

Protein Folding, Dynamics, and Structural Evolution

CHAPTER 9

1 Protein Folding: Theory and Experiment

- A. Protein Renaturation
- B. Determinants of Protein Folding
- C. Folding Pathways

2 Folding Accessory Proteins

- A. Protein Disulfide Isomerase
- B. Peptidyl Prolyl Cis-Trans Isomerase
- C. Molecular Chaperones: The GroEL/ES System

3 Protein Structure Prediction and Design

- A. Secondary Structure Prediction
- B. Tertiary Structure Prediction
- C. Protein Design

4 Protein Dynamics

5 Conformational Diseases: Amyloids and Prions

- A. Amyloid Diseases
- B. Alzheimer's Disease
- C. Prion Diseases

6 Structural Evolution

- A. Structures of Cytochromes *c*
- B. Gene Duplication

In the preceding chapters, we saw how proteins are constructed from their component parts. This puts us in a similar position to a mechanic who has learned to take apart and put together an automobile engine without any inkling of how the engine works. What we need in order to understand the workings of a protein is knowledge of the types of internal motions it can and must undergo in order to carry out its biological function as well as how it achieves its native structure. Put in terms of our deprived auto mechanic, we wish to understand the operations of the “gears” and “levers” with which proteins carry out their function. This is a problem of enormous complexity whose solution we are only beginning to understand. We shall see in later chapters, for example, that even though the catalytic mechanisms of many enzymes of known structure have been studied in great detail, it cannot be said that we fully understand any of these mechanisms. This is because our comprehension of the ways in which a protein's component groups interact is far from complete. As far as proteins are concerned, we have not greatly surpassed our hypothetical mechanic's level of understanding.

In this third of four chapters on protein structure, we consider the temporal behavior of proteins. Specifically, we first take up the problem of how random coil polypeptides fold to their native structures and how this process is facilitated by other proteins. This is followed by a consideration of the progress that has been made in predicting protein structures based on their sequences and in understanding their dynamic properties, that is, the nature and functional significance of their internal motions. Next we consider the diseases caused by proteins taking up alternate conformations. We end by extending the discussions we began in Section 7-3 on protein evolution but do so in terms of the three-dimensional structures of proteins.

1 PROTEIN FOLDING: THEORY AND EXPERIMENT

Solving the so-called **protein folding problem**, that is, determining how and why proteins fold to their native states, is considered to be one of the grand challenges of biochemistry. Early notions of protein folding postulated the existence of “templates” that somehow caused proteins to assume their native conformations. Such an explanation begs the question of how proteins fold because, even if it were true, one would still have to explain how the template achieved its conformation. In fact, *proteins spontaneously fold into their native conformations under physiological conditions*. This implies that *a protein's primary structure dictates its three-dimensional structure*. In general, under the proper conditions, biological structures are **self-assembling** so that they have no need of external templates to guide their formation.

A. Protein Renaturation

Although evidence had been accumulating since the 1930s that proteins could be reversibly denatured, it was not until 1957 that the elegant experiments of Christian Anfinsen on bovine pancreatic **RNase A** put **protein renaturation** on a quantitative basis. RNase A, a 124-residue single-chain protein, is completely unfolded and its four disulfide bonds reductively cleaved in an 8*M* urea solution containing 2-mercaptoethanol (Fig. 9-2). Yet dialyzing away the urea and exposing the resulting solution to O₂ at pH 8

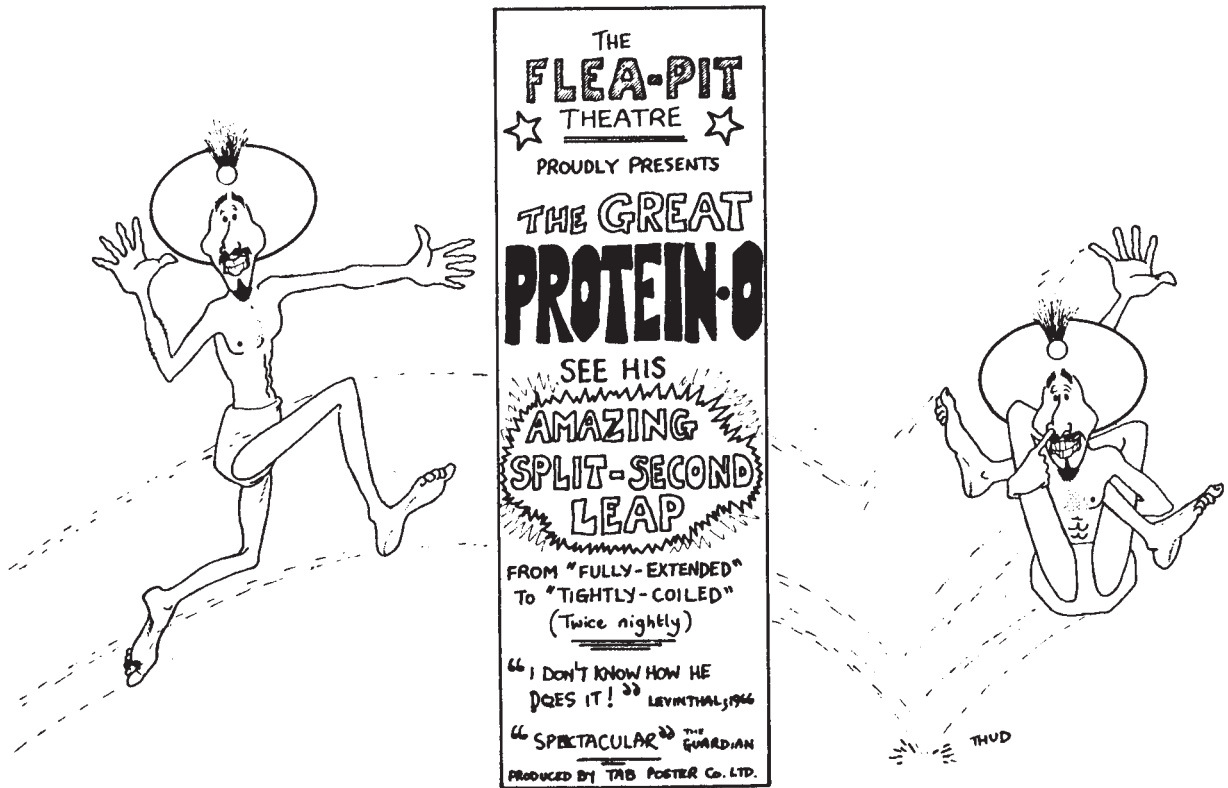


Figure 9-1 [Drawing by T.A. Bramley, in Robson, B., *Trends Biochem. Sci.* 1, 50 (1976). Copyright © Elsevier Biomedical Press, 1976. Used by permission.]

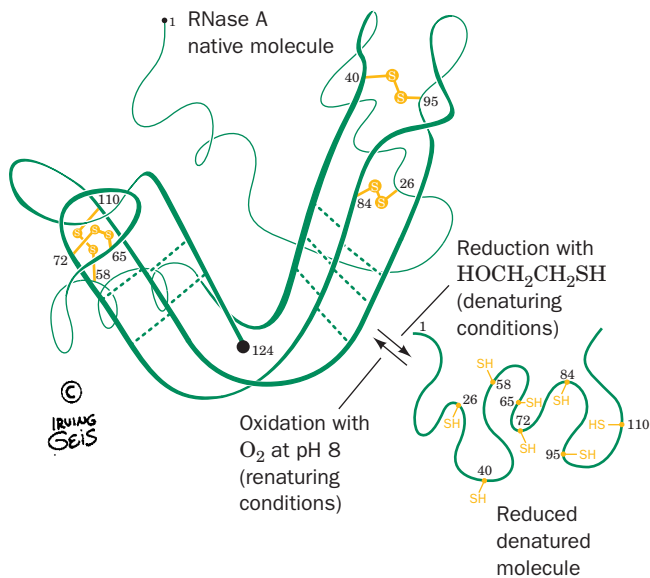


Figure 9-2 Reductive denaturation and oxidative renaturation of RNase A. [Illustration, Irving Geis. Image from the Irving Geis Collection, Howard Hughes Medical Institute. Reprinted with permission.]

yields a protein that is virtually 100% enzymatically active and physically indistinguishable from native RNase A. The protein must therefore have spontaneously renatured. Any reservations that this occurs only because RNase A is really not totally denatured by 8M urea have been satisfied by the chemical synthesis of enzymatically active RNase A (Section 7-5).

The renaturation of RNase A demands that its four disulfide bonds reform. The probability of one of the eight Cys residues from RNase A randomly reforming a disulfide bond with its proper (native) mate among the other seven Cys residues is $\frac{1}{7}$; that of one of the remaining six Cys residues then randomly reforming its proper disulfide bond is $\frac{1}{5}$; etc. The overall probability of RNase A reforming 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 from RNase A do not randomly reform under renaturing conditions.

If the RNase A is reoxidized in 8M urea so that its disulfide bonds reform while the polypeptide chain is a random coil, then after removal of the urea, the RNase A is, as expected, only ~1% enzymatically active. This “scrambled” RNase A can be made fully active by exposing it to a trace of 2-mercaptoethanol, which, over about a 10-h period,

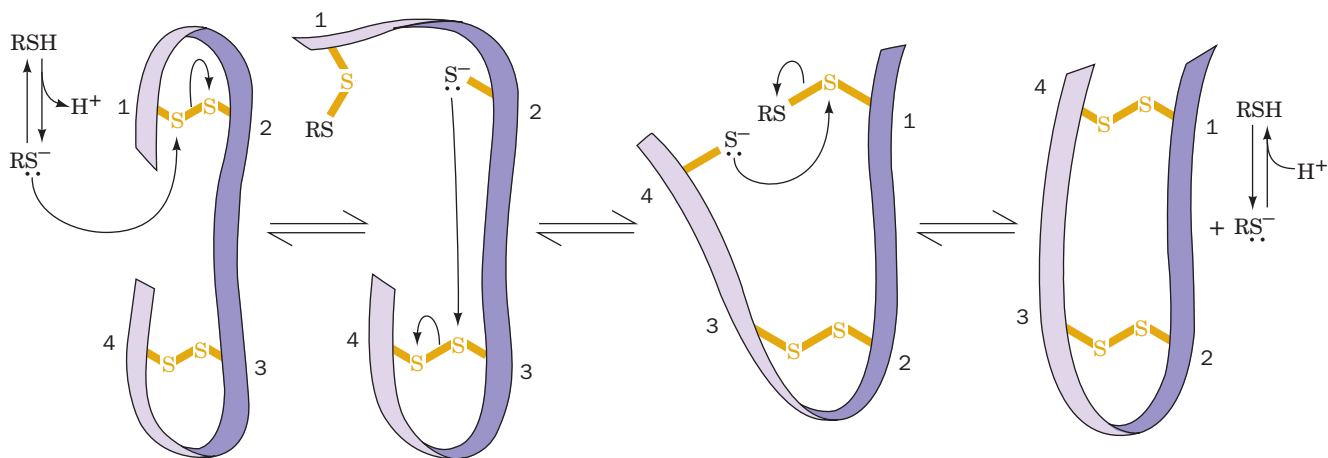


Figure 9-3 Plausible mechanism for the thiol- or enzyme-catalyzed disulfide interchange reaction in a protein. The purple ribbon represents the polypeptide backbone of the protein. The attacking thiol group must be in its ionized thiolate form.

catalyzes disulfide bond interchange reactions until the native structure is achieved (Fig. 9-3). The native state of RNase A under physiological conditions is therefore, most probably, its thermodynamically most stable conformation (if the protein has a conformation that is more stable than the native state, conversion to it must involve such a large activation barrier so as to make it kinetically inaccessible; rate processes are discussed in Section 14-1C).

The time for renaturation of “scrambled” RNase A is reduced to ~2 min through the use of an enzyme, **protein disulfide isomerase (PDI)**, that catalyzes disulfide interchange reactions. (In fact, the supposition that *in vivo* folding to the native state requires no more than a few minutes prompted the search that led to this enzyme’s discovery.) PDI itself contains two active site Cys residues, which must be in the —SH form for the isomerase to be active. The enzyme evidently catalyzes the random cleavage and reformation of a protein’s disulfide bonds (Fig. 9-3), thereby interchanging them as the protein progressively attains thermodynamically more favorable conformations. PDI is further discussed in Section 9-2A.

a. Post-Translationally Modified Proteins May Not Readily Renature

Many “scrambled” proteins are renatured through the action of PDI and are unaffected by it in their native state (their PDI-cleaved disulfide bonds rapidly reform because these native proteins are in their most stable local conformations). In post-translationally modified proteins, however, the disulfide bonds may serve to hold the protein in its otherwise unstable native state. For instance, the 51-residue polypeptide hormone **insulin**, which consists of two polypeptide chains joined by two disulfide bonds (Fig. 7-2), is inactivated by PDI. This observation led to the discovery that insulin is derived from a single-chain, 84-residue precursor named **proinsulin** (Fig. 9-4). Only after its disulfide bonds have formed is proinsulin converted to the two-chained active hormone by the specific proteolytic excision of an internal 33-residue segment known as its C chain. Nevertheless, two

sets of observations suggest that the C chain does not direct the folding of the A and B chains but, rather, simply holds them together while they form their native disulfide bonds: (1) Under proper renaturing conditions, native

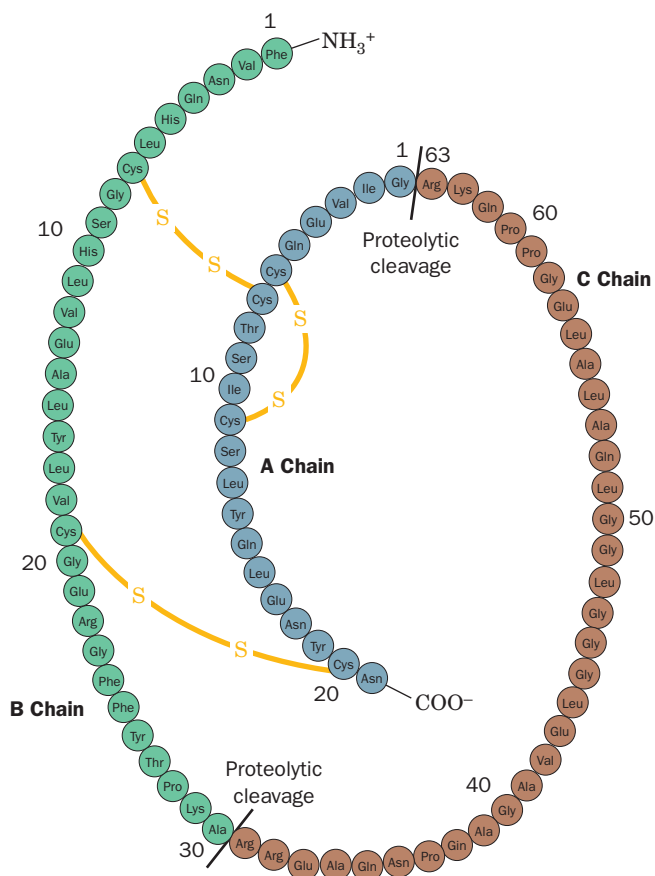


Figure 9-4 Primary structure of porcine proinsulin. Its C chain (brown) is proteolytically excised from between its A and B chains to form the mature hormone. [After Chance, R.E., Ellis, R.M., and Brommer, W.W., *Science* **161**, 165 (1968).]

insulin is obtained from scrambled insulin in 25 to 30% yield, which increases to 75% when the A and B chains are chemically cross-linked; and (2) sequence comparisons of proinsulins from a variety of species indicate that mutations are accepted into the C chain at a rate which is eight times that for the A and B chains.

B. Determinants of Protein Folding

In Section 8-4, we discussed the various interactions that stabilize native protein structures. In this section we extend the discussion by considering how these interactions are organized in native proteins. Keep in mind that only a small fraction of the myriads of possible polypeptide sequences are likely to have unique stable conformations. Evolution has, of course, selected such sequences for use in biological systems.

a. Helices and Sheets May Predominate in Proteins Simply because They Fill Space Efficiently

Why do proteins contain such a high proportion (~60%, on average) of α helices and β pleated sheets? Hydrophobic interactions, although the dominant influence responsible for the compact nonpolar cores of proteins, lack the specificity to restrict polypeptides to particular conformations. Similarly, the observation that polypeptide segments in the coil conformation are no less hydrogen bonded than helices and sheets suggests that the conformations available to polypeptides are not greatly limited by their hydrogen bonding requirements. Rather, as Ken Dill has shown, it appears that helices and sheets form largely as a consequence of steric constraints in compact polymers. Exhaustive simulations of the conformations which simple flexible chains (such as a string of pearls) can assume indicate that the proportion of helices and sheets increases dramatically with a chain's level of compaction (number of intrachain contacts); that is, helices and sheets are particularly compact entities. Thus, most ways to compact a chain involve the formation of helices and sheets. In native proteins, such elements of secondary structure are fine tuned to form α helices and β sheets by short-range forces such as hydrogen bonding, ion pairing, and van der Waals interactions. It is probably these less dominant but more specific forces that "select" the unique native structure of a protein from among its relatively small number of hydrophobically generated compact conformations (recall that most hydrogen bonds in proteins link residues that are close together in sequence; Section 8-4Bb).

b. Protein Folding Is Directed Mainly by Internal Residues

Numerous protein modification studies have been aimed at determining the role of various classes of amino acid residues in protein folding. In one particularly revealing study, the free primary amino groups of RNase A (Lys residues and the N-terminus) were derivatized with 8-residue chains of poly-DL-alanine. Intriguingly, these large, water-soluble poly-Ala chains could be simultaneously coupled to RNase's 11 free amino groups without significantly altering the protein's native conformation or its ability to

refold. Since these free amino groups are all located on the exterior of RNase A, this observation suggests that *it is largely a protein's internal residues that direct its folding to the native conformation*. Similar conclusions have been reached from studies of protein structure and evolution (Section 9-6): Mutations that change surface residues are accepted more frequently and are less likely to affect protein conformations than are changes of internal residues. It is therefore not surprising that the perturbation of protein folding by limited concentrations of denaturing agents indicates that *protein folding is driven by hydrophobic forces*.

c. Protein Structures Are Hierarchically Organized

Large protein subunits consist of domains, that is, of contiguous, compact, and physically separable segments of the polypeptide chain. Furthermore, as George Rose showed, domains consist of subdomains, which in turn consist of sub-subdomains, etc. Conceptually, this means that if a polypeptide segment of any length in a native protein is viewed as a tangle of string, a single plane can be found that divides the string into only two segments rather than many smaller segments (such as would happen if a ball of yarn were cut in this way). This is readily demonstrated by coloring the first $n/2$ residues of an n -residue domain red and the second $n/2$ residues blue. If this process is iterated, as is shown in Fig. 9-5 for high potential iron-sulfur protein (HiPIP), it is clear that at every stage of the process, the red

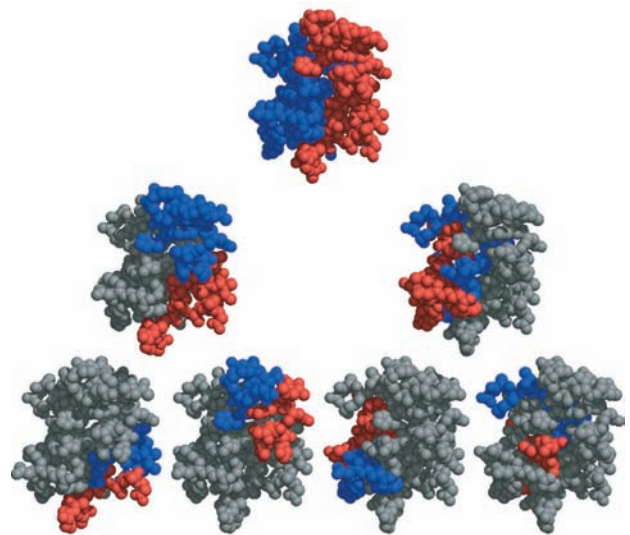


Figure 9-5 Hierarchical organization of globular proteins. Here the X-ray structure of high potential iron-sulfur protein (HiPIP) is represented by its C_{α} atoms shown as spheres. In the top drawing, the first $n/2$ residues of this n -residue protein (where $n = 71$) are colored red and the remaining $n/2$ residues are colored blue. In the second row, the process is iterated such that, on the right, for example, the first and last halves of the second half of the protein are red and blue, with the remainder of the chain gray. In the third row, the process is again iterated. Note that at each stage of this hierarchy, the red and blue regions do not intermingle. [Courtesy of George Rose, Johns Hopkins University, and Robert Baldwin, Stanford University School of Medicine.]

and blue regions do not interpenetrate. Evidently, *protein structures are organized hierarchically*, that is, polypeptide chains form locally compact structures that associate with similar adjacent (in sequence) structures to form larger compact structures, etc. This structural organization is, of course, consistent with the observation that hydrogen bonding interactions in proteins are mostly local (Section 8-4Bb). It also has important implications for how polypeptides fold to form native proteins (Section 9-1C).

d. Protein Structures Are Highly Adaptable

Globular proteins have packing densities comparable to those of organic crystals (Section 8-3Bc) because the side chains in a protein's interior fit together with exquisite complementarity. To ascertain whether this phenomenon is an important determinant of protein structure, Eaton Lattman and Rose analyzed 67 globular proteins of known structure for the existence of preferred interactions between side chains. They found none, thereby indicating that, at least in globular proteins, *the native fold determines the packing but packing does not determine the native fold*. This view is corroborated by the widespread occurrence of protein families whose members assume the same fold even though they may be so distantly related as to have no recognizable sequence similarity (e.g., the α/β barrel proteins; Section 8-3Bh).

The foregoing study indicates that *there are a large number of ways in which a protein's internal residues can pack together efficiently*. This was perhaps most clearly shown by Brian Matthews in an extensive series of studies on **T4 lysozyme** (a product of bacteriophage T4) in which the X-ray structures of over 300 mutant varieties of this 164-residue monomeric enzyme were compared. Replacements of one or a few residues in T4 lysozyme's hydrophobic core were accommodated mainly by local shifts in the protein backbone rather than by any global structural changes. In many cases, T4 lysozyme could accommodate the insertion of up to four residues without a major structural change or even a loss of enzymatic activity. Moreover, assays of the enzymatic activities of 2015 single-residue substitutions in T4 lysozyme indicated that only 173 of these mutants had significantly decreased enzymatic activity. Clearly protein structures are highly resilient.

e. Secondary Structure Can Be Context-Dependent

The structure of a native protein is determined by its amino acid sequence, but to what extent is the conformation of a given polypeptide segment influenced by the surrounding protein? The NMR structure of **protein GB1** (the B1 domain of streptococcal **protein G**, which helps the bacterium evade the host's immunological defenses by binding to the antibody protein **immunoglobulin G**) reveals that this 56-residue domain, which lacks disulfide bonds, consists of a long α helix lying across a 4-stranded mixed β sheet (Fig. 9-6). In mutagenesis experiments by Peter Kim, the 11-residue "chameleon" sequence AWTVEKAFKTF was made to replace either residues 23 to 33 of GB1's α helix (AATAEKVFVQY in GB1; a 7-residue change) to yield Chm- α , or residues 42 to 52 of its C-terminal β hairpin

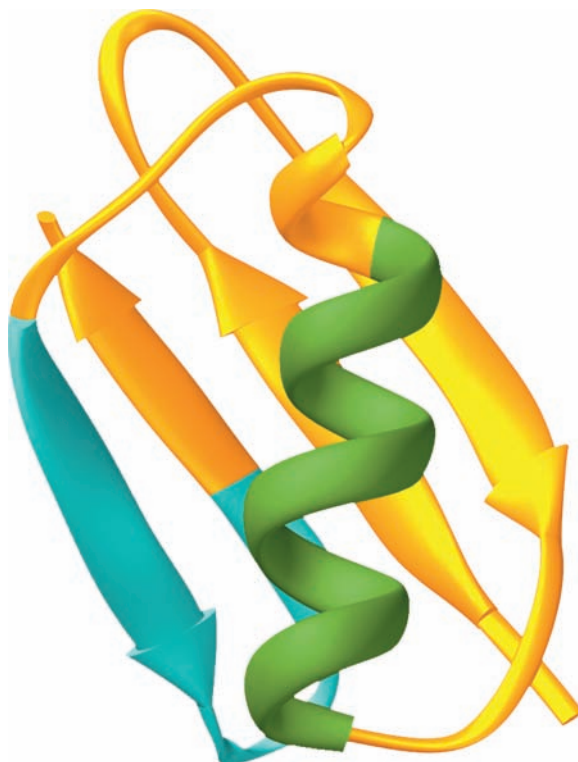


Figure 9-6 NMR structure of protein GB1. Residues 23 to 33 are green and residues 42 to 53 are cyan. The 11-residue chameleon sequence AWTVEKAFKTF can occupy either of these positions without significantly altering the native protein's backbone conformation. [NMR structure by Angela Gronenborn and Marius Clore, National Institutes of Health, Bethesda, Maryland. PDBid 1GB1.]

