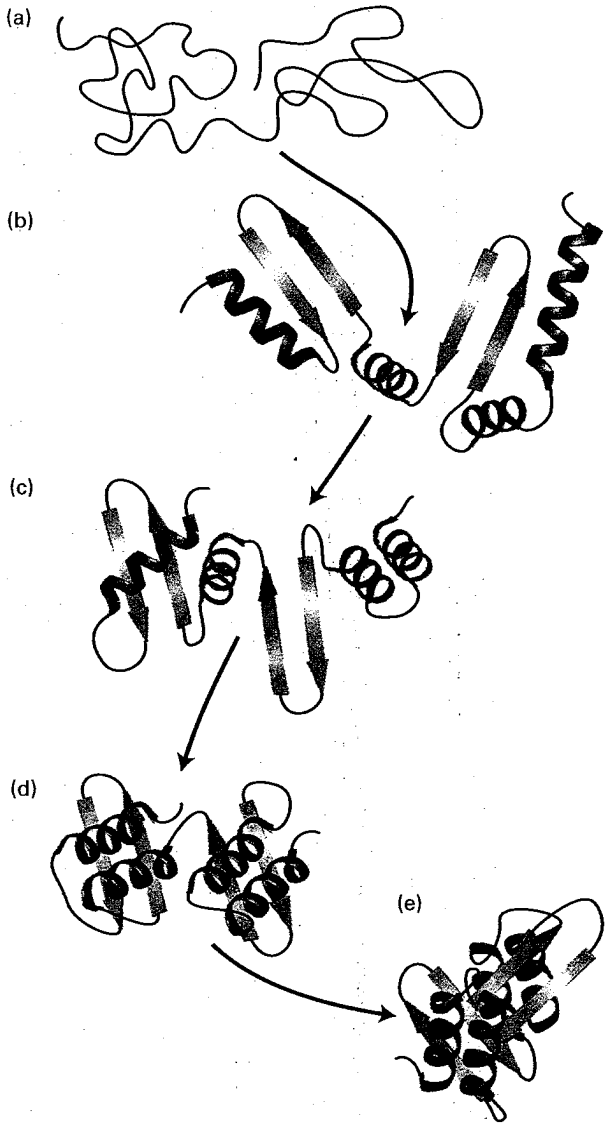
Information Directing a Protein's Folding Is Encoded in Its Amino Acid Sequence

While the constraints of backbone bond angles seem very restrictive, any polypeptide chain containing only a few residues could, in principle, still fold into many conformations. For example, if the ϕ and ψ angles were limited to only eight combinations, an n-residue-long peptide would potentially have 8^n conformations—a very large number for even a small polypeptide of only 10 residues long (about 8.6 million possible conformations)! In general, however, any particular protein adopts only one or just a few very closely related characteristic functional conformations called the *native state;* for the vast majority of proteins, the native state is the most stably folded form of the molecule. In thermodynamic terms, the native state is usually the conformation with the lowest free energy. What features of proteins limit their folding from very many conformations to just one? The properties of the side chains (e.g., size, hydrophobicity, ability

to form hydrogen and ionic bonds), together with their particular sequence along the polypeptide backbone, impose key restrictions. For example, a large side chain such as that of tryptophan might prevent (sterically block) one region of the chain from packing closely against another region, whereas a side chain with a positive charge such as arginine might attract a segment of the polypeptide that has a complementary negatively charged side chain (e.g., aspartic acid). Another example we have already discussed is the effect of the aliphatic side chains in heptad repeats on the formation of coiled coils. Thus, a polypeptide's primary structure determines its secondary, tertiary, and quaternary structure.