(EWTYDDATKTF in GB1; a 5-residue change) to yield Chm- β . Both Chm- α and Chm- β display reversible thermal unfolding typical of compact single-domain globular proteins, and their 2D NMR spectra indicate that each assumes a structure similar to that of native GB1. Yet NMR measurements also demonstrate that the isolated chameleon peptide (Ac-AWTVEKAFKTF-NH₂, where Ac is acetyl) is unfolded in solution, which indicates that this sequence has no strong preference for either an α helix or a β sheet conformation. This suggests that the information specifying α helix or β sheet secondary structures can be nonlocal; that is, context-dependent effects may be important in protein folding (but see Section 9-1Ci).

f. Changing the Fold of a Protein

Proteins that share as little as ~20% sequence identity may be structurally similar. To what degree must a protein's sequence be changed in order to convert its fold to that of another protein? This question was answered, at least for the protein GB1, by the finding that changing 50% of its 56 residues converted its fold to that of **Rop protein** (Rop for *repressor of primer*; a transcriptional regulator). Rop is a homodimer whose 63-residue subunits each form an $\alpha\alpha$ motif (Fig. 8-46c) that dimerizes with its 2-fold axis perpendicular to the helix axes to form a 4-helix bundle

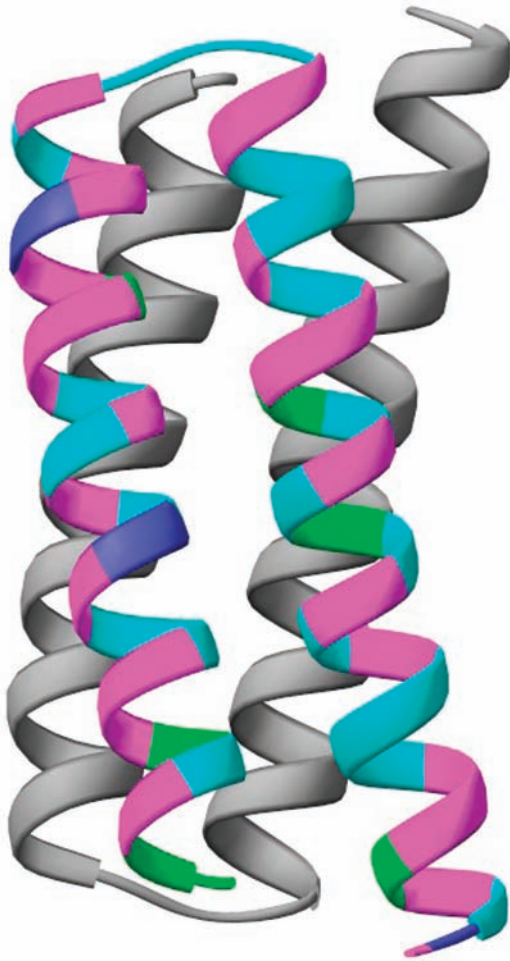


Figure 9-7 X-ray structure of Rop protein, a homodimer of α motifs that associate to form a 4-helix bundle. On a change of 50% of its residues, protein GB1, whose structure is shown in Fig. 9-6, assumes the structure of Rop protein. One of the subunits of the structure shown here is colored according to the sequence of the GB1-derived polypeptide with purple residues identical in both native proteins, magenta residues unchanged from native GB1, cyan residues identical to those in native Rop, and green residues different from those in either native protein. The N-terminus of this subunit is at the lower right. [Based on an X-ray structure by Demetrius Tsernoglou, Università di Roma, Rome, Italy. PDBid 1ROP.]

(Fig. 9-7). Fifty percent of the residues of GB1 were changed based largely on a secondary structure prediction algorithm (Section 9-3Ad), energy minimization, and visual modeling to yield a new polypeptide named Janus (after the two-faced Roman god of new beginnings) that is 41% identical to Rop. In this manner, GB1 residues with high helix-forming propensities were retained, whereas in regions required to be α helical, a number of residues with high β sheet-forming propensities were replaced (helix- and sheet-forming propensities are discussed in Section 9-3Aa); hydrophobic residues were incorporated at the appropriate *a* and *d* positions of a heptad repeat (Fig. 8-26) to form the core of Rop's 4-helix bundle; and residue changes were made to mimic Rop's distribution of surface charges. Fluorescence

and NMR measurements reveal that Janus assumes a stable Rop-like conformation. These studies indicate that not all residues have equally important roles in specifying a particular fold. Indeed, the Janus sequence is more closely related to that of GB1 (50% identity) than to that of Rop (41% identity), even though Janus structurally resembles Rop but not GB1.

9. Many Proteins Are Natively Unfolded

In recent years it has become evident that many entire native proteins and long protein segments (>30 residues) are fully unfolded. Such intrinsically disordered proteins lack specific tertiary structures and are therefore composed of ensembles of conformations. They are characterized by low sequence complexity, a low proportion of the bulky hydrophobic amino acids that form the cores of globular proteins (Val, Leu, Ile, Met, Phe, Trp, and Tyr), and a high proportion of certain polar and charged amino acids (Gln, Ser, Pro, Glu, Lys, Gly, and Ala). Structure prediction techniques based on amino acid sequences (Section 9-3) indicate that an organism's proportion of natively disordered proteins increases with its complexity with ~2% of archeal proteins, ~4% of eubacterial proteins, and ~33% of eukaryotic proteins predicted to contain long disordered regions.

Most natively disordered proteins specifically bind to some other molecule such as a protein, a nucleic acid, or a membrane component, and in doing so fold into stable secondary or tertiary structures. For example, the phosphorylated kinase-inducible domain (pKID) of the transcription factor **cyclic AMP response element-binding protein (CREB)** is disordered when free in solution, but folds to an ordered conformation when it binds to the KID-binding domain of **CREB-binding protein (CBP)**, Fig. 9-8). Apparently, the increased flexibility of natively disordered proteins enables them to perform a relatively unhindered conformational

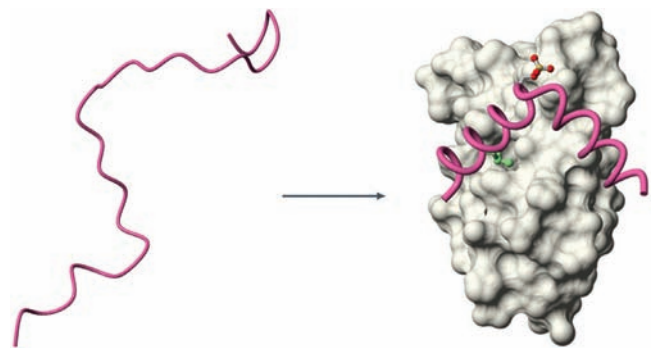


Figure 9-8 The binding of the pKID domain of rat CREB to the KID-binding domain of mouse CBP. The pKID, whose backbone is drawn in worm form (pink), is unstructured when free in solution (left) but forms two perpendicular helices when bound to the KID-binding domain (right). The image on the right shows the NMR structure of the pKID-KID-binding domain complex with the side chains of pKID phosphoSer 133 and Leu 141 drawn in ball-and-stick form with C green, O red, and P yellow and with the KID-binding domain (gray) represented by its solvent-accessible surface. [Courtesy of Peter Wright, Scripps Research Institute, La Jolla, California. PDBid 1KDX.]

search when binding to their target molecules. It has also been suggested that a structured globular protein would have to be two to three times larger than a disordered protein to provide the same size intermolecular interface and hence the use of disordered proteins provides genetic economy and reduces intracellular crowding. Disordered regions may also aid in the transport of proteins across membranes (Section 12-4Ea) and facilitate selective protein degradation (Section 32-6B).

The functions of natively disordered proteins are quite varied. Their most common function appears to be binding to specific DNA sequences to facilitate such processes as replication, transcription, repair, and transposition (Chapter 30). However, they have also been implicated in a variety of other functions including intracellular signal transduction (Chapter 19), forming phosphorylation sites in proteins whose activities are regulated by phosphorylation (Section 18-3C), and in aiding other proteins and RNAs to fold to their native conformations (Section 9-2C).

C. Folding Pathways

How does a protein fold to its native conformation? We, of course, cannot hope to answer this question in detail until we better understand why native protein structures are stable. Moreover, as one might guess, the folding process itself is one of enormous complexity. Nevertheless, as we shall see below, the broad outlines of how proteins fold to their native conformations are beginning to come into focus.

The simplest folding mechanism one might envision is that a protein randomly explores all of the conformations available to it until it eventually “stumbles” onto its native conformation. A back-of-the-envelope calculation first made by Cyrus Levinthal, however, convincingly demonstrates that this cannot possibly be the case: Assume that the $2n$ backbone torsional angles, ϕ and ψ , of an n -residue protein each have three stable conformations. This yields $3^{2n} \approx 10^n$ possible conformations for the protein, which is a gross underestimate, if only because the side chains are ignored. If a protein can explore new conformations at the rate at which single bonds can reorient, it can find $\sim 10^{13}$ conformations per second, which is, no doubt, an overestimate. We can then calculate the time t , in seconds, required for a protein to explore all the conformations available to it:

$$t = \frac{10^n}{10^{13} \text{ s}^{-1}} \quad [9.1]$$

For a small protein of $n = 100$ residues, $t = 10^{87}$ s, which is immensely more than the apparent age of the universe (~ 13.7 billion years = 4.3×10^{17} s).

It would obviously take even the smallest protein an absurdly long time fold to its native conformation by randomly exploring all its possible conformations, an inference known as the **Levinthal paradox**. Yet several proteins fold to their native conformations in microseconds. Therefore, as Levinthal suggested, *proteins must fold by some sort of ordered pathway or set of pathways in which the approach to the native state is accompanied by sharply increasing conformational stability (decreasing free energy).*

a. Rapid Measurements Are Required to Monitor Protein Folding

Folding studies on several small single-domain proteins, including RNase A, cytochrome *c*, and **apomyoglobin** (myoglobin that lacks its heme group), indicate that these proteins fold to a significant degree within one millisecond or less of being brought to native conditions. Hence, if the earliest phases of the folding process are to be observed, denatured proteins must be brought to native conditions in significantly less time. This is most often done using a rapid mixing device such as a **stopped-flow** apparatus (Fig. 9-9) in which a protein solution at a pH that denatures the protein or containing guanidinium chloride or urea at a concentration that does so is rapidly changed in pH or diluted to initiate folding. Most such instruments have “dead times” (the interval between the times when mixing is initiated and meaningful measurements can first be made) of ~ 0.5 ms. However, recently developed ultrarapid mixing devices have dead times of at little as 40 μ s.

An alternative technique involves the refolding of **cold denatured proteins**. [For proteins whose folding has both ΔH and ΔS positive, a decrease in temperature is destabilizing (Table 3-2). Since $\Delta G = \Delta H - T\Delta S$, these proteins are unstable, that is denature, when $T < \Delta H/\Delta S$. For many of these proteins, solution conditions can be found for which this temperature is $>0^\circ\text{C}$.] The refolding of the cold-denatured protein is initiated by a so-called **temperature-jump** in which the solution is heated with an infrared laser pulse by 10 to 30°C in <100 ns.

With either of the above methods, the folding protein must be monitored by a technique that can report rapid structural changes in a protein. The three such techniques that have been most extensively used are (1) **circular dichroism (CD)** spectroscopy, (2) **pulsed HD exchange** followed by 2D-NMR spectroscopy or mass spectrometry, and (3) **fluorescence resonance energy transfer (FRET)**. We discuss these methods below.

b. The Circular Dichroism Spectrum of a Protein Is Indicative of Its Conformation

Polypeptides absorb strongly in the ultraviolet (UV) region of the spectrum ($\lambda = 100$ to 400 nm) largely because their aromatic side chains (those of Phe, Trp, and Tyr) have particularly large molar extinction coefficients (Section 5-3Ca)

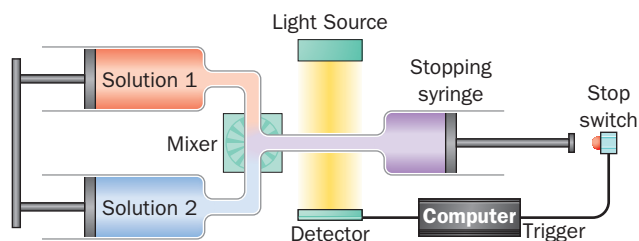


Figure 9-9 A stopped-flow device. The reaction is initiated by simultaneously and rapidly discharging the contents of both syringes through the mixer. On hitting the stop switch, the stopping syringe triggers the computer to commence optically monitoring the reaction (via its UV/visible, fluorescence, or CD spectrum).

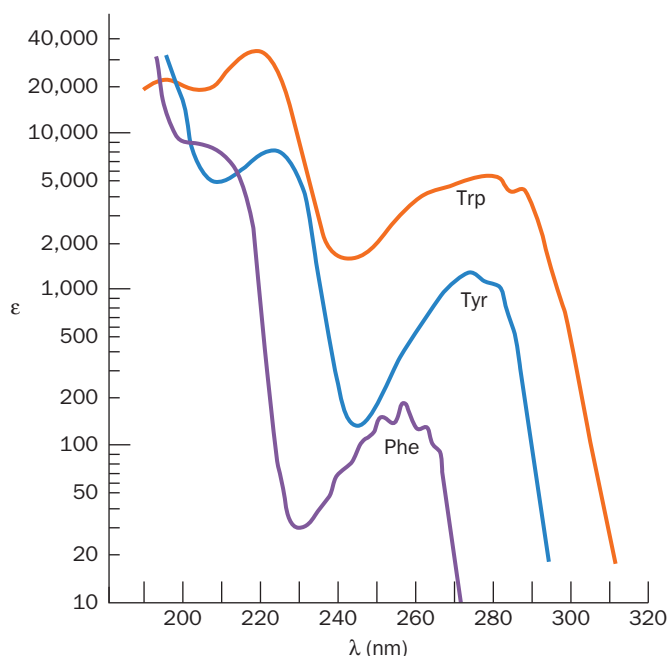


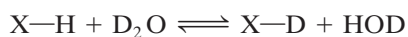
Figure 9-10 UV absorbance spectra of the three aromatic amino acids, phenylalanine, tryptophan, and tyrosine. Note that the molar absorptance, ϵ , is displayed on a log scale. [After Wetlauffer, D.B., *Adv. Prot. Chem.* **7**, 310 (1962).]

in this spectral region (ranging into the tens of thousands; Fig. 9-10). However, polypeptides do not absorb visible light ($\lambda = 400$ to 800 nm), so that they are colorless.

For chiral molecules such as proteins, ϵ has different values for left and right circularly polarized light, ϵ_L and ϵ_R . The variation with λ of the difference in these quantities, $\Delta\epsilon = \epsilon_L - \epsilon_R$, constitutes the **circular dichroism (CD)** spectrum of the solute of interest (for nonchiral molecules $\epsilon_L = \epsilon_R$ and hence they have no CD spectrum). In proteins, α helices, β sheets, and random coils exhibit characteristic CD spectra (Fig. 9-11). Hence the CD spectrum of a polypeptide provides a rough estimate of its secondary structure.

c. Pulsed H/D Exchange Provides Structural Details on How Proteins Fold

Pulsed H/D exchange, a method devised by Walter Englander and Robert Baldwin, is the only known technique that can follow the time course of individual residues in a folding protein. Weakly acidic protons (^1H), such as those of amine and hydroxyl groups (X-H), exchange with those of water, a process known as **hydrogen exchange** that can be demonstrated with the use of deuterated water [D_2O ; deuterium (D or ^2H) is a stable isotope of ^1H]:



Since ^1H has an NMR spectrum in a different frequency range from that of D , the exchange of ^1H for D can be readily followed by NMR spectroscopy. Under physiological conditions, small organic molecules, such as amino acids and dipeptides, completely exchange their weakly acidic protons for D in times ranging from milliseconds to seconds.

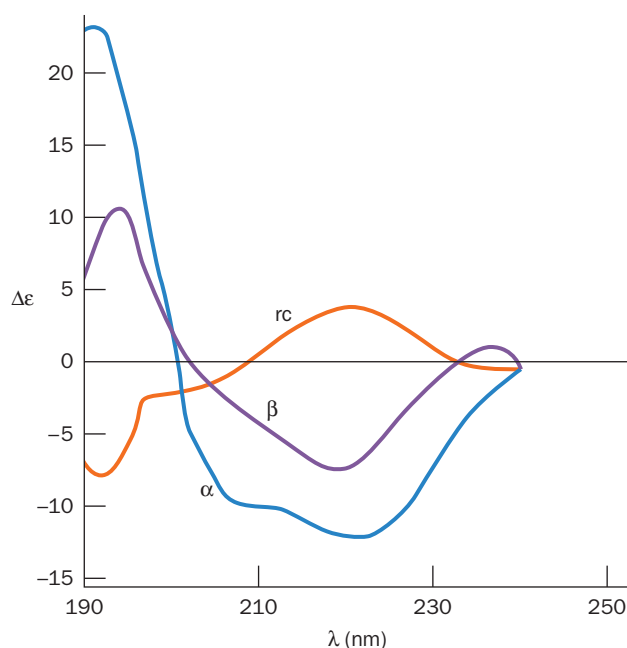


Figure 9-11 Circular dichroism (CD) spectra of polypeptides. Polypeptides in the α helix, β sheet, and random coil (rc) conformations were determined from the CD spectra of proteins of known X-ray structures. By comparing these spectra with the absorption spectra in Fig. 9-10, it can be seen that $\Delta\epsilon = \epsilon_L - \epsilon_R$ is a small difference of two large numbers. [After Saxena, V.P. and Wetlauffer, D.B., *Proc. Natl. Acad. Sci.* **66**, 971 (1971).]

Proteins bear numerous exchangeable protons, particularly those of its backbone amide groups. However, protons that are engaged in hydrogen bonding do not exchange with solvent and, moreover, groups in the interior of a native protein are not in contact with solvent.

Through the use of 2D-NMR (Section 8-3Ac), pulsed H/D exchange can be used to follow the time course of protein folding. The protein of interest, usually with its native disulfide bonds intact, is denatured by guanidinium chloride or urea in D_2O solution such that all of the protein's peptide nitrogen atoms become deuterated (N-D). Folding is then initiated in a stopped-flow apparatus by diluting the denaturant solution with $^1\text{H}_2\text{O}$ while the pH is simultaneously lowered so as to arrest hydrogen exchange (near neutrality, hydrogen exchange reactions are catalyzed by OH^- and, therefore, their rates are highly pH dependent). After a preset folding time, t_f , the pH is rapidly increased (using a third independently triggered syringe; the so-called labeling pulse) to initiate hydrogen exchange. Peptide nitrogen atoms whose D atoms have not formed hydrogen bonds by time t_f exchange with ^1H , whereas those that are hydrogen bonded at t_f , and hence unavailable for hydrogen exchange, remain deuterated. After a short time (10 to 40 ms), the labeling pulse is terminated by rapidly lowering the pH (with a fourth syringe). Folding is then allowed to go to completion and the H/D ratio at each exchangeable site is determined by 2D-NMR (the peaks in the 2D proton NMR spectrum must have been previously assigned). By repeating the analysis for several values of t_f , the

time course of hydrogen bond formation at each residue can be determined.

Pulsed H/D exchange–NMR studies do not directly indicate the structures of the folding intermediates. However, if the native structure of the protein under investigation is known (as it almost always is for proteins whose folding is being investigated) and if it is assumed that the protein folds without forming secondary structures not present in the native protein, then the 2D-NMR spectra reveal the time course of the formation of the elements of the native structure together with how fast they are excluded from the bulk solvent.

The time course of protein folding may also be followed by combining pulsed H/D exchange with mass spectrometry. In this method, a partially deuterated protein is fragmented by pepsin (a protease that functions under the acidic conditions necessary to prevent further hydrogen exchange; Table 7-2), the resulting fragments separated by HPLC, and their degree of deuteration determined by mass spectrometry. This method does not yield the residue-level structural information that NMR provides. However, unlike NMR, it can determine if a sample contains subpopulations of protein fragments with different degrees of deuteration and hence that have followed different folding pathways.

d. Fluorescence Resonance Energy Transfer Monitors Distances

Fluorescence is the phenomenon whereby an electronically excited molecule or group decays to its ground state by emitting a photon. The initial excited state rapidly decays, via nonradiative processes (e.g., heating; Section 24-2Aa), to an excited state of lower energy before the photon is emitted. Hence the molecule or group's emission spectrum has a longer wavelength than its absorption spectrum (Fig. 9-12).

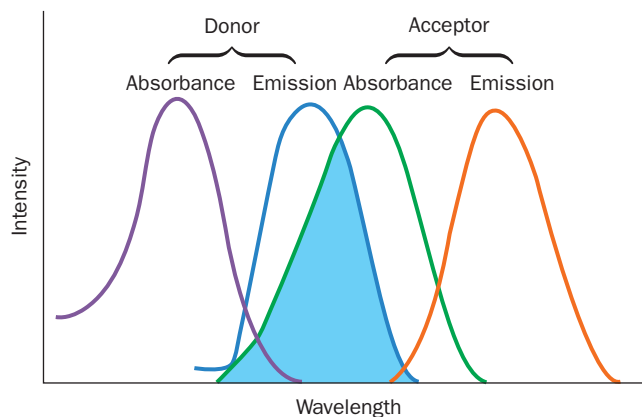
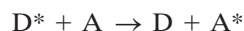


Figure 9-12 Schematic diagram of the absorption and emission spectra of a donor and an acceptor in fluorescence resonance energy transfer (FRET). Note that the absorption spectrum occurs at shorter wavelengths than the corresponding emission spectrum and that the donor's emission spectrum must overlap the acceptor's absorption spectrum (cyan) for FRET to occur.

When two fluorescent molecules or groups, a donor (D) and an acceptor (A), are within 100 Å of each other and D is electronically excited (say, by a laser with a wavelength within its absorption spectrum), some of the excitation energy will be transferred from D to A,



where the asterisk indicates an electronically excited state. A will then fluoresce with its characteristic emission spectrum (Fig. 9-12), whose intensity can be measured. This phenomenon is known as **fluorescence resonance energy transfer (FRET)**. Its efficiency E , the fraction of the energy transferred to the acceptor per donor excitation event, is given by

$$E = \frac{1}{1 + (r/R_0)^6} \quad [9.2]$$

where r is the distance between D and A, and R_0 , their **Förster distance** (named after Theodor Förster, who formulated the theory for the mechanism of long-range nonradiative energy transfer), is the value of r at which the FRET efficiency is 50%. R_0 varies with the degree of overlap between the donor's emission spectrum and the acceptor's absorption spectrum (Fig. 9-12) as well as with the relative orientation of the donor and acceptor. Hence, the intensity of the acceptor's fluorescence is indicative of the distance between D and A as well as their relative orientation.

In proteins, D and A can be the side chains of Trp or Tyr residues. The number and placement of these residues in the protein of interest can be manipulated by site-directed mutagenesis (Section 5-5Gc). Alternatively, fluorescent groups may be covalently linked to reactive side chains such as Cys, which can also be placed via site-directed mutagenesis. FRET measurements can then be used to track how the distances between specific residues vary with time in a folding protein.

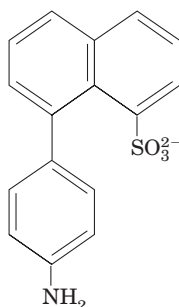
e. The Earliest Protein Folding Events Are Initiated by a Hydrophobic Collapse

Stopped-flow–CD measurements indicate that *for many, if not all, small single-domain proteins, much of the secondary structure that is present in native proteins forms within a few milliseconds of when folding is initiated.* This is called the **burst phase** because subsequent folding events occur over much longer time intervals. Pulsed H/D exchange measurements of these small proteins show that some protection against hydrogen exchange in some secondary structural elements develops by ~5 ms after folding initiation.

Since globular proteins contain a compact hydrophobic core, it seems likely that the driving force in protein folding is a so-called **hydrophobic collapse**, in which the protein's hydrophobic groups coalesce so as to expel most of their surrounding water molecules. The polypeptide's radius of gyration is thereby dramatically reduced (from ~30 to ~15 Å for a 100-residue polypeptide), a phenomenon that is

generally characteristic of polymers on being transferred from a good to a poor solvent.

This hydrophobic collapse mechanism is consistent with the observation that the hydrophobic dye **8-anilino-1-naphthalenesulfonate (ANS)**



8-Anilino-1-naphthalene sulfonate (ANS)

binds to folding proteins. ANS undergoes a significant enhancement of its fluorescence when it occupies a nonpolar environment, an enhancement that is observed within the burst phase when ANS is present in a solution of a folding protein. Since ANS is expected to preferentially bind to hydrophobic groups, this indicates that the hydrophobic core of a protein rapidly forms once folding has been initiated.

The initial collapsed state of a folding protein is known as a **molten globule**. Such a species has a radius of gyration that is only 5 to 15% greater than that of the native protein and has significant amounts of the native secondary structure and overall fold. However, a molten globule's side chains are extensively disordered, its structure fluctuates far more than that of the native protein, and it has only marginal thermodynamic stability. Nevertheless, to continue folding toward its native state, the polypeptide chain need not undergo large rearrangements in the crowded core of the partially folded protein.

f. Nativelike Tertiary Structure Appears During Intermediate Folding Events

After the burst phase, small proteins exhibit increased ANS binding, further changes in their CD spectrum, and enhanced protection against H/D exchange. These intermediate folding events typically occur over a time interval of 5 to 1000 ms. This is the stage at which the protein's secondary structure becomes stabilized and its tertiary structure begins to form. These nativelike elements are thought to take the form of subdomains that are not yet properly docked to each other. Side chains are probably still mobile, so that, at this stage of folding, the protein can be described as an ensemble of closely related and rapidly interconverting structures.

g. Final Folding Events Often Require Several Seconds

In the final stage of folding, a protein achieves its native structure. To do so, the polypeptide must undergo a series of complex motions that permit the attainment of the relatively

rigid native core packing and hydrogen bonding, while expelling the remaining water molecules from its hydrophobic core. For small single-domain proteins, this takes place over a time interval of several seconds or less.

h. Landscape Theory of Protein Folding

The classic view of protein folding was that proteins fold through a series of well-defined intermediates. The folding of a random coil polypeptide was thought to begin with the random formation of short stretches of 2° structure, such as α helices and β turns, that acted as **nuclei** (scaffolding) for the stabilization of additional ordered regions of the protein. Nuclei with the proper nativelike structure then grew by the diffusion, random collision, and adhesion of two or more such nuclei. The stabilities of these ordered regions were thought to increase with size, so, after having randomly reached a certain threshold size, they spontaneously grew in a cooperative fashion until they formed a nativelike domain. Finally, through a series of relatively small conformational adjustments, the domain rearranged to the more compact 3° structure of the native conformation.

The advent of experimental methods that could observe early events in protein folding led to a somewhat different view of how proteins fold. In this so-called **landscape theory**, which was formulated in large part by Peter Wolynes, Baldwin, and Dill, folding is envisioned to occur on an **energy surface** or landscape that represents the conformational energy states available to a polypeptide under the prevailing conditions. The horizontal coordinates of a point on this surface represent a particular conformation of the polypeptide, that is, the values of ϕ and ψ for each of its amino acid residues and the torsion angles for each of its side chains (but here projected onto two dimensions from its multidimensional space). The vertical coordinate of a point on the energy surface represents the polypeptide's internal free energy in this conformation. The above-described measurements indicate that the energy surface of a folding polypeptide is funnel-shaped, with the native state represented by the bottom of the funnel, the global (overall) free energy minimum (Fig. 9-13a). The width of the funnel at any particular height (free energy) above the native state is indicative of the number of conformational states with that free energy, that is, the entropy of the polypeptide.

Polypeptides fold via a series of conformational adjustments that reduce their free energy and entropy until the native state is reached. Since a collection of unfolded polypeptides all have different conformations (have different positions on the folding funnel), they cannot follow precisely the same pathway in folding to the native state. If the polypeptide actually folded to its native state via a random conformational search, as Levinthal conjectured, its energy surface would resemble a flat disk with a single small hole, much like the surface of a golf course (Fig. 9-13b). Thus, it would take an enormously long time for a polypeptide (a golf ball) to achieve the native state (to fall in the hole) via a random conformational search (by rolling about aimlessly on the surface of the golf course).

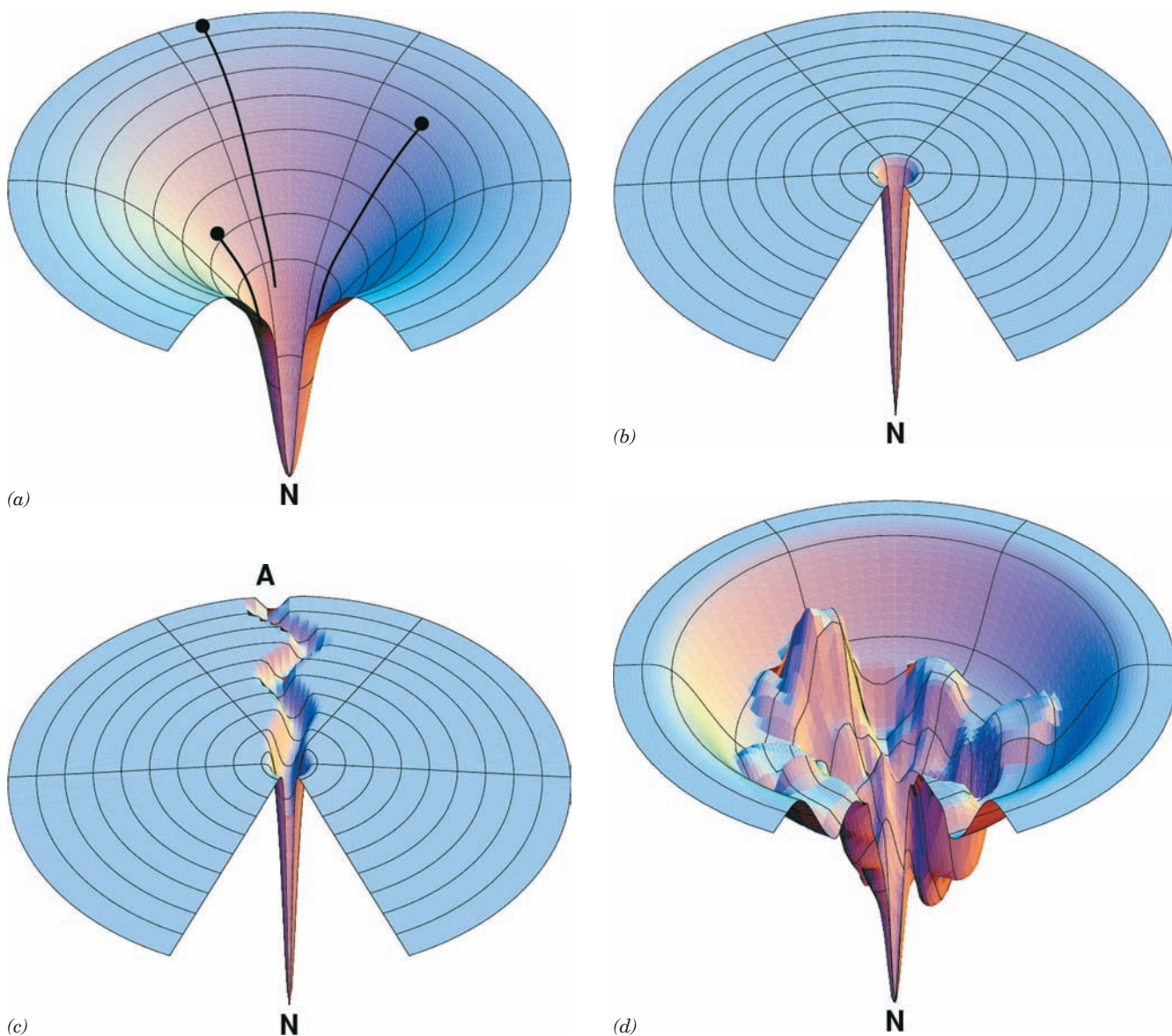


Figure 9-13 Folding funnels. (a) An idealized funnel landscape. As the chain forms increasing numbers of intrachain contacts, its internal free energy (its height above the native state, N) decreases together with its conformational freedom (the width of the funnel). Polypeptides with differing conformations (black dots) follow different pathways (black lines) in achieving the native fold. (b) The Levinthal “golf course” landscape in which the chain must search for the native fold (the hole)

randomly, that is, on a level energy surface. (c) The classic folding landscape in which the chain must search at random on a level energy surface until it encounters the canyon that leads it to the native state. (d) A rugged energy surface containing local minima in which a folding polypeptide can become transiently trapped. The folding funnels of real proteins are thought to have such topographies. [Courtesy of Ken Dill, University of California at San Francisco.]

The energy surface of a protein that follows the classic view of protein folding would have a deep radial groove in its disklike surface that slopes toward the hole representing the native state (Fig. 9-13c). The extent of the conformational search to randomly find this groove would be much reduced relative to the Levinthal model, so that such a polypeptide would readily fold to its native state. However, the conformational search for the pathway (groove) leading to the native state would still take time, so that the polypeptide would require perhaps several seconds to start down the folding pathway.

The observation that many polypeptides acquire significant natively like structure within fractions of a millisecond after folding commences indicates that their energy surfaces are, in fact, funnel-shaped; that is, they tend to slope toward the native conformation at all points. Thus, the various pathways followed by initially unfolded polypeptides in folding to their native state are analogous to the various trajectories that could be taken by skiers initially distributed around the top of a bowl-shaped valley to reach the valley’s lowest point. Apparently, *there is no single pathway or closely related set of*

pathways that a polypeptide must follow in folding to its native conformation.

The foregoing does not imply that the surface of the folding funnel is necessarily smooth, as is drawn in Fig. 9-13a. Indeed, landscape theory suggests that this energy surface has a relatively rugged topography, that is, has many local energy minima and maxima (Fig. 9-13d). Consequently, in following any particular folding pathway, a polypeptide is likely to become trapped in a local minimum until it randomly acquires sufficient thermal energy to surmount this kinetic barrier and continue the folding process. Thus, in landscape theory, the local energy maxima (transition states; Section 14-1C) that govern the rate of folding are not specific structures as the classic theory of protein folding suggests but, rather, are ensembles of structures.

i. Protein Folding Is Hierarchical

The observation that protein structures are hierarchically organized (Section 9-1B) suggests that they also fold in a hierarchic manner. By this it is meant that folding begins with the formation of marginally stable nativelylike microstructures known as **foldons** (e.g., Fig. 9-14) that are local in sequence and that these foldons diffuse and collide with nearby (in sequence) foldons to yield intermediates of increasing complexity and stability that sequentially grow to form the native protein. In contrast, in nonhierarchical folding, a protein's tertiary structure would not only stabilize its local structures but also determine them. Landscape theory is consistent with hierarchical folding, whereas the classic theory of protein folding is more in accord with nonhierarchical folding. Moreover, since a polypeptide *in vivo* begins folding as it is being synthesized, that is, as it is extruded from the ribosome, it would seem that it would most readily achieve its native state if it folded in a hierarchical manner.

Several lines of evidence indicate that proteins, in fact, fold in a hierarchical manner.

1. H/D exchange studies have established the existence of foldons in numerous proteins. Indeed, it appears that foldons rather than individual amino acid residues carry out the unit steps in protein folding pathways.

2. Many peptide fragments excised from proteins either form or exhibit a tendency to form foldons in the absence of long-range (3°) interactions. Moreover, when proteins such as cytochrome *c* and apomyoglobin are brought to a pH sufficiently low to destabilize their native structures, their foldons persist.

3. The boundaries of helices in native proteins are fixed by their flanking sequences (Section 9-3) rather than by 3° interactions.

4. The folding rates of proteins increase, on average, with the degree to which their native contacts are local. Thus fast folders tend to have a high proportion of helices and tight turns, whereas slow folders tend to have a high proportion of β sheets.

In Section 9-1B we saw that in protein GB1 (Fig. 9-6), the 11-residue “chameleon” sequence assumed either an α helix or a β hairpin, depending on its position in the protein.

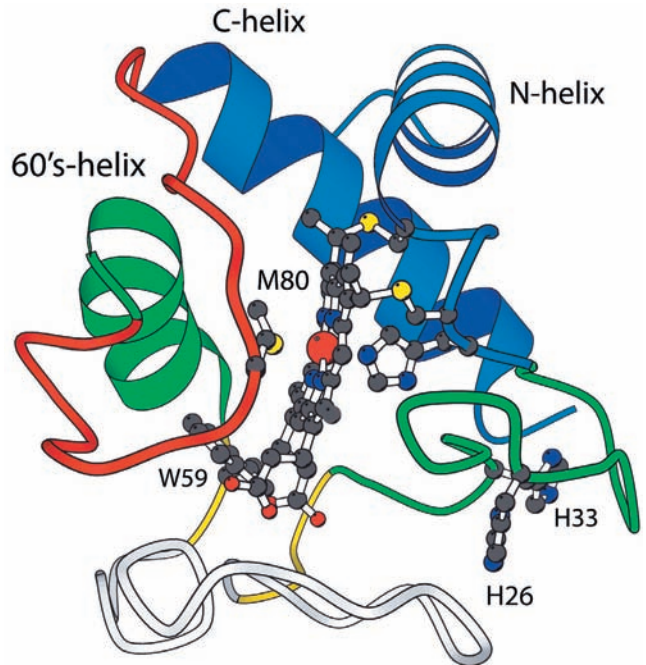


Figure 9-14 Ribbon diagram of cytochrome *c*. Its several foldon units are shown in different colors. Its heme group and several of its functionally important side chains are drawn in ball-and-stick form with C black, N blue, O red, S yellow, and Fe a large red sphere. [Courtesy of Walter Englander, University of Pennsylvania.]

Thus, its conformation appears to be determined by its context rather than by local interactions. However, computer simulations suggest that the conformation of the chameleon sequence is actually determined by local interactions beyond its boundaries.

The folds of native proteins, as we have seen, are highly resistant to sequence changes. Evidently, *the sequence information specifying a particular fold is both distributed throughout the polypeptide chain and highly overdetermined*. It is these characteristics that appear to be responsible for hierarchical folding.

j. Primary Structures Determine Protein Folding Pathways as Well as Structures

The above discussions suggest that *protein primary structures evolved to specify efficient folding pathways as well as stable native conformations*. Evidence corroborating this hypothesis has been obtained by Jonathan King in his study of the renaturation of the **tail spike protein** of bacteriophage P22. The tail spike protein is a trimer of identical 76-kD polypeptides, whose $T_m = 88^\circ\text{C}$. However, certain mutant varieties of the protein fail to renature at 39°C . Nevertheless, at 30°C , these mutant proteins fold to structures whose properties, including their T_m 's, are indistinguishable from that of the wild-type tail spike protein. The amino acid changes causing these temperature-sensitive folding mutations apparently act to destabilize intermediate states in the folding process but do not affect the native protein's stability. This observation suggests that *a protein's amino acid sequence dictates its native structure by specifying how it folds to its native conformation*.

This hypothesis is supported by the observation that, in native proteins, a greater number of polar residues than would be randomly expected occupy helix-capping positions (Section 8-4Bb) even though they do not make helix-capping hydrogen bonds. This suggests that they do so as the helix forms so as to facilitate the protein's proper folding.

2 FOLDING ACCESSORY PROTEINS

Most unfolded proteins renature *in vitro* over periods ranging from minutes to days and, quite often, with low efficiency,

that is, with a large fraction of the polypeptide chains assuming quasi-stable non-native conformations and/or forming nonspecific aggregates. *In vivo*, however, polypeptides efficiently fold to their native conformations as they are being synthesized, a process that normally requires a few minutes or less. This is because all cells contain three types of accessory proteins that function to assist polypeptides in folding to their native conformations and in assembling to their 4° structures: **protein disulfide isomerases**, **peptidyl prolyl cis-trans isomerases**, and **molecular chaperones**. We discuss these essential proteins in this section.

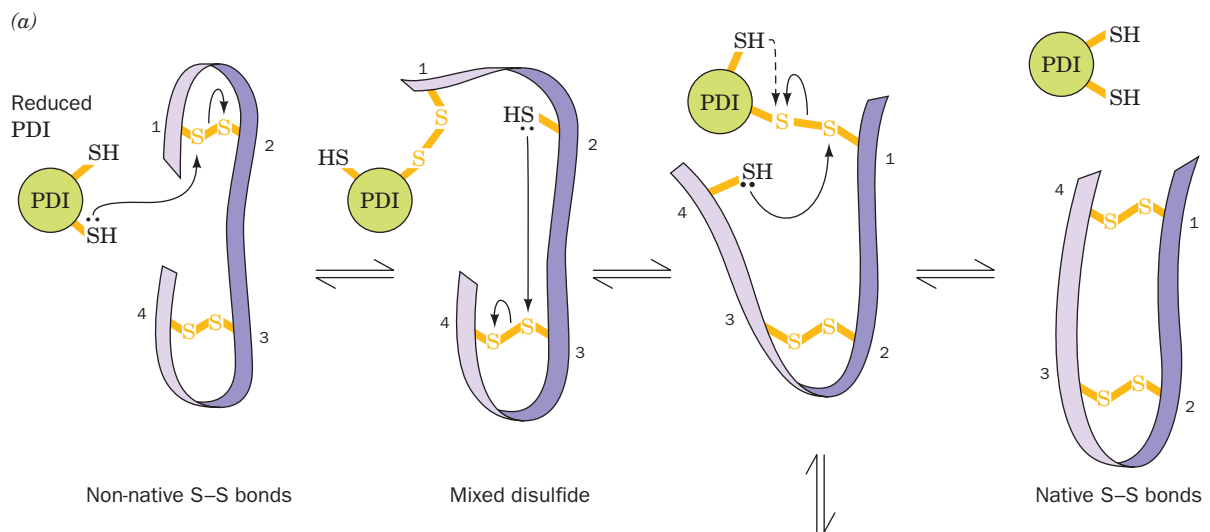
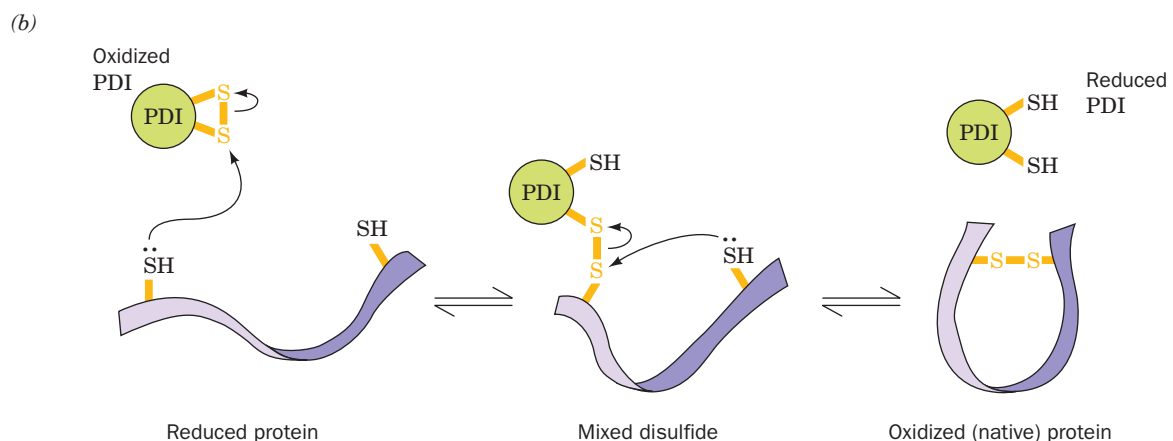


Figure 9-15 Reactions catalyzed by protein disulfide isomerase (PDI). (a) Reduced PDI catalyzes the rearrangement of the non-native disulfide bonds in a substrate protein (purple ribbon) via disulfide interchange to yield the native disulfide bonds (horizontal reactions). If a disulfide bond between PDI and the substrate protein is resistant to disulfide interchange, it is reduced by PDI's second SH group to yield reduced substrate protein and oxidized PDI (vertical reaction and dashed curved arrow). (b) The oxidized PDI-dependent synthesis of disulfide bonds in proteins. The reaction occurs with the intermediate formation of a mixed disulfide between PDI and the protein. The reduced PDI reaction product reacts with cellular oxidizing agents to regenerate oxidized PDI. See the Animated Figures



A. Protein Disulfide Isomerase

Protein disulfide isomerase (PDI), which we encountered in Section 9-1A, is an ~510-residue eukaryotic enzyme that inhabits the lumen of the endoplasmic reticulum, where disulfide-containing proteins fold and are post-translationally processed (Section 12-4B). In its reduced form, PDI catalyzes disulfide interchange reactions, thereby facilitating the shuffling of the disulfide bonds in proteins (Fig. 9-15a, *horizontal reactions*) until they achieve their native pairings, which are resistant to further rearrangement. Moreover, PDI must facilitate the correct folding of those proteins that denature in the absence of their native disulfide bonds. Intriguingly, PDI is also the β subunit of the $\alpha_2\beta_2$ heterotetramer prolyl hydroxylase, the enzyme that hydroxylates the Pro residues of collagen (Section 8-2B). The significance of this latter finding is unknown.

Sequence comparisons indicate that PDI contains four ~100-residue domains that are arranged, from N- to C-terminus, as a–b–b'–a', in which domains a and a' are homologs that are 30% identical in sequence. They are also homologous to the ubiquitous disulfide-containing redox protein **thioredoxin** (Section 28-3Ae), and hence belong to the thioredoxin superfamily. Prokaryotes have enzymes with functions similar to those of PDI that also assume the thioredoxin fold.

PDI's a and a' domains each contain the active site sequence motif Cys-Gly-His-Cys, in which the first Cys residue, in its —SH form, participates in the disulfide interchange reaction diagrammed in Fig. 9-15a (the catalytic motif of the thioredoxin superfamily is Cys-X-X-Cys, where X is any amino acid residue). If the second Cys residue is mutated, PDI's isomerization activity drops to

<1% of the wild type and it accumulates in disulfide-linkage to substrate proteins. This suggests that this second Cys residue functions, in its —SH form, to release PDI from the otherwise stable disulfide bonds that its first Cys residue occasionally forms with substrate proteins, thereby yielding reduced substrate proteins and PDI with a disulfide bond linking its two active site Cys residues (Fig. 9-15a, *vertical reaction*).

The X-ray structure of yeast PDI, determined by William Lennarz and Hermann Schindelin, reveals that it adopts a twisted U-shape in which the N-terminal active site Cys S atoms of the a and a' domains face each other across the top of the U at a distance of 28 Å (Fig. 9-16). As expected, a and a' have folds that are similar to each other (Fig. 9-17, *top*) and to those of other members of the thioredoxin superfamily. Surprisingly, although b and b' exhibit no significant sequence similarity with a and a' or with each other, they also adopt the thioredoxin fold (Fig. 9-17, *bottom*). Nevertheless, both b and b' lack Cys residues and hence cannot participate directly in the catalytic reaction. The b and b' domains share an extensive interface (burying ~700 Å²) and hence appear to be rigidly linked together, whereas the a–b and a'–b' interfaces are negligibly small (burying ~200 Å²). This suggests that the a and a' domains are flexibly linked to a rigid base formed by the b and b' domains, thereby enabling PDI to accommodate a diverse set of substrates of up to ~100 residues within the U.

The inner face of the U has a continuous hydrophobic surface that also surrounds the a and a' active sites. This surface appears to be essential for the binding of PDI to its substrate proteins, which tend to be partially or fully unfolded

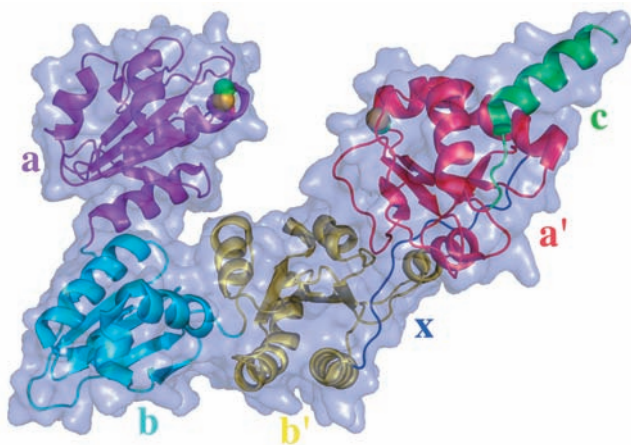


Figure 9-16 X-ray structure of yeast protein disulfide isomerase (PDI). The protein is represented by its transparent molecular surface with its polypeptide chain in ribbon form with its a, b, b', and a' domains colored magenta, cyan, yellow, and red, respectively. The 16-residue loop, x, linking the b' and a' domains is blue and the C-terminal extension, c, is green. The side chains of the N-terminal active site Cys residues of the a and a' domains are drawn in space-filling form with C green and S yellow. [Based on an X-ray structure by William Lennarz and Hermann Schindelin, State University of New York, Stony Brook, New York. PDBid 2B5E.]

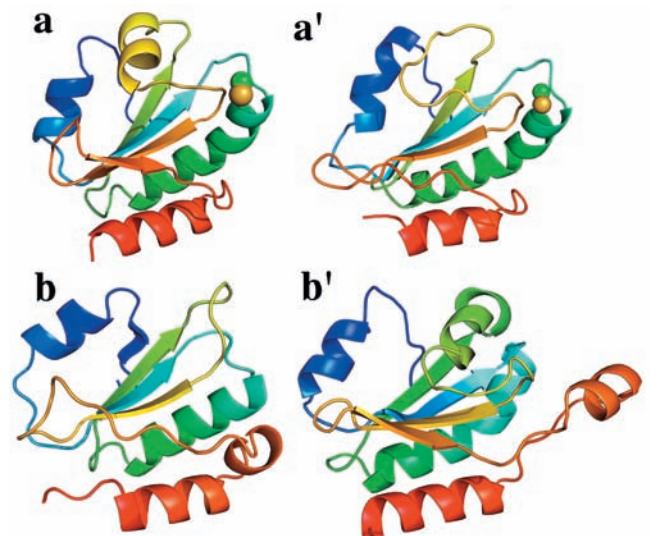


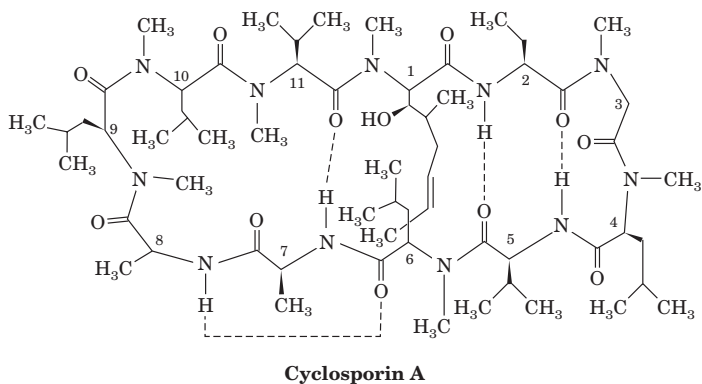
Figure 9-17 Structural comparison of the a, b, b', and a' domains of yeast PDI. The domains are shown in similar orientations and drawn in ribbon form colored in rainbow order from their N-terminus (blue) to their C-terminus (red). The side chains of the N-terminal active site Cys residues of the a and a' domains are drawn in space-filling form with C green and S yellow. [Based on an X-ray structure by William Lennarz and Hermann Schindelin, State University of New York, Stony Brook, New York. PDBid 2B5E.]

and hence have exposed hydrophobic groups. Moreover, as we shall see in Section 9-2C, PDI's hydrophobic surface facilitates the proper folding of its unfolded substrate proteins. Efficient catalysis of disulfide bond rearrangement requires that reduced PDI be intact, thus suggesting that PDI's two active sites act in concert. The isomerase reaction is driven by the release of conformational strain in the unfolded substrate protein as it folds to its native conformation.