The initial evidence that the information necessary for a protein to fold properly is encoded in its sequence came from in vitro studies on the refolding of purified proteins. Various perturbations (such as thermal energy from heat, extremes of pH that alter the charges on amino acid side chains, and chemicals, called *denaturants,* such as urea or guanidine hydrochloride at concentrations of 6–8 M) can disrupt the weak noncovalent interactions that stabilize the native conformation of a protein, leading to its **denaturation.** Treatment with the reducing agents, such as β-mercaptoethanol, that break disulfide bonds can further destabilize disulfide-containing proteins. Under such unfolding or denaturing conditions, entropy increases when a population of uniformly folded molecules is destabilized and converted into a collection of many unfolded, or denatured, molecules that have many different non-native and biologically inactive conformations. As we have seen, there are very many possible non-native conformations (e.g., 8^n-1).

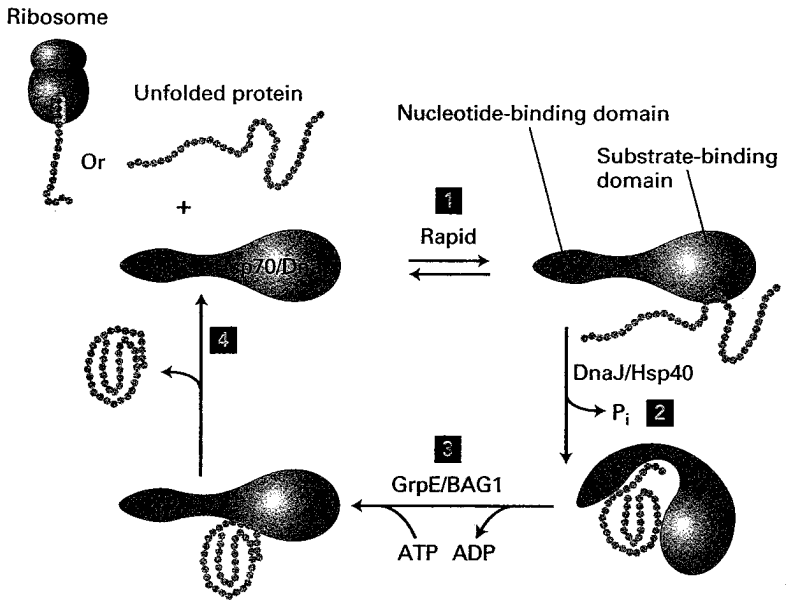
The spontaneous unfolding of proteins under denaturing conditions is not surprising, given the substantial increase in entropy. What is striking, however, is that when a pure sample of a single type of unfolded protein is shifted back to normal conditions (body temperature, normal pH levels, reduction in the concentration of denaturants by dilution or their removal), some denatured polypeptides can spontaneously renature (refold) into their native, biologically active states. This kind of refolding experiment, as well as studies that show synthetic proteins made chemically can fold properly, showed that sufficient information must be contained in the protein's primary sequence to direct correct refolding. Newly synthesized proteins appear to fold into their proper conformations just as denatured proteins do. The observed similarity in the folded, three-dimensional structures of proteins with similar amino acid sequences, noted in Section 3.1, provided additional evidence that the primary sequence also determines protein folding in vivo. It appears that formation of secondary structures and structural motifs occurs early in the folding process, followed by assembly of more compact and complex domains, which then associate into more complex tertiary and quaternary structures (Figure 3-15).



▲ **FIGURE 3-15 Hypothetical protein-folding pathway.** Folding of a monomeric protein follows the structural hierarchy of primary (a) → secondary (b–d) → tertiary (e) structure. Formation of small structural motifs (c) appears to precede formation of more stable domains (d) and the final tertiary structure (e).