Disulfide bonds in native proteins are usually buried and frequently occur in hydrophobic environments. Indeed, it is probably the burial of the correctly paired Cys residues in a native protein that terminates the action of PDI. However, the N-terminal S atoms in the both the a and a' active sites of PDI are exposed on the protein surface. Although their disulfide bonds almost always stabilize proteins (Section 8-4D) and are usually unreactive, oxidized a and a' are less stable than their reduced forms and therefore have highly reactive, that is, strongly oxidizing, disulfide bonds. This permits oxidized PDI to directly introduce disulfide bonds into newly synthesized and hence reduced polypeptides via a disulfide interchange mechanism (Fig. 9-15b). For this latter process to continue, reduced PDI must be reoxidized (its disulfide bond reformed) by cellular oxidizing agents.

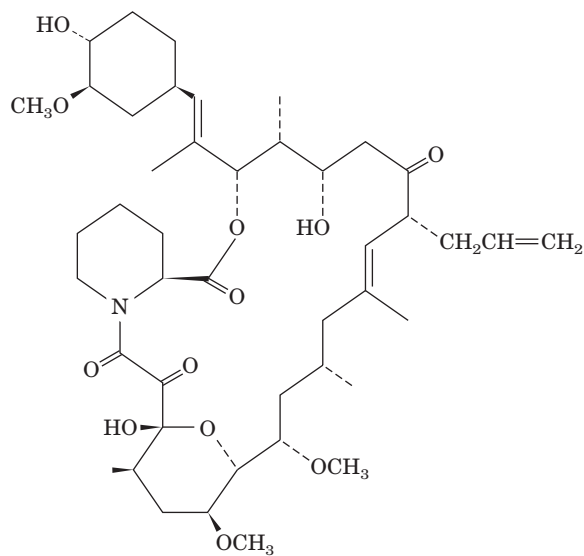
B. Peptidyl Prolyl Cis-Trans Isomerase

Although polypeptides are probably biosynthesized with almost all of their X-Pro peptide bonds (where X is any amino acid residue) in the trans conformation, ~10% of these bonds assume the cis conformation in globular proteins because, as we have seen in Section 8-1A, the energy difference between their cis and trans conformations is relatively small. **Peptidyl prolyl cis-trans isomerases (PPIs;** alternatively known as **rotamases**) catalyze the otherwise slow interconversion of X-Pro peptide bonds between their cis and trans conformations, thereby accelerating the folding of Pro-containing polypeptides. Two structurally unrelated families of PPIs, collectively named the **immunophilins**, have been characterized: the **cyclophilins** (so named because they are inhibited by the immunosuppressive drug **cyclosporin A**,



a fungally produced 11-residue cyclic peptide) and the fam-

ily for which the 12-kD **FK506 binding protein (FKBP12)** is prototypic (**FK506**



is a fungally produced macrocyclic lactone that is also an immunosuppressive drug; medicinal chemists tend to identify the often huge numbers of related drug candidates they deal with by serial numbers rather than by trivial names).

The X-ray structure of human cyclophilin in complex with succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide reveals that this model substrate binds to the enzyme with its Ala-Pro peptide bond in the cis conformation and that it could not do so if it had the trans conformation. This suggests that the enzyme predominantly catalyzes the trans to cis isomerization of peptidyl-prolyl amide bonds. In addition, the Arg 55 → Ala mutation in cyclophilin reduces its enzymatic activity 100-fold. This, together with the observation that Arg 55 is positioned so that it could hydrogen bond to the N atom of the Ala-Pro peptide bond (although it does not do so in the crystal structure) suggests that the formation of a hydrogen bond from Arg 55 to this N atom facilitates the cis-trans isomerization by deconjugating and hence weakening the peptidyl-prolyl amide bond.

a. Cyclosporin A and FK506 Are Clinically Important Immunosuppressive Agents

Cyclosporin A and FK506 are highly effective agents for the treatment of autoimmune disorders and for preventing organ-transplant rejection. Indeed, until the advent of cyclosporin A in the early 1980s, the long-term survival of a transplanted organ (and its recipient) was a rare occurrence. The more recently discovered FK506 is an even more potent immunosuppressant. The immunosuppressive properties of both cyclosporin A and FK506 stem from the abilities of their respective complexes with cyclophilin and FKBP12 to prevent the expression of genes involved in the activation of **T lymphocytes** (the immune system cells responsible for **cellular immunity**; the immune response is discussed in Section 35-2) by interfering with these cells'

intracellular signaling pathways. Enigmatically, there is no obvious relationship between the immunophilins' immunosuppressive properties and rotamase activities: Both cyclosporin A and FK506 are effective immunosuppressants at concentrations far below those of the cyclophilin and FKBP12 in cells; and mutational changes that destroy cyclophilin's rotamase activity do not eliminate its ability to bind cyclosporin A or the ability of the resulting complex to interfere with T lymphocyte signaling. This conundrum is explained in Section 19-3Ff.

C. Molecular Chaperones: The GroEL/ES System

Newly synthesized and hence unfolded proteins contain numerous solvent-exposed hydrophobic groups. Moreover, proteins *in vivo* fold in the presence of extremely high concentrations of other macromolecules (~300 g/L, which occupy ~25% of the available volume). Consequently, unfolded proteins *in vivo*, particularly larger proteins (those >15 kD), have a great tendency to form both intramolecular and intermolecular aggregates. **Molecular chaperones**, which are also known as **heat shock proteins** (so named because their rates of synthesis increase at elevated temperatures), are proteins that function to prevent or reverse such improper associations, particularly in multidomain and multisubunit proteins. They do so by binding to an unfolded or aggregated polypeptide's solvent-exposed hydrophobic surfaces and subsequently releasing them, often repeatedly, in a manner that facilitates their proper folding and/or 4° assembly. Most molecular chaperones are **ATPases** (enzymes that catalyze ATP hydrolysis), which bind to unfolded polypeptides and apply the free energy of ATP hydrolysis to effect their release in a favorable manner. Thus it appears, as John Ellis has pointed out, that molecular chaperones function analogously to their human counterparts: *They inhibit inappropriate interactions between potentially complementary surfaces and disrupt unsuitable liaisons so as to facilitate more favorable associations.*

The molecular chaperones comprise several unrelated classes of proteins that have somewhat different functions including:

1. The **heat shock proteins 70 (Hsp70)**, which are ~70-kD monomeric proteins that are highly conserved in both prokaryotes and eukaryotes (in which different species occur in the cytosol, the endoplasmic reticulum, mitochondria, and chloroplasts; the *E. coli* Hsp70 is called **DnaK** because it was discovered through the isolation of mutants that do not support the growth of bacteriophage λ and hence was initially thought to participate in DNA replication). They function in an ATP-driven process to reverse the denaturation and aggregation of proteins (processes that are accelerated at elevated temperatures), to facilitate the proper folding of newly synthesized polypeptides as they emerge from the ribosome, to unfold proteins in preparation for their transport through membranes (Section 12-4Ea), and to subsequently help them refold. Hsp70 works in association with the **cochaperone** protein **Hsp40**

(**DnaJ** in *E. coli*) to bind and release small hydrophobic regions of misfolded proteins.

2. **Trigger factor**, which is a ribosome-associated prokaryotic protein. It prevents the intra- and intermolecular aggregation of newly synthesized polypeptides as they emerge from the ribosome by shielding their hydrophobic segments. Unlike most other chaperones, trigger factor does not bind ATP. Trigger factor and the Hsp70/40 system appear to have redundant functions: *E. coli* can tolerate the loss of either one but the loss of both is lethal above 30°C and is accompanied by the massive aggregation of newly synthesized proteins. Trigger factor and Hsp70/40 are the first chaperones that newly synthesized polypeptides encounter. Subsequently, many of the resulting partially folded proteins are handed off to other chaperones, such as those listed below, to complete the folding process. Eukaryotes lack a homolog of trigger factor but contain other small chaperones that may have similar functions.

3. The **chaperonins**, which are heat shock proteins that form large, multisubunit, cage-like assemblies that are universal components of prokaryotes and eukaryotes. They bind improperly folded globular proteins via their exposed hydrophobic surfaces and then, in an ATP-driven process, induce the protein to fold while enveloping it in an internal cavity, thereby protecting the folding protein from nonspecific aggregation with other unfolded proteins (see below). There are two classes of chaperonins: the **Group I chaperonins**, which occur in eubacteria, mitochondria, and chloroplasts, and the **Group II chaperonins**, which occur in archaea and eukaryotes.

4. The **Hsp90** proteins, which are homodimeric, ATP-dependent, eukaryotic proteins of ~730-residue subunits that mainly facilitate the late stage folding of proteins involved in signaling, including **steroid hormone receptors** (Section 34-3Bn) and **receptor tyrosine kinases** (Section 19-3A). Like other chaperones, they do so by binding to exposed hydrophobic surfaces of their substrate proteins so as to prevent nonspecific aggregation. Unlike other chaperones, however, Hsp90 proteins have a regulatory role in that they induce conformational changes in natively substrate proteins that result in their activation or stabilization. They do so through their interactions with a large variety of cochaperones. Hsp90 proteins are among the most abundant proteins in eukaryotes, constituting 1 to 2% of their soluble proteins under normal conditions and 4 to 6% under stressful conditions that destabilize proteins such as high temperatures.

5. The **nucleoplasmins**, which are decameric, acidic, nuclear proteins whose presence is required for the proper *in vivo* assembly of **nucleosomes** (particles in which eukaryotic DNA is packaged) from their component DNA and histones (Section 34-1B).

In the following paragraphs we concentrate on the structure and function of the chaperonins, as these are the best characterized molecular chaperones. This discussion also constitutes our introduction to the dynamic functions of proteins, that is, to proteins as molecular machines.

a. The GroEL/ES System Forms a Large Cavity in Which Substrate Protein Folds

Group I chaperonins consist of two families of proteins that work in concert: (1) the **Hsp60** proteins (**GroEL** in *E. coli* and **Cpn60** in chloroplasts), which, as electron microscopic images first revealed, consist of 14 identical ~60-kD subunits arranged in two apposed rings of 7 subunits each (Fig. 9-18); and (2) the **Hsp10** proteins (**GroES** in *E. coli* and **Cpn10** in chloroplasts), which form single heptameric rings of identical ~10-kD proteins. These proteins, which are essential to the survival of *E. coli* under all conditions tested, facilitate the folding of improperly folded proteins to their native conformations (their discovery in *E. coli* as being necessary for the growth of certain bacteriophages is why they have the designation “Gro”).

The X-ray structure of GroEL (Fig. 9-19), determined by Arthur Horwich and Paul Sigler, shows, as expected, that GroEL's 14 identical 547-residue subunits associate to form a porous thick-walled hollow cylinder that consists of two 7-fold symmetric rings of subunits stacked back to back with 2-fold symmetry to yield a complex with D_7 symmetry (Section 8-5B). Each GroEL subunit consists of three domains: a large equatorial domain (residues 1–135 and 410–547) that forms the waist of the protein and holds

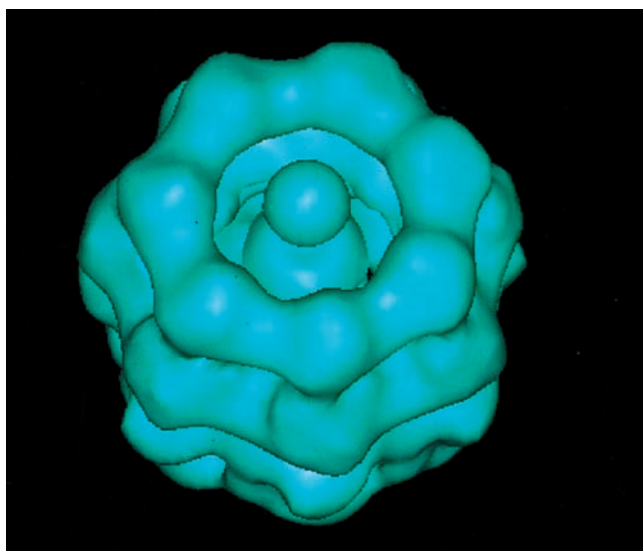
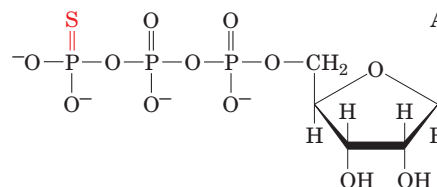


Figure 9-18 Electron micrograph–derived 3D image of the **Hsp60** chaperonin from the photosynthetic bacterium *Rhodobacter sphaeroides*. Hsp60 consists of 14 identical ~60-kD subunits arranged to form two apposed rings of 7 subunits, each surrounding a central cavity. The image of Hsp60, which is viewed with its 7-fold axis tipped toward the viewer, indicates that each subunit consists of two major domains, one in contact with the opposing heptameric ring, and the other at the end of the cylindrical protein molecule. The spherical density occupying the protein's central cavity is thought to represent a bound polypeptide. The cavity provides a protected microenvironment in which a polypeptide can fold. [Courtesy of Helen Saibil and Steve Wood, Birkbeck College, London, U.K.]

its subunits together through both intra- and inter-ring interactions, a loosely structured apical domain (residues 191–376) that forms the open ends of the GroEL cylinder, and a small intermediate domain (residues 136–190 and 377–409) that connects the equatorial and apical domains. The X-ray structure suggests that GroEL encloses an ~45-Å-diameter central channel that runs the length of the complex. We shall see below that this channel, in part, forms the chambers in which partially folded proteins fold to their native states. However, both electron microscopy–based images and neutron scattering studies indicate that the channel is obstructed in its equatorial region, so that proteins cannot pass between two GroEL rings. The obstruction is apparently caused by each subunit's N-terminal 5 residues and C-terminal 22 residues, which are not seen in the X-ray structure and hence are almost certainly disordered.

The X-ray structure of GroEL with **ATP γ S** bound to each subunit (ATP γ S is a poorly hydrolyzable analog of ATP in which S replaces one of the O atoms substituent to P γ)

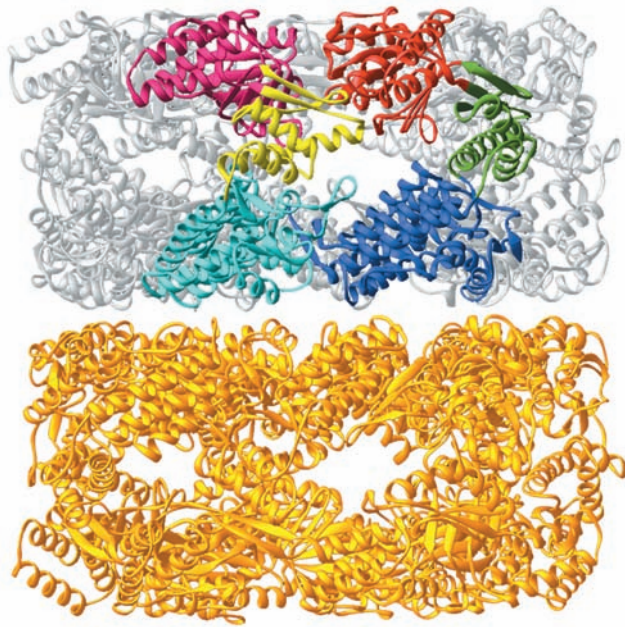


ATP γ S

indicates that ATP binds to a pocket in the equatorial domain that opens onto the central channel. The residues forming this pocket are highly conserved among chaperonins. The only significant differences between the structures of the GroEL–ATP γ S complex and that of GroEL alone are modest movements of the residues in the vicinity of the ATP pocket.

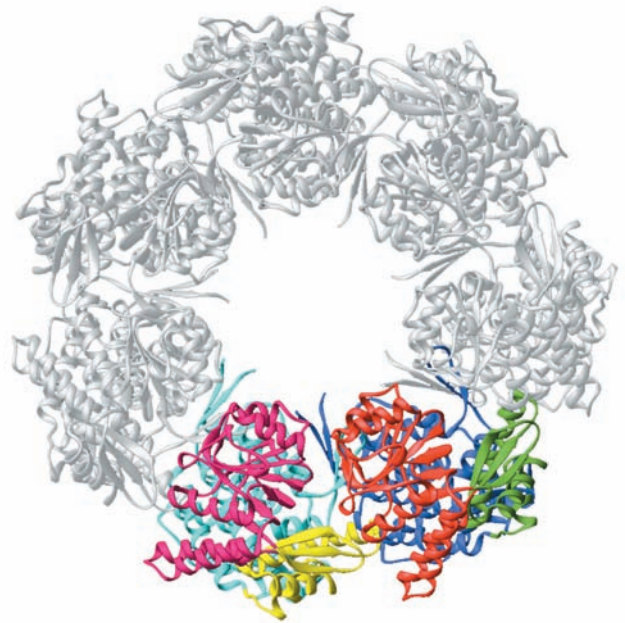
The X-ray structure of GroES (Fig. 9-20), determined by Lila Gierasch and Johann Deisenhofer, shows that this protein's 7 identical 97-residue subunits form a domelike structure with C_7 symmetry. Each GroES subunit consists of an irregular antiparallel β barrel from which two β hairpins project. One of these β hairpins (residues 47–55) extends from the top of the β barrel toward the protein's 7-fold axis, where it interacts with the other such β hairpins to form the roof of the dome. The second β hairpin (residues 16–33) extends from the opposite side of the β barrel outward from the bottom outer rim of the dome. This so-called mobile loop is seen in only one of GroES's 7 subunits; it is apparently disordered in the other subunits in agreement with the results of NMR studies of uncomplexed GroES in solution. The inner surface of the GroES dome is lined with hydrophilic residues.

Both electron microscopic and neutron scattering studies reveal that partially unfolded proteins bind in the mouth of the GroEL barrel in a manner reminiscent of a cork in a champagne bottle (Fig. 9-18). Mutations that impair polypeptide binding to GroEL all map to a poorly



(a)

Figure 9-19 X-ray structure of GroEL. (a) Side view perpendicular to the 7-fold axis in which the seven identical subunits of the lower ring are gold and those of the upper ring are silver, with the exception of the two subunits nearest the viewer, whose equatorial, intermediate, and apical domains are colored blue, green, and red on the right subunit and cyan, yellow, and magenta on the left subunit. The two rings of the complex



(b)

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

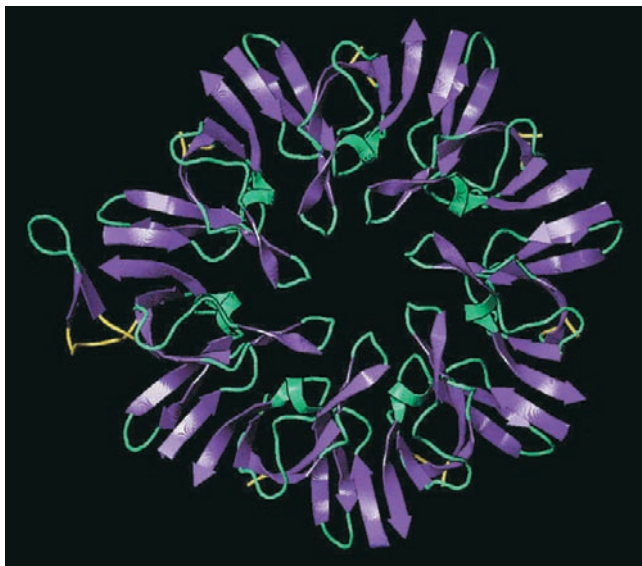


Figure 9-20 X-ray structure of GroES as viewed along its 7-fold axis. The mobile loop of only one of the protein's 7 identical subunits (*left*) is visible in the structure. The polypeptide segments that flank the mobile loop are yellow. [Courtesy of Johann Deisenhofer, University of Texas Southwest Medical Center, Dallas.]

resolved (and presumably flexible) segment at the top of the apical domain that, in the structure of GroEL alone, faces the central channel. In fact, changing any of nine highly conserved hydrophobic residues in this region to a hydrophilic residue abolishes polypeptide binding. It therefore seems likely that these residues provide the binding site(s) for non-native polypeptides. Interestingly, mutations of these same residues also abolish the binding of GroES.

The X-ray structure of the GroEL-(ADP)₇-GroES complex (Fig. 9-21), also determined by Horwich and Sigler, provides considerable insight into how this chaperonin carries out its function. In this complex, a GroES heptamer and the 7 ADPs are bound to the same GroEL ring (the so-called *cis* ring; the opposing GroEL ring is known as the *trans* ring) such that the GroES cap closes over the GroEL *cis* ring barrel like a lid on a pot, thereby forming a bullet-shaped complex with *C*₇ symmetry. The *trans* ring subunits have conformations that closely resemble those in the structure of GroEL alone. In contrast, the apical and intermediate domains of the *cis* ring have undergone large *en bloc* movements relative to their positions in GroEL alone (Fig. 9-22). This widens and elongates the *cis* cavity in a way that more than doubles its volume (from 85,000 to 175,000 Å³; Fig. 9-21c), thereby permitting it to enclose a partially folded substrate protein of up to ~70 kD. *These en bloc movements are concerted, that is, they occur simultaneously*

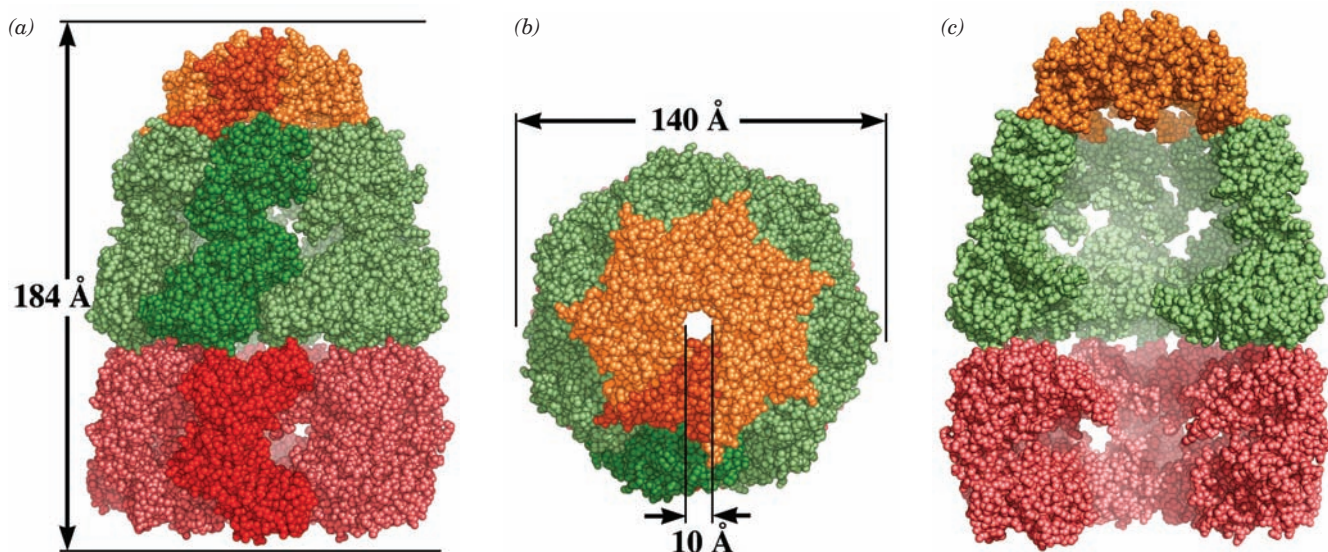


Figure 9-21 X-ray structure of the GroEL-(ADP)₇-GroES complex. (a) A space-filling drawing as viewed perpendicularly to the complex's 7-fold axis with the GroES ring orange, the cis ring of GroEL green, and the trans ring of GroEL red with one subunit in each ring shaded more brightly. The dimensions 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 7-fold axis. (c) As in Part a but with the two GroEL subunits closest to the viewer in both the cis and the 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.]

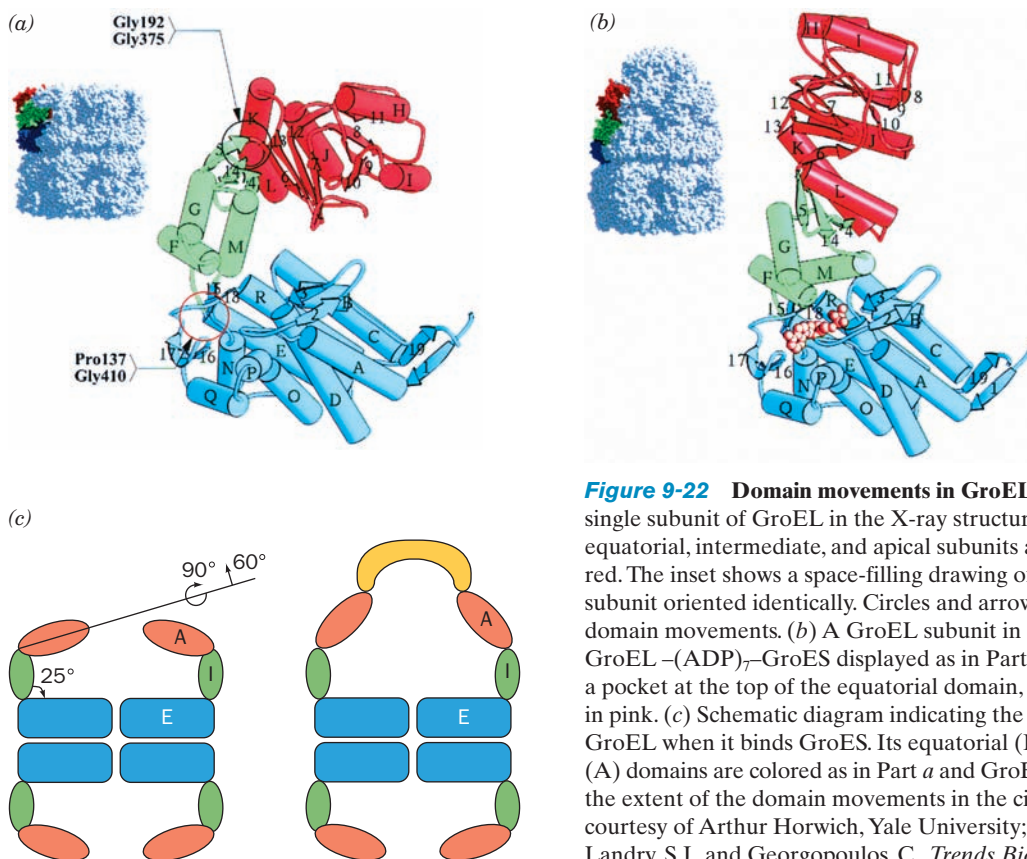


Figure 9-22 Domain movements in GroEL. (a) Ribbon diagram of a single subunit of GroEL in the X-ray structure of GroEL alone. Its equatorial, intermediate, and apical subunits are colored blue, green, and red. The inset shows a space-filling drawing of GroEL with the colored subunit oriented identically. Circles and arrows indicate the pivot points for domain movements. (b) A GroEL subunit in the X-ray structure of GroEL-(ADP)₇-GroES displayed as in Part a. The ADP, which is bound in a pocket at the top of the equatorial domain, is shown in space-filling form in pink. (c) Schematic diagram indicating the conformational changes in GroEL when it binds GroES. Its equatorial (E), intermediate (I), and apical (A) domains are colored as in Part a and GroES is yellow. The arrows indicate the extent of the domain movements in the cis ring of GroEL. [Parts a and b courtesy of Arthur Horwich, Yale University; Part c after Richardson, A., Landry, S.J., and Georgopoulos, C., *Trends Biochem. Sci.* **23**, 138 (1998).]

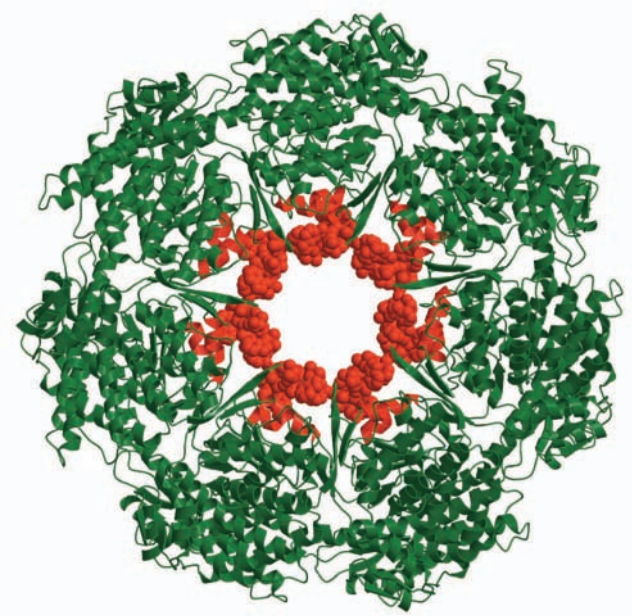


Figure 9-23 Apical domain of GroEL in complex with a tight-binding 12-residue polypeptide (SWMTPWGFLHP). To generate this drawing, the C_{α} atoms of the apical domain in the X-ray structure of the complex were superimposed on those of the apical domains in the X-ray structure of GroEL alone (Fig. 9-19). Each apical subunit is represented by a ribbon diagram in which the two helices involved in binding the polypeptide (helices H and I in Fig. 9-22a) are red and the remainder of the subunit is green. The polypeptides are shown in space-filling form in red. [Courtesy of Lingling Chen, Yale University. PDBid 1DKD.]

in all seven subunits of a GroEL ring, most probably because if one GroEL subunit did not undergo these conformational shifts, it would mechanically block its adjacent subunits from doing so.

In forming the GroEL-(ADP)₇-GroES complex, the ADP becomes completely enclosed by protein through the collapse of the intermediate domain onto the equatorial domain (Fig. 9-22b). This movement activates GroEL's ATPase function by shifting the side chain of its catalytically essential Asp 398, which extends from the L helix of the equatorial domain, into its catalytically active position near the ADP's β phosphate group. Electron microscopy studies at 10-Å resolution by Horwich and Helen Saibil reveal that similar movements occur when ATP binds to GroEL.

The hydrophobic groups lining the inner surface of the trans ring's apical domain, which extend from its H and I helices and an underlying loop (Fig. 9-22), presumably bind to the improperly exposed hydrophobic groups of substrate proteins. Indeed, an X-ray structure of the apical domain of GroEL in complex with a 12-residue peptide that binds strongly to GroEL reveals that this peptide binds to these exposed hydrophobic groups (Fig. 9-23). However, in the cis ring of the GroEL-(ADP)₇-GroES complex, these hydrophobic groups participate either in binding GroES via its flexible loops or in stabilizing the newly formed interface between the rotated and elevated apical domains. Consequently, *these hydrophobic groups are no longer exposed on the inner surface of the cis cavity (Fig. 9-24), thereby depriving a substrate protein of its binding sites.*

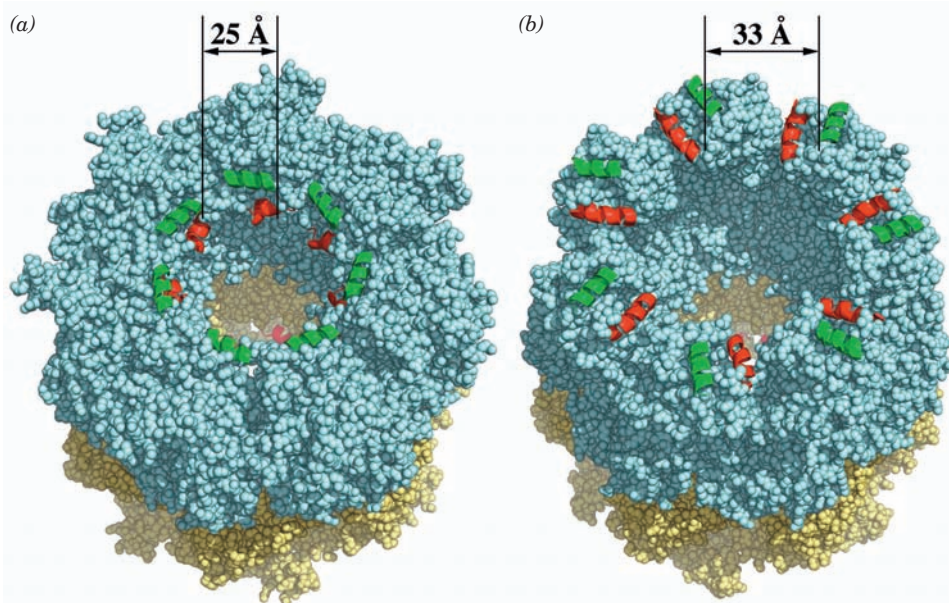


Figure 9-24 Movements of the polypeptide-binding helices of GroEL. (a) A space-filling drawing of GroEL in the structure of GroEL alone and (b) in the structure of GroEL-(ADP)₇-GroES. The GroEL cis and trans rings are pale cyan and pale yellow and the cis ring's H and I helices (Figs. 9-22a,b), which form most of the hydrophobic binding sites for improperly folded proteins, are drawn in cartoon form and colored green and red, respectively. On

the addition of GroES and ATP to GroEL, neighboring binding sites separate by 8 Å and non-neighboring sites separate by up to 20 Å. A substrate protein initially bound to two of these sites will likely be forcibly stretched and hence partially unfolded before being released as the binding sites become occluded. [After drawings by George Lorimer, University of Maryland; and Walter Englander, University of Pennsylvania. PDBids 1OEL and 1AON.]

b. GroEL/ES Undergoes Coordinated Conformational Changes That Are Paced by ATP Binding and Hydrolysis

The binding of ATP and GroES to the cis ring of GroEL strongly inhibits their binding to the trans ring. The X-ray structure of the GroEL-(ADP)₇-GroES complex suggests that this occurs through concerted small conformational shifts in the GroEL equatorial domains that apparently prevent the trans ring from assuming the conformation of the cis ring. However, once the cis ring has hydrolyzed its bound ATP (which it is committed to do once its nucleotide binding sites close off and its ATPase active sites form), the trans ring can bind ATP and the resulting conformational shifts release GroES from the cis ring. This explains why a mutant form of GroEL that has only a single ring (and hence is known as SR1) can bind substrate protein and GroES but does not release them after it hydrolyzes its bound ATP. *The proper functioning of GroEL requires two rings, even though their central cavities are unconnected.*

A mutant form of GroEL, D398A (in which Asp 398 has been changed to Ala), binds but cannot hydrolyze ATP. In the presence of ATP, D398A GroEL binds GroES together with substrate protein. However, it does not release GroES or the protein when the trans ring is exposed to ATP, as is the case when the cis ring can hydrolyze ATP. Evidently, *ATP's γ -phosphate group provides strong contacts that stabilize the GroEL-GroES interaction. When the ATP in the cis ring is hydrolyzed, the resulting phosphate group is released and these interactions are lost.*

c. ATP Hydrolysis in the Cis Ring Must Occur before Substrate Protein and GroES Can Bind to the Trans Ring

The foregoing indicates that *events in the cis and trans rings of the GroEL-GroES complex are coordinated through concerted conformational changes in one ring that influence the conformation of the opposing ring.* What is the sequence of events in the trans ring relative to those in the cis ring, that is, at what stage of the folding cycle in the cis ring do substrate protein and GroES bind to the trans ring? Horwich answered this question using fluorescence labeling techniques. D398A GroEL that had been mixed with ADP and GroES so as to form a stable complex [D398A GroEL-(ADP)₇-GroES] was then mixed with a substrate protein to which a fluorescent group had been covalently linked. When this mixture was subjected to gel filtration chromatography (Section 6-3B), the label migrated with the GroEL, thereby indicating that the substrate protein had bound to the complex's trans ring. However, when the initial complex was instead made with ATP (recall that D398A GroEL cannot hydrolyze ATP), the substrate protein did not associate with the GroEL. In similar experiments, fluorescently labeled GroES associated with preformed D398A GroEL-(ADP)₇-GroES in the presence of ATP but not with preformed D398A GroEL-(ATP)₇-GroES. Evidently, *the cis ring of the GroEL-GroES complex must hydrolyze its bound ATP before the trans ring can bind either substrate protein or GroES + ATP.*

d. The GroEL/ES System Functions as a Two-Stroke Engine

Taken together, all of the preceding observations indicate how the GroEL/ES system functions (Fig. 9-25):

1. A GroEL ring that is binding 7 ATP and an improperly folded substrate protein via the hydrophobic patches on its apical domains (Fig. 9-25, *upper left*) binds GroES. This induces a conformational change in the now cis GroEL ring, thereby releasing the substrate protein into the resulting enlarged and closed cavity, where the substrate protein commences folding. The cavity, which is now lined only with hydrophilic groups, provides the substrate protein with an isolated microenvironment that prevents it from nonspecifically aggregating with other unfolded proteins (a so-called **Anfinsen cage**).

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_i (where P_i is the symbol for inorganic phosphate) and the P_i is released. The absence of ATP's γ -phosphate group weakens the interactions that bind GroES to GroEL.

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

4. The binding of substrate protein and ATP to the trans ring induces the cis ring to release its bound GroES, 7 ADP, and the now possibly natively folded substrate protein. This leaves only ATP and substrate protein bound to the previous trans ring of GroEL, which becomes the cis ring on binding GroES when the complex again cycles through Step 1.

Substrate protein that has not achieved its native state or is not committed to do so is readily recaptured by GroEL. Substrate protein that has achieved its native fold lacks exposed hydrophobic groups and hence cannot bind to GroEL. It is the irreversible hydrolysis of ATP that drives the folding cycle in only the direction indicated in Fig. 9-25.

e. GroEL Unfolds Its Substrate Proteins before Facilitating Their Refolding

How does the foregoing cycle promote the proper folding of an improperly folded protein? Two models, not mutually exclusive, have received the most consideration:

1. The Anfinsen cage model, in which the GroEL/ES complex provides the substrate protein with a protected microenvironment in which it can fold to its native conformation without interference by nonspecific aggregation with other misfolded proteins. Moreover, the confinement of the substrate protein to the relatively small volume of the cis ring cavity eliminates nonproductive folding pathways involving extended conformations, and the hydrophilic character of the cavity walls promotes productive folding pathways by favoring the burial of hydrophobic residues. In terms of landscape theory (Section 9-1Ch), this would smooth the walls of the folding funnel (Fig. 9-13d) and thus facilitate the folding of the substrate protein toward its global free energy minimum, that is, its native state.

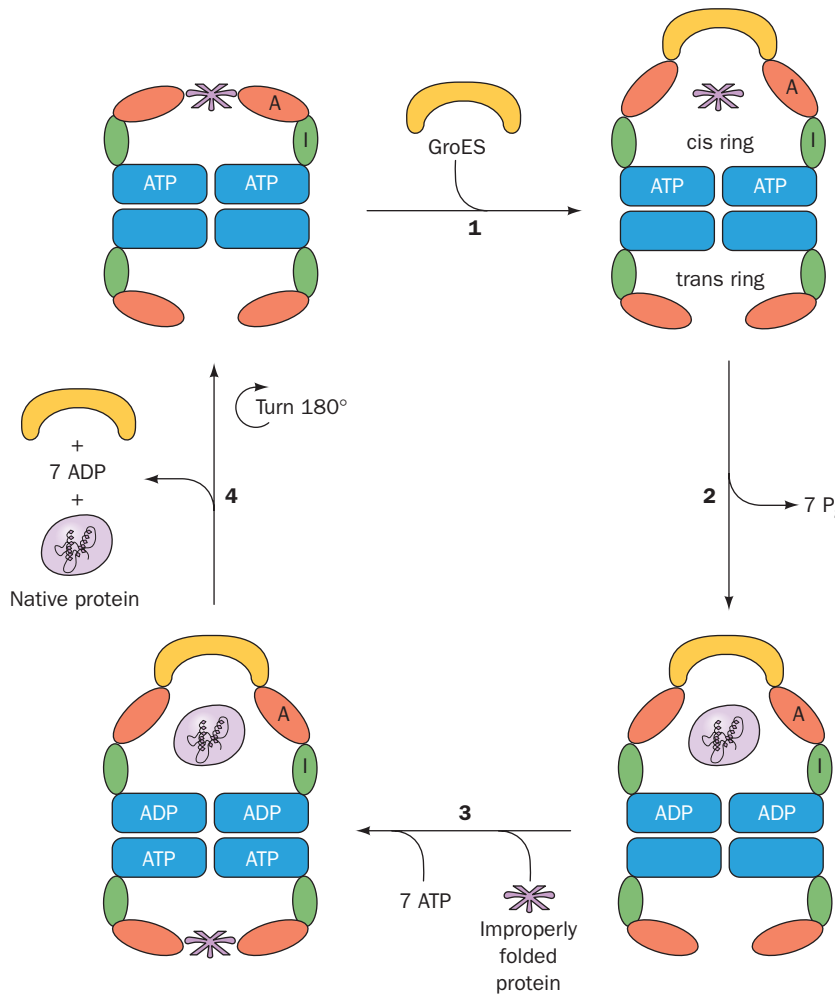


Figure 9-25 Reaction cycle of the GroEL/ES chaperonin system in protein folding. See the text for an explanation.

2. The iterative annealing model, in which the ATP-driven unfolding of a misfolded and conformationally trapped substrate protein followed by its release permits it to resume folding to its native state. This would occur through the binding of a misfolded protein to the hydrophobic patches on two or more of the GroEL cis rings's seven apical domains, followed by the stretching and ultimate release of the protein as GroEL changes conformation on binding ATP and GroES [recall that these patches are further apart in the GroEL-(ADP)₇-GroES complex than they are in GroEL alone; Fig. 9-24]. In terms of landscape theory, this stretching would expel the substrate protein (raise its free energy) from a local energy minimum in which it had become trapped and thereby permit it to continue, but not necessarily complete, its conformational journey down the folding funnel toward its native state.

Fluorescence resonance energy transfer (FRET; Section 9-1Cd) measurements by Hays Rye indicate that the forced unfolding of substrate protein by GroEL enhances its rate of folding. **Ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCO;** Sections 24-3Ac and 24-3C) from *Rhodospirillum rubrum*,

which requires GroEL/ES to fold to its native state, was covalently labeled on its N- and C-terminal domains by acceptor and donor fluorescent probes, respectively (which does not affect RuBisCO's stability or its GroEL-mediated folding rate). FRET measurements indicated that on binding to the trans ring of a GroEL-(ADP)₇-GroES complex, the fluorescently labeled RuBisCO's end-to-end distance increases slightly. However, on the subsequent addition of ATP, this distance greatly increases within 0.2 s and then decreases to less than its original value over a ~5 s period. This indicates that the initial binding of ATP to GroEL significantly unfolds its bound substrate protein, which then folds to a more compact state within the now cis cavity of GroEL/ES. Further measurements indicate that the fraction of unfolded RuBisCO that folds to its native state increases with the extent of its unfolding previous to being released into the cis cavity.

As diagrammed in Fig. 9-25, the GroEL/ES system releases its substrate protein after each reaction cycle, whether or not the protein is properly folded. In contrast, SR1, the single-ring mutant form of GroEL, in its complex with GroES, cannot release its bound substrate protein.

Nevertheless, the trapped substrate protein refolds nearly quantitatively to its native state over a period of several minutes, about the same rate as it does so in the cycling system. Evidently, the efficiency with which a substrate protein folds to its native state varies with the length of time that it spends in the cis cavity (Anfinsen cage). Then why hasn't a GroEL/ES system evolved that allows an unfolded protein to complete its folding before it is released? The answer may be that the release of the substrate protein from the cis cavity with each turn of the GroEL/ES cycle is a protective mechanism that prevents irretrievably damaged proteins from permanently clogging GroEL. In a cycling system, a substrate protein spends only a fraction of the time in a cis cavity. Thus, since forced unfolding increases folding efficiency, it is a major contributor to the GroEL/ES system's multilayered protein folding mechanism. Moreover, forced unfolding explains how the GroEL/ES system is able to facilitate the folding of several proteins that are too large to completely fit inside the GroEL cavity.

A variety of experiments indicate that substrate proteins bound to the open ring of GroEL alone are largely unstructured. For example, NMR measurements indicate that the 21-kD enzyme **dihydrofolate reductase (DHFR; Section 28-3Bd)** bound to GroEL or SR1 has no stable structure, and hydrogen exchange measurements (Section 9-1Cc) on several substrate proteins bound to GroEL indicate that they exhibit little or no secondary structure. Moreover, FRET measurements on the 41-kD **maltose binding protein** bound to the trans ring of the GroEL–GroES complex reveal that it undergoes a rapid conformational expansion on ATP addition (as does RuBisCO), and NMR measurements indicate that DHFR inside the SR1–GroES cavity follows the same folding trajectory as does DHFR free in solution. Thus GroEL/ES-mediated folding appears to be an all-or-none process rather than an iterative one in which the substrate protein progressively acquires more native-like structure with each round of folding. This suggests that each time a substrate protein binds to the trans ring of GroEL–GroES, it is raised to the top of its folding funnel in an ATP-driven process from which it commences folding via a different trajectory.

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. This may seem like a profligate use of ATP, but it is only a fraction of the 1200 ATPs expended in ribosomally synthesizing a 300-residue protein from its component amino acids (4 ATPs per residue; Sections 32-2C and 32-3D), not to mention the far greater number of ATPs required to synthesize these amino acids (Section 26-5).

f. GroEL/ES Is Required for the Folding of ~85 *E. coli* Proteins *in vivo*

The GroEL/ES system only interacts *in vivo* with a subset of *E. coli* proteins. Ulrich Hartl identified these proteins by modifying GroES to have a C-terminal His₆ segment (a His-Tag) and isolating the resulting GroEL–GroES–

substrate protein complexes from *E. coli* lysates by metal chelation affinity chromatography (Section 6-3Dg). These complexes were separated by SDS–PAGE (Section 6-4C) and the substrate proteins identified by mass spectrometry (Section 7-1I).

Approximately 250 of *E. coli*'s ~2400 cytosolic proteins were found to be associated with GroEL/ES. Of these, ~165 proteins either show little tendency to aggregate during folding or can utilize other chaperone proteins such as trigger factor or DnaK/J to fold to their native states. However, the remaining ~85 proteins have an absolute dependence on the GroEL/ES system for folding, that is, they invariably aggregate in the absence of GroEL/ES. Thirteen of these proteins are indispensable for *E. coli* viability, thereby explaining why GroEL/ES is also essential for *E. coli* viability. About 75% to 80% of the GroEL/ES binding sites are occupied by the ~85 GroEL/ES-dependent proteins, even though they have only low to intermediate abundance in the *E. coli* cytosol.

What are the characteristics of proteins that are obligate substrates of GroEL/ES? Analysis, using the SCOP database (Section 8-3Cd), of those proteins of known structure or with homologs of known structure revealed that many of them contain α/β domains (Section 8-3Bh). In particular, ~35% by mass of all GroEL/ES substrate proteins contain α/β barrels (also called TIM barrels; Section 8-3Bh), even though they comprise only ~6% of the cytosol's total protein mass. These proteins, whose molecular masses range from 23 to 54 kD, are stabilized by numerous long range (in sequence) interactions and hence would be expected to have particularly rugged folding funnels with many local free energy minima that could trap the unaided protein.

What are the substrate protein sequence motifs that bind to GroEL? During the GroEL/ES cycle, the GroES mobile loops (sequence GGIVLTGSA) displace these motifs (Section 9-2Ca), thus suggesting that they have similar sequences. Moreover, to be stretched by GroEL, a substrate protein must have at least two such motifs separated by at least 10 residues. By searching GroEL's ~250 substrate proteins for motifs with these characteristics, George Lorimer and Devarajan Thirumalai found that they have the consensus sequence P_HHH_P_H, where P, H, and _ respectively represent polar, hydrophobic, and any residues and where the core sequence is P_HHH. This is corroborated by the observation that in natively folded substrate proteins of known structure, nearly all of these sequence motifs are buried (<50% of their surface area is solvent-accessible), although since they occur in helices, sheets, and loops, they apparently have little other structural preferences.

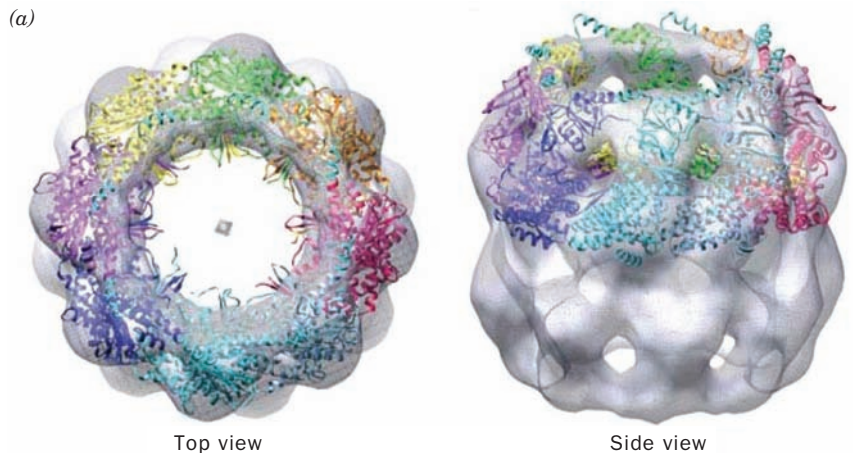
The obligate GroEL/ES substrate proteins belong to fold classes that tend to have a greater number of superfamilies than do other *E. coli* proteins. This suggests that GroEL/ES may have facilitated the evolutionary diversification of certain protein folds, perhaps by “buffering” mutations that would otherwise cause severe aggregation. Indeed, the GroEL/ES system likely played an essential role in the evolution of the α/β barrel into the most versatile structural platform for enzymatic functions (Section 8-3Bh).

g. Group II Chaperonins Have Built-In Lids

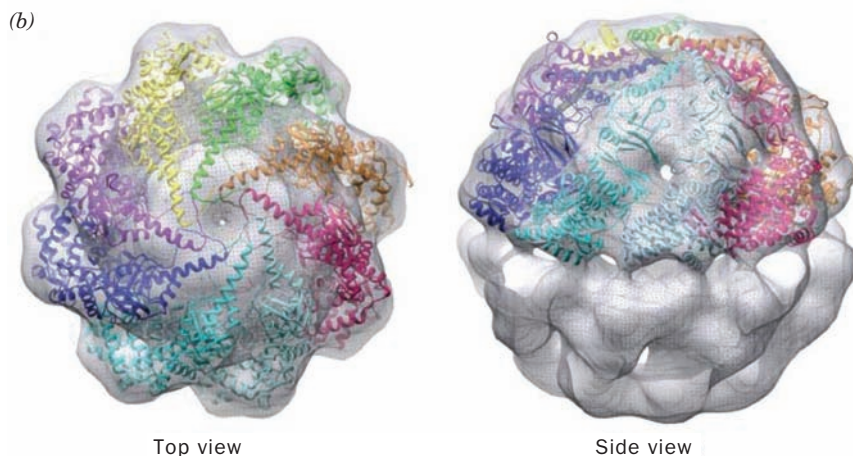
Group II chaperonins structurally and functionally resemble Group I chaperonins but consist of back-to-back rings of 8 or 9 subunits and have no corresponding cochaperones such as GroES. Archaeal Group II chaperonins, which are called **thermosomes**, consist of 1 to 3 different types of subunits. Eukaryotic Group II chaperonins, which are named **TRiC** (for *TCP-1 ring complex*; TCP for *T-complex polypeptide*) or alternatively **CCT** (for chaperonin-containing TCP-1), have dual octameric rings, each consisting of eight genetically distinct but homologous subunits arranged in a specific order. Like GroEL/ES, each of the TRiC subunits couple the hydrolysis of ATP to the folding of substrate proteins. Around 10% of eukaryotic cytosolic proteins transiently interact with TRiC, many of which have an absolute requirement for TRiC-aided folding. These include a variety of essential structural and regulatory proteins including the muscle proteins **actin** and **myosin** (Section 35-3A), the major microtubule components **tubulins α** and **β** (Section 35-3G), and proteins that participate in signal transduction (Chapter 19) and cell cycle

regulation (Section 34-4D). The only characteristic that these proteins appear to have in common is that they form homo- or hetero-oligomeric complexes.