Folding of Proteins in Vivo Is Promoted by Chaperones

The refolding of a denatured protein is presumed to mimic many aspects of the folding of a newly synthesized polypeptide. However the conditions inside a cell are not the same as those in test tubes used for in vitro refolding experiments with purified proteins. The presence of other biomolecules, including many other proteins at very high concentration, some of which are themselves nascent and in the process of folding, can potentially interfere with the autonomous, spontaneous folding of a protein. Furthermore, although protein folding into the native state can occur in vitro, this does not happen for all unfolded molecules in a timely fashion. Given such impediments, cells require a faster, more efficient mechanism for folding proteins into their correct shapes than sequence alone provides. Without such help, they would waste much energy in



◀ **FIGURE 3-16 Chaperone-mediated protein folding.** Many proteins fold into their proper three-dimensional structures with the assistance of Hsp70-like proteins. These molecular chaperones transiently bind to a nascent polypeptide as it emerges from a ribosome or to proteins that have otherwise unfolded. In the Hsp70 cycle, a substrate unfolded protein binds in rapid equilibrium to the open conformation of the substrate-binding domain (SBD) of Hsp70, to which an ATP is bound in the nucleotide-binding domain (NBD) (step 1). Accessory proteins (DnaJ/Hsp40) stimulate the hydrolysis of ATP and conformational change in Hsp70, resulting in the closed form, in which the substrate is locked into the SBD; here proper folding is facilitated (step 2). Exchange of ATP for the bound ADP, stimulated by other accessory proteins (GrpE/BAG1), converts the Hsp70 back to the open form (step 3), releasing the properly folded substrate (step 4).

the synthesis of improperly folded, nonfunctional proteins, which would have to be destroyed to prevent their disrupting cell function. Cells clearly have such mechanisms, since more than 95 percent of the proteins present within cells have been shown to be in their native conformations. The explanation for the cell's remarkable efficiency in promoting the proper protein folding encoded in primary sequence is that cells make a set of proteins, called **chaperones**, that facilitate protein folding. The importance of chaperones is highlighted by the observations that they are evolutionarily conserved, they are found in all organisms from bacteria to humans, and some are highly homologous and use almost identical mechanisms to assist protein folding. Chaperones, which in eukaryotes are located in every cellular compartment and organelle, bind to the target proteins whose folding they will assist. Two general families of chaperones are recognized:

- **Molecular chaperones,** which bind and stabilize unfolded or partly folded proteins, thereby preventing these proteins from aggregating and being degraded
- **Chaperonins,** which form a small folding chamber into which an unfolded protein can be sequestered, giving it time and an appropriate environment to fold properly

One reason that chaperones are needed for intracellular protein folding is that they help prevent aggregation of unfolded proteins. Unfolded and partly folded proteins tend to aggregate into large, often water insoluble masses, from which it is extremely difficult for a protein to dissociate and then fold into its proper conformation. In part this aggregation is due to the exposure of hydrophobic side chains that have not yet had a chance to be buried in the inner core of the folded protein. These exposed hydrophobic side chains on different molecules will stick to one another owing to the hydrophobic effect (Chapter 2) and thus promote aggregation. When a newly synthesized molecule begins to fold, it is at risk of aggregating before it completes its proper folding. Molecular chaperones bind to the target polypeptide or sequester it from other partially or fully unfolded proteins, thereby preventing aggregation and thus giving the nascent protein time to fold properly.

Molecular Chaperones The heat-shock protein Hsp70 and its homologs (Hsp70 in the cytosol and mitochondrial matrix, BiP in the endoplasmic reticulum, and DnaK in bacteria) are molecular chaperones. They were first identified by their rapid appearance after a cell has been stressed by heat shock (*Hsp* stands for "*h*eat-shock *p*rotein"). Hsp70 and its homologs are the major chaperones in all organisms. When bound to ATP, Hsp70-like proteins assume an open form in which an exposed hydrophobic pocket transiently binds to exposed hydrophobic regions of an incompletely folded or partially denatured target protein (Figure 3-16). Hydrolysis of the bound ATP causes the molecular chaperone to assume a closed form that appears to facilitate the target protein's folding, in part by preventing unfolded proteins from aggregating. The exchange of ATP for the protein-bound ADP causes a conformational change in the chaperone that releases the target protein.

Additional proteins, such as the co-chaperone Hsp40 in eukaryotes (DnaJ in bacteria), help increase efficiency of