Cryo-electron microscopic (cryo-EM) studies of bovine testes TRiC by Wah Chiu and Judith Frydman (Fig. 9-26) revealed that, unlike GroEL, its overall shape is closer to spherical than cylindrical. [In cryo-EM, the sample is cooled to near liquid N₂ temperatures (-196°C) so rapidly (in a few milliseconds) that the water in the sample does not have time to crystallize but, rather, assumes a vitreous (glasslike) state. Consequently, the sample remains hydrated and hence retains its native shape to a greater extent than in conventional electron microscopy (in which the sample is vacuum dried).] Like GroEL, TRiC's subunits each consist of equatorial, intermediate, and apical domains. However, the tip of each TRiC apical domain has a helical extension that GroEL lacks. TRiC has two conformational states: an open state (Fig. 9-26a) in which each helical extension is oriented more or less tangentially to the inner portion of the ring, and a closed state (Fig. 9-26b) in which each helical extension has swung to a more radial



Open TRiC model versus EM open state



Closed TRiC model versus EM closed state

Figure 9-26 Structure of bovine testes

TRiC. (a) Its open state and (b) its closed state. Surface diagrams of the cryo-EM-based structures at ~ 16 Å resolution (transparent gray) are shown as viewed along the protein's 8-fold axis (left) and as viewed from the side with the 8-fold axis tipped toward the viewer (right). The X-ray structure of a homologous archaeal chaperonin was modeled into the upper ring of the cryo-EM-based image with each subunit represented by a differently colored ribbon diagram. [Courtesy of Judith Frydman, Stanford University.]

orientation so as to form an iris-like lid that closes off the TRiC cavity in a manner similar to the way that GroES closes off the GroEL cavity. A lidless mutant of TRiC still hydrolyzes ATP and binds unfolded actin with wild-type affinity but is unable to induce its folding. This suggests that the TRiC lid has a GroES-like function in coupling ATP hydrolysis to the productive folding of substrate protein. Interestingly, the putative substrate binding sites in TRiC are not all hydrophobic as some of its subunits are lined with polar residues. Perhaps each different substrate protein interacts with a specific combination of binding sites on the apical domains in the open state of TRiC.

h. The Concept of Self-Assembly Must Take Accessory Proteins into Account

Many proteins can fold/assemble to their native conformations in the absence of accessory proteins, albeit often with low efficiency. Moreover, accessory proteins are not components of the native proteins whose folding/assembly they facilitate. Hence, accessory proteins must mediate the proper folding/assembly of a polypeptide to a conformation/complex governed solely by the polypeptide's amino acid sequence. Nevertheless, the concept that proteins are self-assembling entities must be modified to incorporate the effects of accessory proteins.

3 PROTEIN STRUCTURE PREDICTION AND DESIGN

Since the primary structure of a protein specifies its three-dimensional structure, it should be possible, at least in principle, to predict the native structure of a protein from a knowledge of only its amino acid sequence. This might be done using theoretical methods based on physicochemical principles, or by empirical methods in which predictive schemes are distilled from the analyses of known protein structures. Theoretical methods, which usually attempt to determine the minimum energy conformation of a protein, are mathematically quite sophisticated and require extensive computations. The enormous difficulty in making such calculations sufficiently accurate and yet computationally tractable has, so far, limited their success. Nevertheless, an understanding of how and why proteins fold to their native structures must ultimately be based on such theoretical methods. In this section we outline various methods that have been used to predict the secondary and tertiary structures of proteins and end with a discussion of a related technique, that of designing proteins that will have a particular structure.

A. Secondary Structure Prediction

The most reliable way to determine the secondary structure taken up by a polypeptide is to map its amino acid sequence onto that of a homolog of known structure. If, however, no such structure is available, the above-mentioned predictive methods must be employed. Here we discuss the use of empirical methods for secondary structure prediction. The theoretical methods discussed in the following

section to predict a polypeptide's tertiary structure will, of necessity, also predict its secondary structure.

a. The Chou–Fasman Method

Empirical methods have had reasonable success in secondary structure prediction. Clearly, certain amino acid sequences limit the conformations available to a polypeptide chain in an easily understood manner. For example, a Pro residue cannot fit into the interior portions of a regular α helix or β sheet because its pyrrolidine ring would fill the space normally occupied by part of an abutting segment of chain and because it lacks the backbone N—H group with which to contribute a hydrogen bond. Likewise, steric interactions between several consecutive amino acid residues with side chains branched at C_β (e.g., Ile and Thr) will destabilize an α helix. Furthermore, there are more subtle effects that may not be apparent without a detailed analysis of known protein structures. Here we shall discuss simple empirical methods for predicting the positions of α helices, β sheets, and reverse turns in proteins of known sequence.

The empirical structure prediction scheme developed by Peter Chou and Gerald Fasman can be readily applied by hand and is reasonably reliable. Its use requires two definitions. The frequency, f_α , with which a given residue occurs in an α helix in a set of protein structures is defined as

$$f_\alpha = \frac{n_\alpha}{n} \quad [9.3]$$

Table 9-1 Propensities and Classifications of Amino Acid Residues for α Helical and β Sheet Conformations

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

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

where n_α is the number of amino acid residues of the given type that occur in α helices and n is the total number of residues of this type in the set. The propensity of a particular amino acid residue to occur in an α helix is defined as

$$P_\alpha = \frac{f_\alpha}{\langle f_\alpha \rangle} \quad [9.4]$$

where $\langle f_\alpha \rangle$ is the average value of f_α for all 20 residues. Accordingly, a value of $P_\alpha > 1$ indicates that a residue occurs with greater than average frequency in an α helix. The propensity, P_β , of a residue to occur in a β sheet is similarly defined.

Table 9-1 contains a list of α and β propensities based on the analysis of 29 X-ray structures. In accordance with its value of a given propensity, a residue is classified as a strong former (H), former (h), weak former (I), indifferent former (i), breaker (b), or strong breaker (B) of that secondary structure. Using these data, Chou and Fasman formulated the following empirical rules (the **Chou-Fasman method**) to predict the secondary structures of proteins:

1. A cluster of four helix-forming residues (H_α or h_α , with I_α counting as one-half h_α) out of six contiguous residues will nucleate a helix. The helix segment propagates in both directions until the average value of P_α for a tetrapeptide segment falls below 1.00. A Pro residue, however, can occur only at the N-terminus of an α helix.

2. A cluster of three β sheet formers (H_β or h_β) out of five contiguous residues nucleates a sheet. The sheet is propagated in both directions until the average value of P_β for a tetrapeptide segment falls below 1.00.

3. For regions containing both α - and β -forming sequences, the overlapping region is predicted to be helical if its average value of P_α is greater than its average value of P_β ; otherwise a sheet conformation is assumed.

These easily applied empirical rules predict the α helix and β sheet strand positions in a protein with an average reliability of $\sim 50\%$ and, in the most favorable cases, $\sim 80\%$ (Fig. 9-27; note, however, that because proteins consist, on average, of $\sim 31\%$ α helix and $\sim 28\%$ β sheet, random predictions of these secondary structures would average $\sim 30\%$ correct).

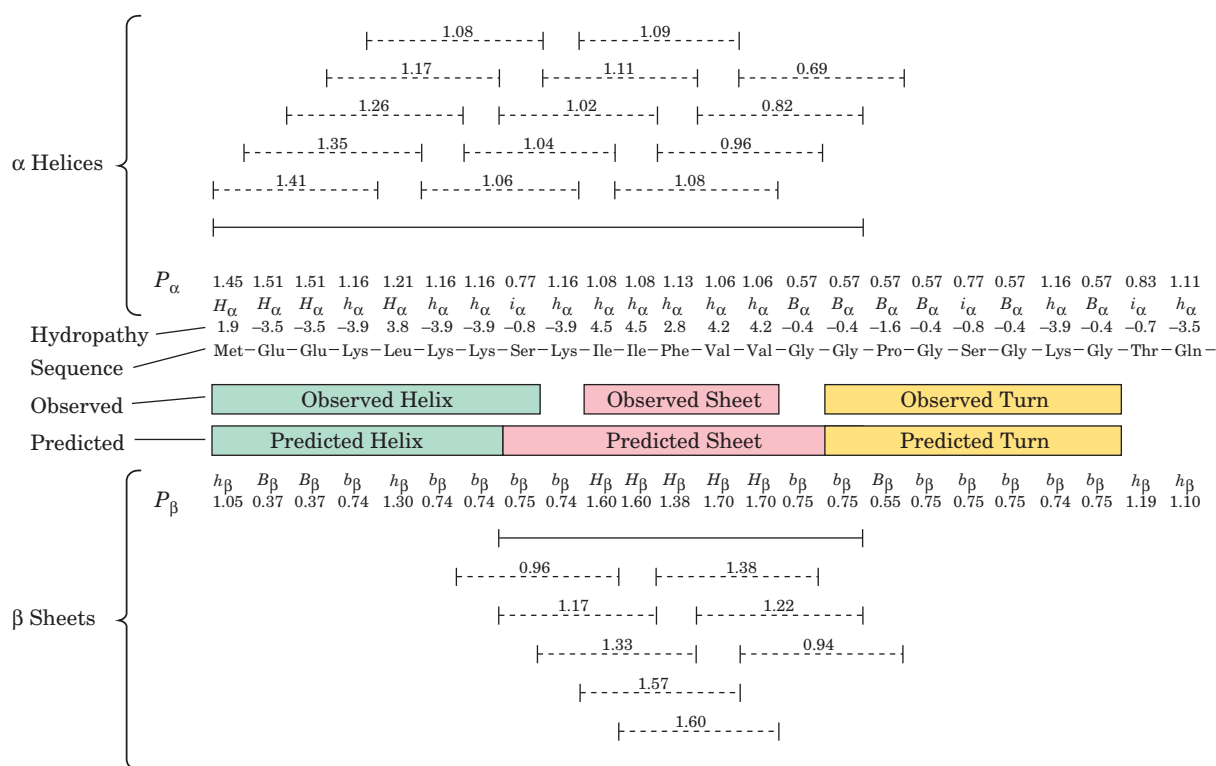


Figure 9-27 Secondary structure prediction. The prediction of α helices and β sheets was made by the Chou-Fasman method and the prediction of reverse turns by the method of Rose for the N-terminal 24 residues of adenylate kinase. The helix and sheet propensities and classifications are taken from Table 9-1. The solid lines indicate all hexapeptide sequences that can nucleate an α helix (top) and all pentapeptide sequences that can nucleate a β sheet (bottom), as is explained in the text. The average helix and sheet propensities for each tetrapeptide segment in the helix

and sheet regions are given above the corresponding dashed lines. Twelve of the 15 residues are observed to have their predicted secondary structures (middle), so that the prediction accuracy, in this case, is 80%. Reverse turns are predicted to occur in sequences in which the hydrophathy (Table 8-6) is a minimum and which do not occur in regions predicted to be helical. The region that matches this criterion is observed to have a reverse turn. [After Schultz, G.E. and Schirmer, R.H., *Principles of Protein Structure*, p. 121, Springer-Verlag (1979).]

b. Reverse Turns Are Characterized by a Minimum in Hydrophobicity Along a Polypeptide Chain

The positions of reverse turns can also be predicted by the Chou–Fasman method. However, since a reverse turn usually consists of four consecutive residues, each with a different conformation (Section 8-1D), their prediction algorithm is necessarily more cumbersome than those for sheets and helices.

Rose has proposed a simpler empirical method for predicting the positions of reverse turns. Reverse turns nearly always occur on the surface of a protein and, in part, define that surface. Since the core of a protein consists of hydrophobic groups and its surface is relatively hydrophilic, reverse turns occur at positions along a polypeptide chain where the hydrophathy (Table 8-6) is a minimum. Using these criteria for partitioning a polypeptide chain, we can deduce the positions of most reverse turns by inspection (Fig. 9-27). Since this method often predicts reverse turns to occur in helical regions (helices are all turns), it should be applied only to regions that are not predicted to be helical.

c. Physical Basis of α Helix Propensity

Why do amino acid residues have such different propensities for forming α helices? This question has been answered, in part, by Matthews through the structural and thermodynamic analysis of T4 lysozyme (Section 9-1Bd) in which Ser 44, a solvent-exposed residue in the middle of a 12-residue (3.3-turn) α helix, was mutagenically replaced, in turn, by all 19 other amino acids. The X-ray structures of 13 of these variant proteins revealed that, with the exception of Pro, the substitutions caused no significant distortion to the α helix backbone and, hence, that differences in α helix propensities are unlikely to arise from strain. However, for 17 of the amino acids (all but Pro, Gly, and Ala), the stability of the α helix increases with the amount of side chain hydrophobic surface that is buried (brought out of contact with the solvent) when residue 44 is transferred from a fully extended state to an α helix. The low α helix propensity of Pro is due to the strain generated by its presence in an α helix, and that of Gly arises from the entropic cost associated with restricting this most conformationally flexible of residues to an α helical conformation (compare Figs. 8-7 and 8-9) and its lack of hydrophobic stabilization. The high α helix propensity of Ala, however, is caused by its lack of a γ substituent (possessed by all residues but Gly and Ala) and hence the absence of the entropic cost associated with conformationally restricting such a group within an α helix together with its small amount of hydrophobic stabilization.

d. Computer-Based Secondary Structure Prediction Algorithms

A number of sophisticated computer-based secondary structure prediction algorithms have been developed. Most of them, like the Chou–Fasman method, employ sets of parameters whose values are determined by the analysis of (learning from) a set of nonhomologous proteins with known structures, in some cases coupled with energy mini-

mization techniques. These algorithms are typically $\sim 60\%$ accurate in predicting which of three conformational states, helix, sheet, or coil, a given residue in a protein adopts. However, a significant increase in accuracy has been gained (to over 80%) by employing evolutionary information through the use of multiple sequence alignments. This is because knowledge of the distribution of residue identities at and around each position in a series of homologous and presumably structurally similar proteins provides a better indication of the protein's structural tendencies than does a single sequence.

Several secondary structure prediction algorithms are freely available over the Web. Among them is **Jpred3** (<http://www.compbio.dundee.ac.uk/www-jpred/>), which classifies residue conformations as being either helical (H), extended/ β sheet (E), or coil (–) with 81.5% reliability. It requires as input either the sequence of a single polypeptide or a multiple sequence alignment. However, if Jpred3 is supplied with only a single sequence, it will first use PSI-BLAST (Section 7-4Bi) to construct a multiple sequence alignment.

Although we have seen that secondary structure is mainly dictated by local sequences, we have also seen that tertiary structure can influence secondary structure (Section 9-1Be). The inability of sophisticated secondary structure prediction schemes to surpass $\sim 80\%$ reliability is therefore partially explained by their failure to take tertiary interactions into account.

B. Tertiary Structure Prediction

The sequence databases (Section 7-4A) contain the sequences of ~ 7 million polypeptides, and the rapid rate at which entire genomes are being sequenced (Section 7-2C) promises that many more such sequences will soon be known. Yet, only a small fraction of the $\sim 70,000$ protein structures in the PDB (Section 8-3B) are unique because many of them are of the same protein binding different small molecules, mutant forms of the same protein, or closely related proteins. Moreover, around 40% of the **open reading frames (ORFs; nucleic acid sequences that appear to encode proteins)** in known genome sequences specify proteins whose function is unknown. Consequently, formulating a method to reliably predict the native structure of a polypeptide from only its sequence is a major goal of biochemistry. In the following paragraphs we discuss the progress that has been made in achieving this goal.

There are currently several major approaches to tertiary structure prediction. The simplest and most reliable approach, **comparative or homology modeling**, aligns the sequence of interest with the sequences of one or more homologous proteins of known structure, compensating for amino acid substitutions as well as insertions and deletions (indels) through modeling and energy minimization calculations. For proteins with as little as 30% sequence identity, this method can yield a root-mean-square deviation (rmsd) between the predicted and observed positions of corresponding C_α atoms of the “unknown” protein (once its structure has been determined) of as little as ~ 2.0 Å. However, the accuracy of this method decreases precipitously

(the rmsd's rapidly increase) as the degree of sequence identity drops below 30%. Conversely, for polypeptides that are >60% identical, a homology model may have rmsd's of ~ 1 Å (the accuracy of the atomic positions in an ~ 2.5 -Å-resolution X-ray structure).

There are numerous instances of proteins that are structurally similar even though their sequences have diverged to such an extent that they have no apparent similarity. **Fold recognition** or **threading** is a computational technique that attempts to determine the unknown fold of a protein by ascertaining whether its sequence is compatible with any of the members of a library of known protein structures. It does so by placing the “unknown” protein's residues along the backbone of a known protein structure, determining the stability of the side chains of the unknown protein in that arrangement, and then sliding (threading) the sequence of the unknown protein along that of the known protein by one residue and repeating the calculation, etc., while allowing for the possibility of indels. If the “correct” fold can be found (and there is no guarantee that the fold of the unknown protein will resemble that of any member of the library), the resulting model can be improved via homology modeling. This method has yielded encouraging results, although it cannot yet be considered to be reliable. Of course, as sequence alignment algorithms (Section 7-4B) improve in their ability to recognize distant homologs, sequences that previously would have been candidates for fold recognition can instead be directly treated by comparative modeling.

Since the native structure of a protein depends only on its amino acid sequence, it should be possible, in principle, to predict the structure of a protein based only on its physicochemical properties (e.g., solvent interactions, atomic volume, charge, hydrogen bonding properties, van der Waals interactions, and bond torsion angle potentials for all of its atoms). A major problem faced by such *de novo* (Latin: from the beginning; synonymously *ab initio*) **methods** is that polypeptide chains have astronomical numbers of non-native low-energy conformations, so that it requires extensive and highly detailed calculations to determine a polypeptide's lowest energy conformation.

To assess the effectiveness of the numerous *de novo* algorithms that have been formulated, as well as other structure prediction schemes, a **Critical Assessment of Structure Prediction (CASP)** has been held every 2 years starting in 1994. CASP participants are provided with the sequences of proteins whose structures will soon be determined by X-ray crystallography or NMR spectroscopy and submit their predicted structures of these proteins for comparison with the subsequently experimentally determined structures. Over the years, *de novo* methods have steadily improved from being little better than random guesses to predicting the folding topologies of <200-residue proteins with a success rate of $\sim 20\%$ and occasionally with near-atomic accuracy. **Rosetta**, the most consistently successful *de novo* algorithm in the past few CASPs, which was formulated by David Baker, is enormously computer-intensive. To satisfy its computational needs, Baker has organized a distributed computer network known as Rosetta@home that uses the

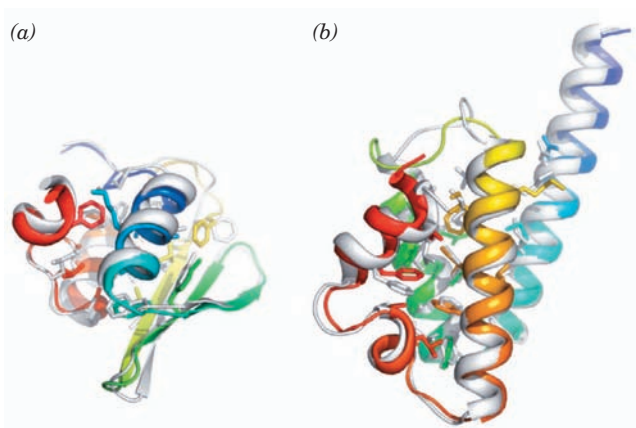


Figure 9-28 Examples of successful modeling predictions by **Rosetta**. Each panel shows the superposition of a predicted model (gray) with the corresponding experimentally determined X-ray structure colored in rainbow order from its N-terminus (blue) to its C-terminus (red) with core side chains drawn as sticks. (a) A protein of unknown function from *Thermus thermophilus* HB8 (PDBid 1WHZ). The backbones are aligned with an accuracy of 1.6 Å over 70 residues. (b) Protein BH3980 (10176605) from *Bacillus halodurans* (PDBid 2HH6). The backbones are aligned with an accuracy of 1.4 Å over 90 residues. [Courtesy of Gautam Dantas, Washington University School of Medicine.]

otherwise idle time of nearly 100,000 volunteered computers so that an average of $\sim 500,000$ CPU-hours can be devoted to predicting the structure of each domain. Examples of successful protein structure predictions by Rosetta are shown in Fig. 9-28.

C. Protein Design

Although we have not yet fully solved the protein folding problem, considerable progress has been made in solving the inverse problem: generating polypeptide sequences to assume specific three-dimensional structures, that is, **protein design**. This is probably because a polypeptide can be “overengineered” to take up a desired conformation. Consequently, protein design has provided insights into protein folding and stability, and it promises to yield useful proteins that are “made to order.” Protein design begins with a target structure such as a 4-helix bundle and attempts to find an amino acid sequence that will form this structure. The designed polypeptide is then synthesized and its structure elucidated.

Successful design requires not only that the desired fold be stable but that other folds be significantly less stable (by ~ 15 – 40 $\text{kJ} \cdot \text{mol}^{-1}$). Otherwise a sequence that has been found to be the most stable in the desired conformation may actually be more stable in other conformations. Before such **negative design** concepts were implemented, efforts to design proteins typically yielded an ensemble of molten globulelike states rather than the desired folds.

Most successful protein design projects have redesigned naturally occurring proteins so as to enhance their stability or provide them with new functionalities. Because of the

strict steric constraints in the cores of globular proteins, this has largely yielded proteins whose internal side chain packings resemble those of the original proteins. Consequently, designing a protein with a novel fold should be a more rigorous test of protein design methods. Indeed, it is unlikely that a polypeptide of arbitrary sequence will have a stable native structure.

Baker designed a topologically novel, 93-residue, α/β protein he named **Top7** as follows. A rough two-dimensional model of the target protein was created and structural constraints that defined its topology (e.g., hydrogen bonds and reverse turns) were identified. Rosetta was then used to generate 172 three-dimensional, backbone-only models with the required topology by assembling 3- and 9-residue fragments with the required secondary structures from the Protein Data Bank (PDB). Side chains were initially placed by considering all sets of energetically allowed torsion angles (which are known as **rotamers**) for each type of side chain but Cys at the 71 core positions and for only polar residues at the remaining 22 surface positions and using Rosetta to identify the lowest energy structures. These models were then structurally optimized through 15 cycles of using Rosetta to calculate the lowest energy backbone conformation for a fixed amino acid sequence followed by sequence redesign as previously described, ultimately yielding Top7. Although the structural differences between Top7 and its initial model were small (their backbones have an rmsd of 1.1 Å), they had dramatic changes in sequence (with only 31% of the residues in Top7 being identical to those in its initial model).

A gene for Top7 with a C-terminal His tag was synthesized (Section 7-6A) and expressed and the protein was purified by metal chelation affinity chromatography (Section 6-3Dg) followed by anion exchange chromatography (Section 6-3A). Top7 is highly soluble in aqueous solution and is monomeric as indicated by gel filtration chromatography (Section 6-3B). It is remarkably stable: Its circular dichroism (CD; Section 9-1Ca) spectrum at 98°C closely resembles that at 25°C. The X-ray structure of Top7 is all but identical, within experimental error, to the structure of the designed model: Their rmsd over all backbone atoms is 1.17 Å and many of their core side chains are effectively superimposable (Fig. 9-29). Evidently, protein folds that have not been observed in nature are not only physically possible but can be highly stable.

Baker also used the principles of protein design to generate enzymes that catalyze nonbiological reactions. He did so by grafting designed constellations of side chains onto the surfaces of naturally occurring proteins so as to form the desired active sites. As we shall see in Chapter 15, the catalytic activities of enzymes are extraordinarily sensitive to the positions of their catalytic groups.

4 PROTEIN DYNAMICS

The fact that X-ray studies yield time-averaged “snapshots” of proteins may leave the false impression that proteins have fixed and rigid structures. In fact, as is becoming increasing

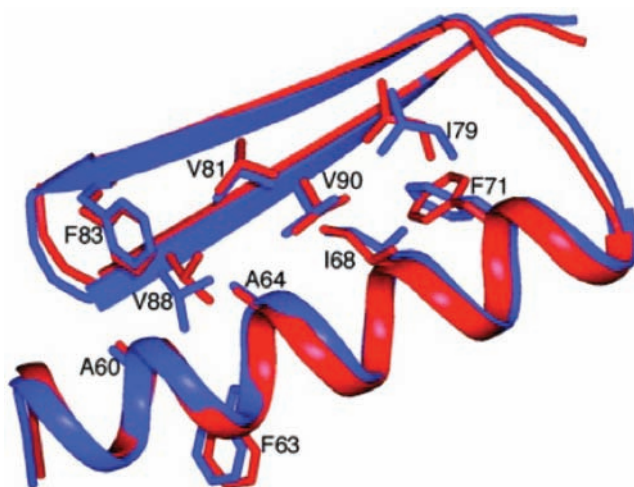


Figure 9-29 Superposition of the designed model of Top7 (blue) with its X-ray structure (red). Core side chains are drawn as sticks. [Courtesy of Gautam Dantas, Washington University School of Medicine. PDBid 1QYS.]

clear, *proteins are flexible and rapidly fluctuating molecules whose structural mobilities have considerable functional significance*. For example, X-ray studies indicate that the heme groups of myoglobin and hemoglobin are so surrounded by protein that there is no clear path for O₂ to approach or escape from its binding pocket. Yet we know that myoglobin and hemoglobin readily bind and release O₂. These proteins must therefore undergo conformational fluctuations, **breathing motions**, that permit O₂ reasonably free access to their heme groups (Fig. 9-30). The three-dimensional structures of myoglobin and hemoglobin undoubtedly evolved the flexibility to facilitate the diffusion of O₂ to its binding pocket.

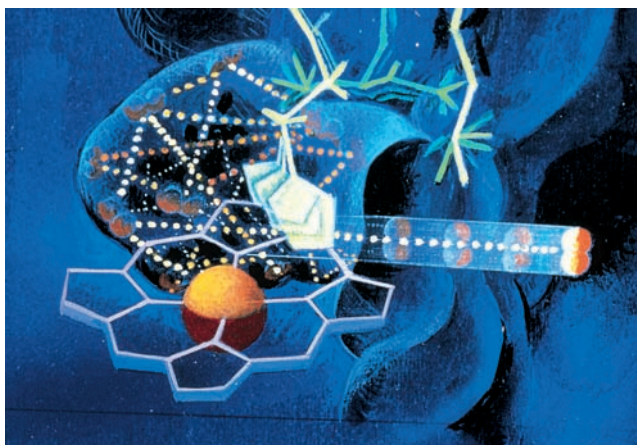


Figure 9-30 Conformational fluctuations in myoglobin. An artist's conception of the “breathing” motions in myoglobin that permit the escape of its bound O₂ molecule (double red spheres). The dotted lines trace a trajectory an O₂ molecule might take in worming its way through the rapidly fluctuating protein before finally escaping. O₂ binding presumably resembles the reverse of this process. [Illustration, Irving Geis. Image from the Irving Geis Collection, Howard Hughes Medical Institute. Reprinted with permission.]

The intramolecular motions of proteins have been classified into three broad categories according to their coherence:

1. Atomic fluctuations, such as the vibrations of individual bonds, which have time periods ranging from 10^{-15} to 10^{-11} s and spatial displacements between 0.01 and 1 Å.

2. Collective motions, in which groups of covalently linked atoms, which vary in size from amino acid side chains to entire domains, move as units with time periods ranging from 10^{-12} to 10^{-3} s and spatial displacements between 0.01 and >5 Å. Such motions may occur frequently or infrequently compared with their characteristic time period.

3. Triggered conformational changes, in which groups of atoms varying in size from individual side chains to complete subunits move in response to specific stimuli such as the binding of a small molecule, for example, the binding of ATP to GroEL (Section 9-2Ca). Triggered conformational changes occur over time spans ranging from 10^{-9} to 10^3 s and result in atomic displacements between 0.5 and >10 Å.

In this section, we discuss how these various motions are characterized and their structural and functional significance. We shall mainly be concerned with atomic fluctuations and collective motions; triggered conformational changes are considered in later chapters in connection with specific proteins.

a. Proteins Have Mobile Structures

X-ray crystallographic analysis is a powerful technique for the analysis of motion in proteins; it reveals not only the average positions of the atoms in a crystal, but also their mean-square displacements from those positions. X-ray analysis indicates, for example, that myoglobin has a rigid core surrounding its heme group and that the regions toward the periphery of the molecule have a more mobile character. Similarly, the apical domain of GroEL and the mobile loop of GroES are both highly flexible in the individual proteins, but when they interact in the GroEL–GroES–(ADP)₇ complex, they become significantly more rigid (Fig. 9-31; Section 9-2Ca). Indeed, as we have seen (Section 9-1Bg), portions of the binding sites of many proteins rigidify on binding their target molecules.

Molecular dynamics simulations, a theoretical technique pioneered by Martin Karplus, has revealed the nature of the atomic motions in proteins. In this technique, the atoms of a protein of known structure and its surrounding solvent are initially assigned random motions with velocities that are collectively characteristic of a chosen temperature. Then, after a time step of ~ 1 femtosecond ($1 \text{ fs} = 10^{-15} \text{ s}$) the aggregate effects of the various interatomic

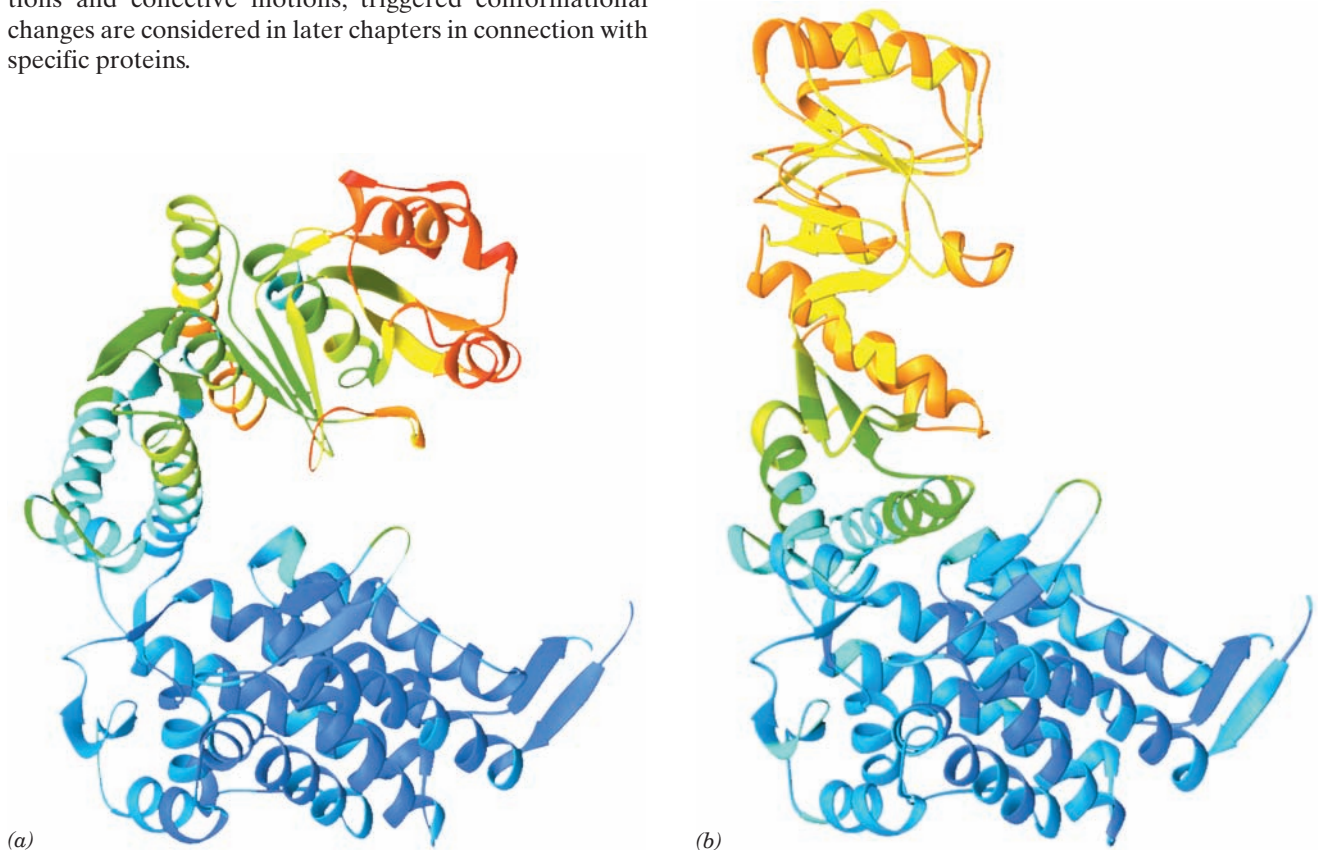


Figure 9-31 The mobility of the GroEL subunit. (a) In the X-ray structure of GroEL alone and (b) in the X-ray structure of GroEL–GroES–(ADP)₇. The polypeptide backbone is colored in rainbow order according to its degree of thermal motion, with blue being the least mobile (cool) and red being the most mobile (hot). The subunits are oriented as in Fig. 9-22a,b. Note that the

outer end of the apical domain, which functions to bind both substrate protein and the mobile loop of GroES (Section 9-2Ca), is more mobile in GroEL alone (*red and red-orange*) than it is in GroEL–GroES–(ADP)₇ (*orange and yellow*). [Based on X-ray structures by Axel Brünger, Arthur Horwich, and Paul Sigler, Yale University. PDBids (a) 1OEL and (b) 1AON.]

forces in the system (those due to departures from ideal covalent bond lengths, angles, and torsion angles as well as noncovalent interactions) on the velocities of each of its atoms are calculated according to Newton's equations of motion. Since all the atoms in the system will have moved after this time step (by a distance that is only a small fraction of a bond length), the interatomic forces (their potential field) on each atom will likewise have changed (although by only a small amount). Then, using this altered potential field together with the new positions and velocities of the atoms, the calculation is repeated for an additional time step. This computationally intensive process has been iterated for up to $\sim 1 \mu\text{s}$ for ~ 100 -residue proteins (a time that is increasing with the available computational power), thereby yielding a record of the positions and velocities of all the atoms in the system over this time period.

Molecular dynamics simulations (e.g., Fig. 9-32) have revealed that *a protein's native structure really consists of a large collection of conformational substates that have essentially equal stabilities*. These substates, which each have slightly different atomic arrangements, randomly interconvert at rates that increase with temperature. Consequently, the interior of a protein typically has a fluidlike character for structural displacements of up to $\sim 2 \text{ \AA}$, that is, over excursions that are somewhat larger than a bond length.

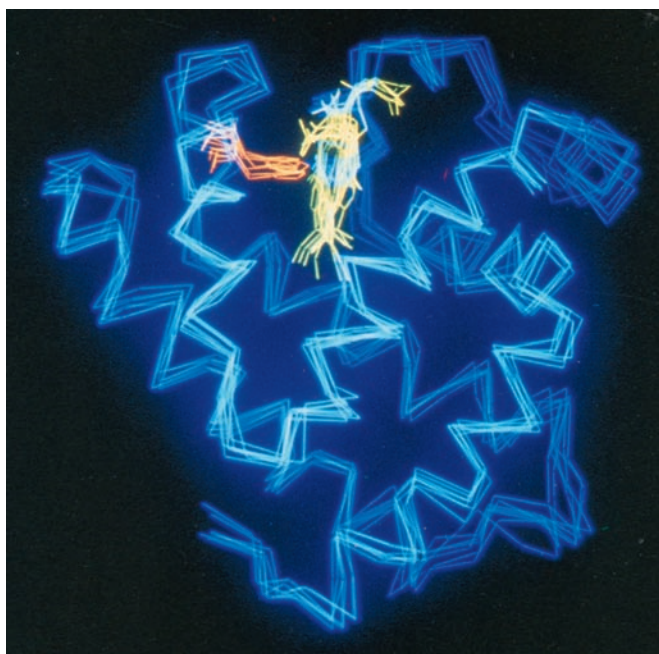
Gregory Petsko and Dagmar Ringe have demonstrated the functional significance of the internal motions in proteins. Both experimental and theoretical evidence indicates that below $\sim 220 \text{ K}$ (-53°C), collective motions in proteins are arrested, leaving atomic fluctuations as the dominant intramolecular motions. For example, X-ray studies have shown that, at 228 K , the enzyme RNase A, in its crystalline form, readily binds an unreactive substrate analog (protein crystals generally contain large solvent-filled channels through which small molecules rapidly diffuse; at low temperatures, the water is

prevented from freezing by the addition of an antifreeze such as methanol). Yet, when the same experiment is performed at 212 K , the substrate analog does not bind to the enzyme, even after 6 days of exposure. Likewise, at 228 K , substrate-free solvent washes bound substrate analog out of the crystal within minutes but, if the temperature is first lowered to 212 K , the substrate analog remains bound to the crystalline enzyme for at least 2 days. Evidently, RNase A assumes a glasslike state below 220 K that is too rigid to bind or release substrate. In terms of landscape theory, this is interpreted as the protein being trapped in a single energy well.

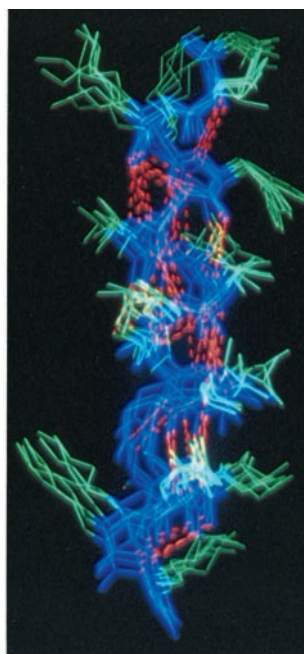
b. Protein Core Mobility Is Revealed by Aromatic Ring Flipping

The rate at which an internal Phe or Tyr ring in a protein undergoes 180° "flips" about its $C_\beta-C_\gamma$ bonds is indicative of the surrounding protein's rigidity. This is because, in the close packed interior of a protein, these bulky asymmetric groups can move only when the surrounding groups move aside transiently (although note that these rings have the shape of flattened ellipsoids rather than thin disks).

NMR spectroscopy can determine the mobilities of protein groups over a wide range of time scales. Consequently, the rate at which a particular aromatic ring in a protein flips is best inferred from an analysis of its NMR spectrum (infrequent motions such as ring flipping are not detected by X-ray crystallography since this technique only reveals the average structure of a protein). NMR measurements indicate that the ring flipping rate varies from over 10^6 s^{-1} to one of immobility ($< 1 \text{ s}^{-1}$) depending on both the protein and the location of the aromatic ring within the protein. For example, **bovine pancreatic trypsin inhibitor (BPTI)** is a 58-residue monomeric protein that has eight Phe and Tyr residues. At 4°C , four of these Phe and Tyr rings flip at rates $> 5 \times 10^4 \text{ s}^{-1}$, whereas the remaining four rings flip at rates



(a)



(b)

Figure 9-32 The internal motions of myoglobin as determined by a molecular dynamics simulation. Several "snapshots" of the molecule calculated at intervals of $5 \times 10^{-12} \text{ s}$ are superimposed. (a) The C_α backbone and the heme group. The backbone is shown in blue, the heme in yellow, and the His residue liganding the Fe in orange. (b) An α helix. The backbone is shown in blue, the side chains in green, and the helix hydrogen bonds as dashed orange lines. Note that the helices tend to move in a coherent fashion so as to retain their shape. [Courtesy of Martin Karplus, Harvard University.]

ranging between 30 and $<1 \text{ s}^{-1}$. These ring-flipping rates sharply increase with temperature, as expected.

c. Infrequent Motions Can Be Detected through Hydrogen Exchange

Conformational changes occurring over time spans of more than several seconds can be chemically characterized through hydrogen exchange studies (Sections 9-1Cc). These show that the exchangeable protons of native proteins exchange at rates that vary from milliseconds to many years (Fig. 9-33). Protein interiors, as we have seen (Section 8-3B), are largely excluded from contact with their surrounding aqueous solvent, and, moreover, protons cannot exchange with solvent while they are engaged in hydrogen bonding. The observation that the internal protons of proteins do, in fact, exchange with solvent must therefore be a consequence of transient local unfolding or “breathing” that physically and chemically exposes these exchangeable protons to the solvent. Hence, *the rate at which a particular proton undergoes hydrogen exchange is a reflection of the conformational mobility of its surroundings*. This hypothesis

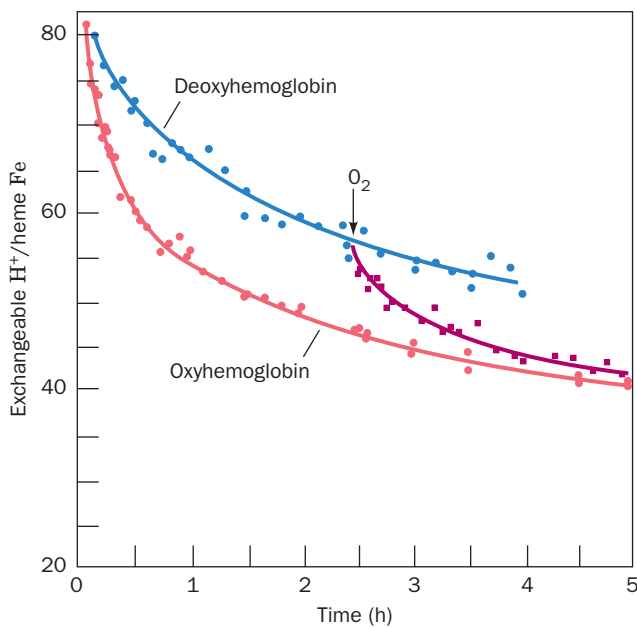


Figure 9-33 The hydrogen–tritium “exchange-out” curve for hemoglobin that has been pre-equilibrated with tritiated water. The vertical axis expresses the ratio of exchangeable protons to heme Fe atoms. Exchange-out was initiated by replacing the protein’s tritiated water solvent with untritiated water through rapid gel filtration (Section 6-3B). As the exchange-out proceeded, additional gel filtration separations were performed and the amount of tritium remaining bound to the protein was measured. At the arrow, O_2 was added to exchanging deoxyhemoglobin (hemoglobin lacking bound O_2). The changing slopes of these curves indicate that the hydrogen exchange rates of the ~ 80 exchangeable protons of each hemoglobin subunit vary by factors of many decades and that O_2 binding increases the exchange rates for ~ 10 of these protons (the structural changes that O_2 binding induces in hemoglobin are discussed in Section 10-2). [After Englander, S.W. and Mauel, C., *J. Biol. Chem.* **247**, 2389 (1972).]

is corroborated by the observation that the hydrogen exchange rates of proteins decrease as their denaturation temperatures increase and that these exchange rates are sensitive to the proteins’ conformational states (Fig. 9-33).

5 CONFORMATIONAL DISEASES: AMYLOID AND PRIONS

Most proteins in the body maintain their native conformations or, if they become partially denatured, are either renatured through the auspices of molecular chaperones (Section 9-2C) or are proteolytically degraded (Section 32-6). However, ~ 35 different, often fatal, human diseases are associated with the extracellular deposition of normally soluble proteins in certain tissues in the form of insoluble aggregates known as **amyloid** (starchlike; a misnomer because it was originally thought that this material resembled starch). These include **Alzheimer’s disease** and **Parkinson’s disease**, neurodegenerative diseases that mainly strike the elderly; the **transmissible spongiform encephalopathies (TSEs)**, a family of infectious neurodegenerative diseases that are propagated in a most unusual way; and the **amyloidoses**, a series of diseases caused by the deposition of an often mutant protein in organs such as the heart, liver, or kidney. The deposition of amyloid interferes with normal cellular function, resulting in cell death and eventual organ failure.

Although the various types of amyloidogenic proteins are unrelated and their native structures have widely different folds, their amyloid forms have remarkably similar core structures: Each consists of an array of ~ 10 -nm-diameter **amyloid fibrils** (Fig. 9-34) in which, as infrared, NMR, and X-ray diffraction methods indicate, certain segments

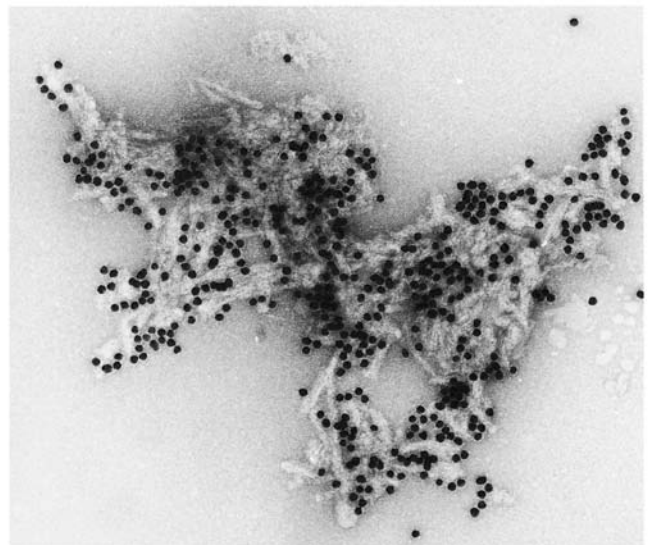


Figure 9-34 An electron micrograph of amyloid fibrils of the protein PrP 27–30 (Section 9-5Ce). They are visually indistinguishable from amyloid fibrils formed by other proteins. The black dots are colloidal gold beads that are coupled to anti-PrP antibodies that are adhering to the PrP 27–30. [Courtesy of Stanley Prusiner, University of California at San Francisco Medical Center.]

of the proteins form extended β sheets whose planes extend parallel to the fibril axis so that their β strands are perpendicular to the fibril axis (see below). Thus, *these proteins each have two radically different stable conformations, their native forms and their amyloid forms.*

We begin this section with a discussion of the amyloidoses as exemplified by **islet amyloid polypeptide (IAPP)**; also called **amylin**) and certain mutants forms of lysozyme. We then consider Alzheimer's disease and finally the TSEs and their bizarre mode of propagation.

A. Amyloid Diseases

Many amyloidogenic proteins are mutant forms of normally occurring proteins. These include lysozyme (an enzyme that hydrolyzes bacterial cell walls; Section 15-2) in the disease **lysozyme amyloidosis**, **transthyretin** [Fig. 8-66; a blood plasma protein that functions as a carrier for the water-insoluble thyroid hormone **thyroxin** (Section 19-1D) as well as retinol through its association with retinol binding protein (Section 8-3Bg)] in **familial amyloidotic polyneuropathy**, and **fibrinogen** (the precursor of **fibrin**, which forms blood clots; Section 35-1A) in **fibrinogen amyloidosis**. Most such diseases do not present (become symptomatic) until the third to seventh decades of life and typically progress over 5 to 15 years ending in death.

a. IAPP Forms a Cross- β Spine Structure

IAPP is a 37-residue peptide that is associated with **type 2 diabetes mellitus** (also called **non-insulin dependent and maturity-onset diabetes mellitus**; Section 27-4B), an often fatal disease that afflicts ~ 100 million mainly older people worldwide. IAPP is expressed and secreted by the pancreatic β islet cells, which also synthesize the polypeptide hormone insulin (Fig. 7-2; whose lack is responsible for **type I** or **juvenile-onset diabetes mellitus**). Although the role of IAPP in the development of type 2 diabetes is unclear, the pancreases of 95% of individuals with type 2 diabetes contain amyloid deposits of IAPP with the extent of this deposition increasing with the severity of the disease. Interestingly, mouse IAPP, which differs from human IAPP in 6 of its 37 residues, does not form amyloid and mice do not develop type 2 diabetes, although transgenic mice that express human IAPP sometimes do so.

Attempts to crystallize IAPP, either in its native conformation or its amyloid form, have been unsuccessful. However, through the use of structure prediction techniques (Section 9-3B), Baker and David Eisenberg identified two of its segments that have high fibril-forming potential: **NNFGAIL** and **SSTNVG**, which comprise IAPP's residues 21 to 27 and 28 to 33 (5 of the 6 residue differences between human and mouse IAPP occur in the segment 23–29). Both of these peptides form amyloidlike fibrils as well as very thin needle-shaped crystals.

The X-ray structure of **SSTNVG**, determined by Eisenberg, reveals that this hexapeptide forms an extended parallel β sheet with two such sheets facing each other such that their protruding side chains interdigitate so tightly that they completely exclude water (Fig. 9-35a). The X-ray

structure of **NNFGAIL** is similar but contains a pronounced bend in its backbone, such that the interface between the two β sheets is formed between main chains (Fig. 9-35b). Such structures, which are known as **cross- β spines**, are assumed by a variety of other amyloid-forming peptides, although many of them contain antiparallel rather than parallel β sheets.

Figures 9-35c–e exhibit a model of the IAPP amyloid fibril based on the forgoing X-ray structures. Its is a 4-sheet,

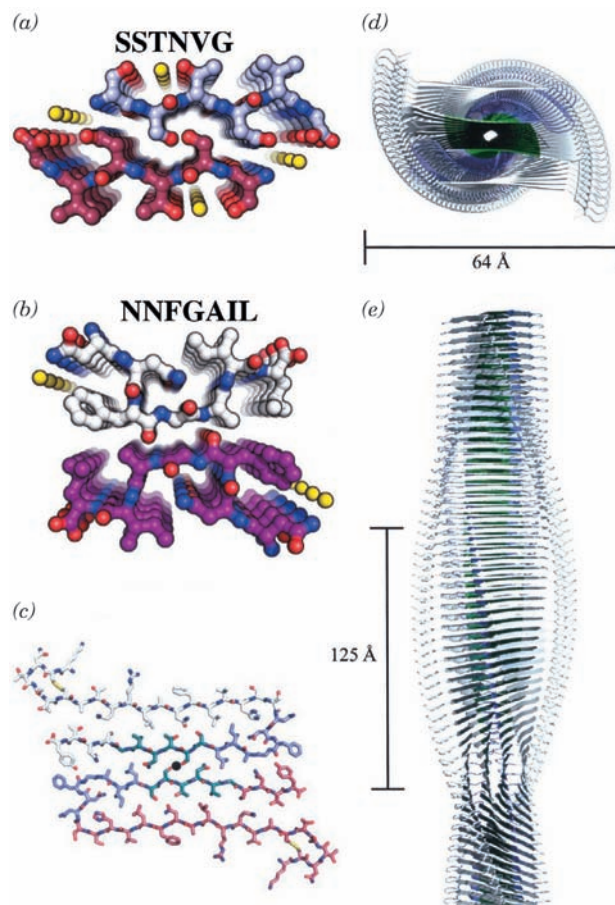


Figure 9-35 The IAPP amyloid fibril. (a) The X-ray structure of **SSTNVG** drawn in ball-and-stick form as viewed along the planes of the β sheets that they form. Atoms are colored according to type with C on one chain white and the other magenta, N blue, O red, and water molecules represented by yellow spheres. (b) The X-ray structure of **NNFGAIL** viewed and colored as in Part a. (c) Model of the fibril as viewed along its axis (black dot). Atoms are colored according to type with the C atoms of the **SSTNVG** segment green, those of the **NNFGAIL** segment light blue, those of the N-terminal 20 residues and the C-terminal 4 residues light red or white, N blue, O red, and S atoms, which form disulfide bonds, yellow. (d) Schematic view along the fibril axis in which the **SSTNVG** segments are green, the **NNFGAIL** segments are light blue, and the modeled residues are white. (e) View perpendicular to the fibril axis represented as in Part d. The helix has a gentle left-handed twist of 3.4 Å per layer so that it makes a quarter turn every 125 Å. [Courtesy of David Eisenberg, UCLA. PDBids 3DG1 for **SSTNVG** and 3DGJ for **NNFGAIL**.]

left-handed helix which has a pitch (rise per turn) of 500 Å. The SSTNVG segment's cross- β spine is centered on the fibril axis. It is extended by NNFGAIL segments that also form hairpin turns. IAPP's final 4 residues were modeled to complete the inner β strands and its initial 20 residues were modeled to form the outer strands. The calculated X-ray diffraction pattern of this assembly closely resembles the observed diffraction pattern of IAPP fibrils.

Other amyloid-forming proteins have segments that form similar cross- β spines. However, the structures of the loops connecting the β strands must vary with the identity of the protein.

b. Amyloidogenic Lysozyme Variants Have Conformationally Flexible Native Structures

There are two known amyloidogenic variants of the 130-residue human lysozyme, I56T and D67H. These form amyloid fibrils that are deposited in the viscera (internal organs), usually resulting in death by the fifth decade. The amyloid fibrils consist exclusively of the variant lysozymes, thereby explaining why these mutations are dominant. Structural studies on these variant proteins have shed light on how they form amyloid fibrils.

The X-ray structures of both mutant lysozymes resemble that of the wild-type enzyme. However, the replacement of Asp 67 by His interrupts a network of hydrogen bonds that stabilizes the domain containing the structure's only β sheet (its so-called β domain), resulting in the movements of the β sheet and an adjoining loop away from each other by displacements of up to 11 Å (Fig. 9-36). Although the replacement of Ile 56 by Thr causes only subtle changes in the protein structure, it insinuates a hydrophilic residue in a critical hydrophobic interface that links the protein's two domains.

The melting temperatures (T_m 's) of both variants are at least 10°C less than those of the wild-type enzyme, and both variants eventually lose all enzymatic activity when incubated at physiological temperature and pH (37°C and 7.4), conditions under which wild-type lysozyme remains fully active. The variants also aggregate on heating *in vitro*, and a variety of physical measurements indicate that, in doing so, they form amyloidlike fibrils. In hydrogen exchange experiments (Section 9-1Cc), wild-type lysozyme strongly protects 55 protons from exchange with D₂O under conditions (37°C and pH 5) in which these protons are essentially unprotected in the amyloidogenic variants, thereby confirming that the native protein's tertiary structure is greatly loosened in the two mutant forms. This suggests that the partially folded, aggregation-prone forms are in dynamic equilibrium with the native conformation, even under conditions in which the native state is thermodynamically stable [keep in mind that the ratio of unfolded (U) to native (N) protein molecules in the reaction $N \rightleftharpoons U$ is governed by Eq. [3.17]: $[U]/[N] = e^{-\Delta G^\circ/RT}$, where ΔG° is the standard free energy of unfolding, so that as ΔG° decreases, the proportion of U increases]. It has therefore been proposed that lysozyme fibrillogenesis is initiated by the association of the β domains of two partially unfolded

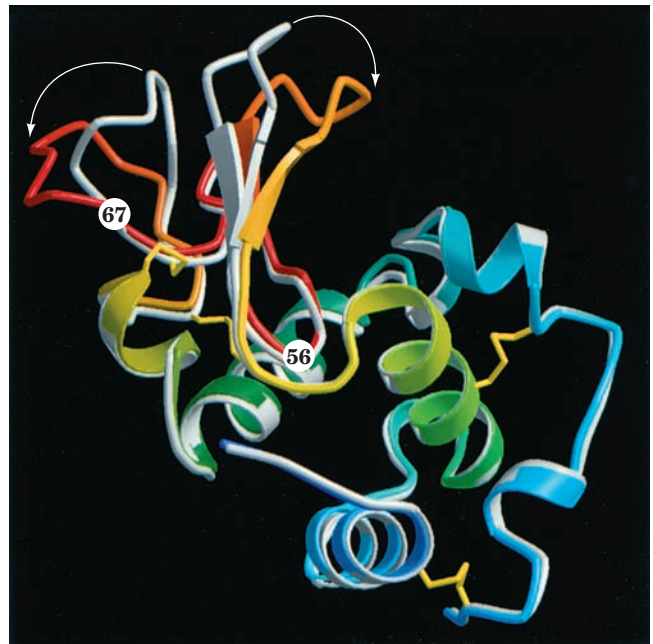


Figure 9-36 Superpositions of wild-type human lysozyme and its D67H mutant. Wild-type lysozyme is gray and its D67H mutant is color-ramped in rainbow order from blue at its N-terminus to red and back to blue at its C-terminus. The white arrows indicate the conformational shifts of residues 45 to 54 and 67 to 75 in the D67H mutant relative to those in the wild-type protein. The four disulfide bonds present in both the wild-type and mutant protein are shown in yellow. The positions of residues 56 and 67 are indicated. [Courtesy of Margaret Sunde, University of Oxford, UK and Colin Blake, University of Oxford, UK.]

lysozyme variants 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 in a process that may involve the conformational conversion of α helices to β strands. Such an autocatalytic refolding process may be a general mechanism for amyloid fibrillogenesis. However, the several decades that many hereditary amyloid diseases require to become symptomatic suggest that the spontaneous generation of an amyloid nucleus is a rare event, that is, has a high free energy of activation (activation barriers and their relationship to reaction rates are discussed in Section 14-1C).

B. Alzheimer's Disease

Alzheimer's disease (AD), a neurodegenerative condition that afflicts ~20 million mainly elderly people worldwide (~10% of those over the age of 65 and ~50% of those over 85), causes devastating mental deterioration and eventual death. It is characterized by brain tissue containing abundant amyloid plaques (deposits) surrounded by dead and dying neurons. In addition, many neuronal cell bodies contain abnormal ~20-nm-diameter fibers known as **neurofibrillary tangles**. The amyloid plaques consist mainly of amyloid fibrils of a 40- to 42-residue peptide named **amyloid- β**

peptide (A β) [the neurofibrillary tangles, which we shall not further discuss, consist of a hyperphosphorylated form of a protein named **tau** that is normally associated with microtubules (Section 1-2Ae)].

The sequence of the gene encoding A β , which was identified via reverse genetics (Section 7-2D) based on the sequence of A β , reveals that A β is a segment of a 770-residue transmembrane protein named **A β precursor protein (β PP)**; transmembrane proteins are discussed in Section 12-3A). β PP has a receptorlike sequence (Section 19-2B) although its normal function is unknown. A β is excised from β PP in a multistep process through the actions of two membrane-anchored proteolytic enzymes dubbed β - and γ -**secretases**.

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

The age-dependence of AD suggests that **β -amyloid** deposition is an ongoing process, at least in the later decades of life. Indeed, there are several rare variants of the β PP gene with mutations in their A β regions that result in the onset of AD as early as the fourth decade of life. These mutations have been shown to affect the proteolytic processing of β PP in a way that increases the rate of A β production. A similar phenomenon is seen in **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. Individuals with Down's syndrome invariably develop AD by their fortieth year. This is because the gene encoding β PP is located on chromosome 21, and hence individuals with Down's syndrome produce β PP and presumably A β at an accelerated rate.

A second gene that has been implicated in the premature onset of AD encodes the cholesterol transport protein **apolipoprotein E (apoE)**; (Section 12-5Bd). The *apoE* gene has several normally occurring variants (alleles) in the population, one of which, ***apoE4***, is a major risk factor for both the development of AD and its earlier onset. Moreover, AD victims with *apoE4* have significantly higher densities of β -amyloid plaques in their brain tissue than AD victims with other apoE variants. These observations motivated experiments showing that ApoE4 induces enhanced aggregation of synthetic A β *in vitro*. This suggests that ApoE4 facilitates A β aggregation *in vivo* (although another possibility is that ApoE4 inhibits the clearance of A β from the extracellular spaces).

There is, at present, no known treatment that arrests the progress of AD. However, the foregoing suggests several strategies for therapeutic intervention, including decreasing the rate of production of A β through the administration of substances that inhibit the action of β - or γ -secretase and through the administration of agents that

would interfere with the formation of β -amyloid fibrils from soluble A β .

C. Prion Diseases

Certain infectious diseases that affect the mammalian central nervous system were originally classified as being 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, so named for the tendency of infected sheep to scrape off their wool [they rub against fences in an effort to stay upright due to ataxia (loss of muscle coordination)]; **bovine spongiform encephalopathy (BSE or mad cow disease)**, which similarly afflicts cattle; and **kuru**, a degenerative brain disease that occurred among the Fore people of Papua New Guinea (kuru means trembling) and that was transmitted by ritual cannibalism. There is also a sporadic (apparently spontaneously arising) human disease with similar symptoms, **Creutzfeldt-Jakob disease (CJD)**, a rare, progressive, cerebellar disorder, which resembles and may be identical to kuru. These diseases, all of which are ultimately fatal, have similar symptoms, which suggests that they are closely related. Since, in all of these diseases, neurons develop large vacuoles that gives brain tissue a spongelike microscopic appearance, they are collectively known as **transmissible spongiform encephalopathies (TSEs)**. None of the TSEs exhibit any sign of an inflammatory process or fever, which indicates that the immune system, which is not impaired by the disease, is not activated by it.

The classic technique for isolating the agent causing an infectious disease involves the fractionation of diseased tissue as monitored by assays for the disease. The long incubation time for scrapie, the most extensively studied "slow virus" disease, enormously hampered initial efforts to characterize its disease agent. Indeed, in the early work on scrapie in the 1930s, an entire herd of sheep and several years of observation were necessary to evaluate the results of a single fractionation. Assays for scrapie were greatly accelerated, however, by the discovery that Syrian hamsters, after intracerebral inoculation of the scrapie agent, develop the disease in a time, minimally 60 days, that decreases as the dose given is increased. Using a hamster assay, Stanley Prusiner purified the scrapie agent to a high degree and was instrumental in characterizing it.

a. Scrapie Is Caused by Prion Protein

The scrapie agent apparently is a single species of protein. This astonishing conclusion was established by the observations that the scrapie agent is inactivated by substances that modify proteins, such as proteases, detergents, phenol, urea, and reagents that react with specific amino acid side chains, whereas it is unaffected by agents that alter nucleic acids, such as nucleases, UV irradiation, and substances that specifically react with nucleic acids. For example, scrapie agent is inactivated by treatment with **diethylpyrocarbonate**, which carboxyethylates the His residues of proteins (Fig. 9-37a), but is unaltered by the cytosine-specific reagent **hydroxylamine** (Fig. 9-37b). In

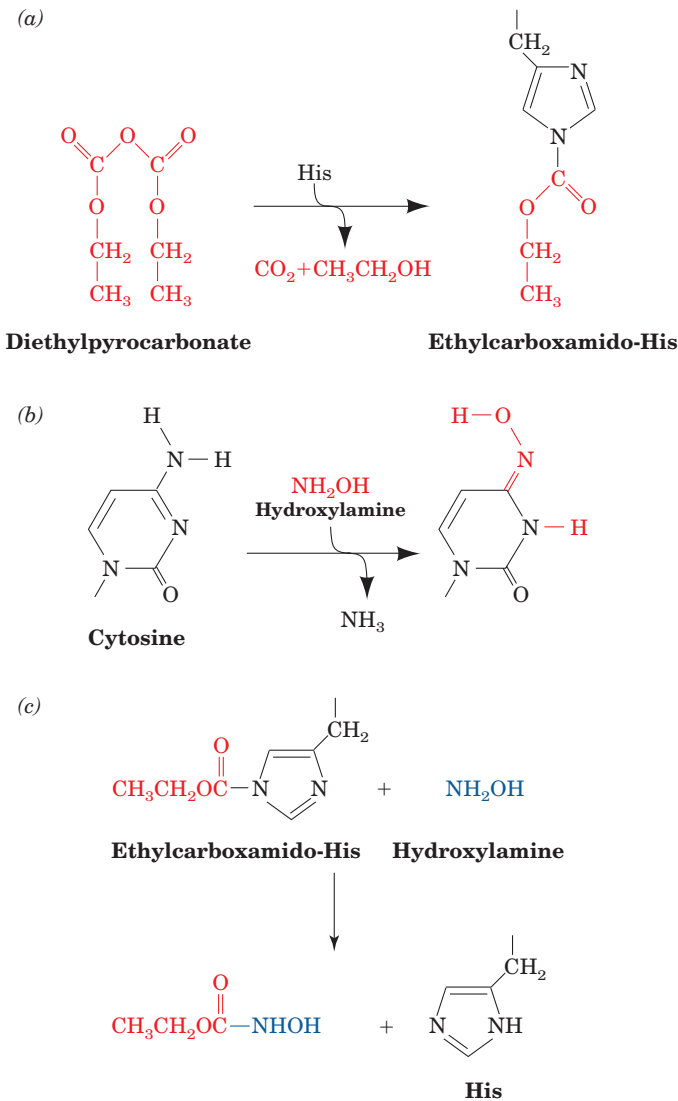


Figure 9-37 Evidence that the scrapie agent is a protein.

(a) Scrapie agent is inactivated by treatment with diethylpyrocarbonate, which specifically reacts with His side chains. (b) Scrapie agent is unaffected by treatment by hydroxylamine, which reacts with cytosine residues. (c) However, hydroxylamine rescues diethylpyrocarbonate-inactivated scrapie reagent, presumably by the reaction shown.

fact, the infectivity of diethylpyrocarbonate-inactivated scrapie agent is restored by treatment with hydroxylamine, presumably by the reaction shown in Fig. 9-37c.

The novel properties of scrapie agent, which distinguish it from viruses and plasmids, have resulted in its being termed a **prion** (for *proteinaceous infectious particle* that lacks nucleic acid). The scrapie protein, which is named **PrP** (for *Prion Protein*), consists of 280 mostly hydrophobic residues. This hydrophobicity, as we shall see below, causes partially proteolyzed PrP to aggregate as clusters of rodlike particles. There is a close resemblance between these clusters and the amyloid fibrils that are seen on electron microscopic examination of prion-infected brain tissue (Fig. 9-34). In fact, brain tissue from CJD victims contains

protease-resistant protein that cross-reacts with antibodies raised against scrapie PrP.

b. PrP Is a Widely Expressed Product of a Normal Cellular Gene That Has No Known Function

The bizarre composition of prions immediately raises the question: How are they synthesized? Three possibilities have been suggested:

1. Despite all evidence to the contrary, prions contain a nucleic acid genome that is somehow shielded from detection; that is, prions are conventional viruses. The enormous and still growing body of information concerning the nature of prions, however, makes this notion increasingly untenable.

2. Prions might somehow specify their own amino acid sequence by “reverse translation” to yield a nucleic acid that is normally translated by the cellular system. Such a process, of course, would directly contravene the “central dogma” of molecular biology (Section 5-4), which states that genetic information flows unidirectionally from nucleic acids to proteins. Alternatively, prions might directly catalyze their own synthesis. Such protein-directed protein synthesis is likewise unknown (although many small bacterial polypeptides are enzymatically rather than ribosomally synthesized).

3. Susceptible cells carry a gene that codes for the corresponding PrP. Infection of such cells by prions activates this gene and/or alters its protein product in some autocatalytic way.

The latter hypothesis seems to be the most plausible mechanism of prion replication. Indeed, the use of oligonucleotide probes complementary to the PrP gene (which is named *Prn-p* for *prion protein*), as inferred from the amino acid sequence of PrP’s N-terminus (Section 7-2D), established that the brains of both scrapie-infected and normal mice contain *Prn-p*. The most surprising discovery, however, is that *Prn-p* is transcribed at similar levels in both normal and scrapie-infected brain tissue. Moreover, the use of the above probes has revealed that *Prn-p* genes occur in all vertebrates so far tested, including humans, as well as in invertebrates such as *Drosophila*. This evolutionary conservation suggests that PrP, a membrane-anchored protein (via glycosylphosphatidylinositol groups; Section 12-3Bc) that occurs mainly on neuron surfaces, has an important function. Thus it came as a further surprise that knockout mice (Section 5-5H) in which both *Prn-p* genes have been disrupted appear to be normal and that mating two such *Prn-p*^{0/0} mice gives rise to normal *Prn-p*^{0/0} progeny (although there is some evidence that *Prn-p*^{0/0} mice develop neurological abnormalities late in life). Nevertheless, evidence is accumulating that PrP is normally a cell-surface signal receptor, although the identity of its corresponding signal and its consequences are as yet unknown.

c. Scrapie Disease Requires the Expression of the Corresponding PrP^C Protein

Prn-p^{0/0} mice remain completely free of scrapie symptoms after inoculation with a dose of mouse scrapie PrP

(PrP^{Sc}; Sc for scrapie) that causes wild-type (*Prn-p*^{+/+}) mice to die of scrapie within 6 months after inoculation. Evidently, PrP^{Sc} induces the conversion of normal PrP (**PrP^C**; C for cellular) to PrP^{Sc}. This unorthodox notion, the so-called **prion hypothesis**, is supported by the observation that when wild-type mice are inoculated with PrP^{Sc} that has been continuously passaged (incubated) in hamsters, the incubation time for developing disease symptoms is, at first, 500 days but then, in all further passages in mice, diminishes to 140 days. Conversely, when PrP^{Sc} that has been passaged in mice is inoculated into hamsters, the incubation time is first 400 days but subsequently shortens to 75 days. This suggests that the conversion of host PrP^C (whose sequence in mice differs from that in hamsters) to PrP^{Sc} by a foreign PrP^{Sc} is a rare event; once it has occurred, however, the newly formed host PrP^{Sc} catalyzes the conversion much more efficiently. Indeed, after inoculation with hamster PrP^{Sc}, transgenic mice expressing hamster PrP have incubation times that are reduced to between 48 and 250 days, depending on the transgenic line.

The foregoing experiments provide indirect support for the prion hypothesis. However, direct support has recently been provided by the demonstration that PrP^{Sc} induces the conversion of PrP^C to PrP^{Sc} in a cell-free system.

d. Mutant *Prn-p* Genes Give Rise to Prion Diseases

Three dominantly inherited neurodegenerative disorders in humans have been traced to mutations in the *Prn-p* gene. These are **familial CJD**, **Gerstmann–Sträussler–Scheinker syndrome (GSS)**, and **fatal familial insomnia (FFI)**. All of them are extremely rare. In fact, FFI has been found in only five families. The mutant PrP^{Sc}s causing these diseases are nevertheless infectious.

e. PrP^{Sc} Is a Stable Conformational Variant of PrP^C

The NMR structure of residues 23 to 230 of the 280-residue human PrP^C, determined by Kurt Wüthrich, consists of a flexibly disordered (and hence unobserved) 98-residue N-terminal “tail” and a 110-residue C-terminal globular domain containing three α helices and a short 2-stranded antiparallel β sheet (Fig. 9-38a). As expected, this structure closely resembles those of the homologous mouse and hamster PrP^Cs.

How does PrP^{Sc} differ from PrP^C? The direct sequencing of PrP^{Sc} indicates that its amino acid sequence is identical to that deduced from the *Prn-p* gene sequence, thereby eliminating any post-transcriptional sequence variation as a possible cause for the pathogenic properties of PrP^{Sc}. Furthermore, mass spectrometric studies on PrP^{Sc} designed to reveal previously uncharacterized post-translational modifications indicated that, in fact, PrP^{Sc} and PrP^C are chemically identical. Thus, although the possibility that only a small fraction of PrP^{Sc} is chemically modified has not been eliminated, it seems more likely that PrP^{Sc} and PrP^C differ in their secondary and/or tertiary structures. Unfortunately, the insolubility of PrP^{Sc} (see below) has precluded its structural determination. However, CD measurements show that, in fact, the conformations of PrP^{Sc} and PrP^C are quite different: PrP^C has a high (~40%) α helix content but little (~3%) β sheet content (in good agreement with the NMR structure of its globular domain), whereas PrP^{Sc} has a lesser (~30%) α helix content but a high (~45%) β sheet content. In a plausible model of PrP^{Sc} (Fig. 9-38b), its N-terminal region has refolded to form a so-called **β helix** in which the polypeptide strand forms a left-handed helix containing three parallel β sheets. Only the two C-terminal helices of PrP^C, which are

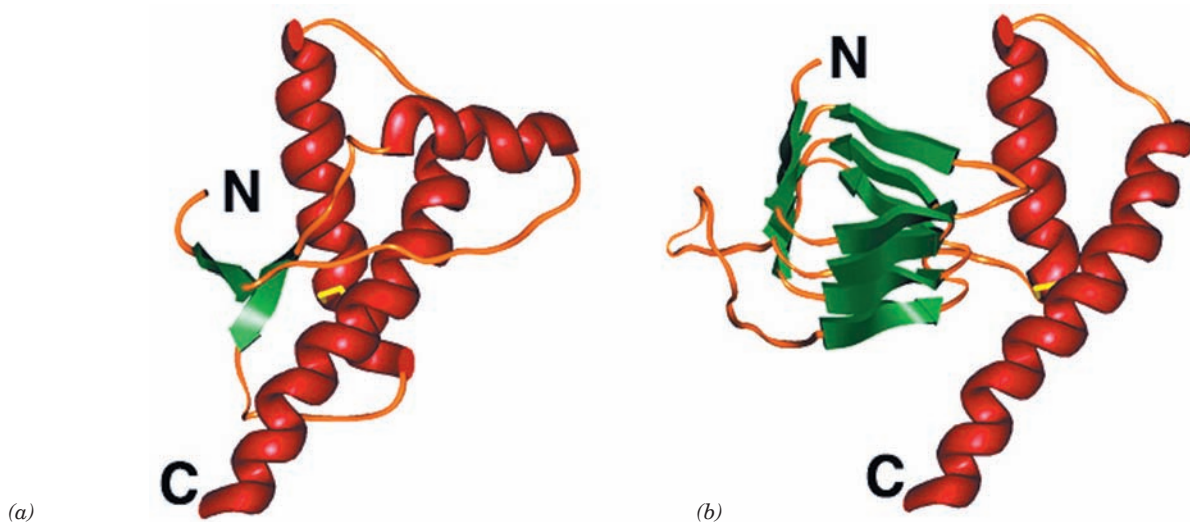


Figure 9-38 Prion protein conformations. (a) The NMR structure of human prion protein (PrP^C). The protein is drawn in ribbon form colored according to its secondary structure with helices red, β sheets green, and other segments orange. Its disulfide bond is shown in stick form in yellow. Its N-terminal “tail” (residues 23–121) is flexibly disordered (the protein’s

N-terminal 23 residues had been post-translationally excised). (b) A plausible model for the structure of PrP^{Sc} represented as in Part a. [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.]

joined by a disulfide bond, maintain their original conformation. The high β sheet content of PrP^{Sc} would, presumably, facilitate the aggregation of PrP^{Sc} as amyloid fibrils. *Evidently, the PrP^C \rightarrow PrP^{Sc} conformational change is autocatalytic; that is, PrP^{Sc} induces PrP^C to convert to PrP^{Sc}.* In fact, PrP^{Sc} in a cell-free system has been shown to catalyze the conversion of PrP^C from an uninfected source to PrP^{Sc}.

In cells, PrP^{Sc} is deposited in cytosolic vesicles rather than being anchored to the cell-surface membrane as is PrP^C. Both PrP^C and PrP^{Sc} are subject to eventual proteolytic degradation in the cell (Section 32-6). However, although PrP^C is completely degraded, PrP^{Sc} only loses its N-terminal 67 residues to form a 27- to 30-kD protease-resistant core, known as **PrP 27–30**, which still exhibits a high β sheet content. *PrP 27–30 then aggregates to form the amyloid plaques that appear to be directly responsible for the neuronal degeneration characteristic of prion diseases.*

According to the prion hypothesis, sporadically occurring prion diseases such as CJD (which strikes one person per million per year) arise from the spontaneous although infrequent conversion of sufficient quantities of PrP^C to PrP^{Sc} to support the autocatalytic conformational isomerization reaction. This model is corroborated by the observation that transgenic mice that overexpress wild-type *Prn-p* invariably develop scrapie late in life. The prion hypothesis similarly explains inherited prion diseases such as FFI as arising from a lower free energy barrier and hence higher rate for the conversion of the mutant PrP^C to PrP^{Sc} relative to that of normal PrP^C.

f. Prions Have Different Strains

Prions from different sources, when passaged in mice or hamsters, reproducibly exhibit characteristic incubation times, neurological symptoms, and neuropathologies. *Evidently, there are different strains of prions, each of whose corresponding PrP^{Sc}s must have a different stable conformation and induce PrP^C to take up this conformation.* The existence of different prion strains (as many as 30 for scrapie in sheep and at least 4 for CJD in humans) was cited as evidence against the prion hypothesis. However, there is now ample physical evidence that the PrP^{Sc}s in different prion strains have different structures.

BSE or mad cow disease was first reported in the U.K. in late 1985. It soon became an epidemic that, in total, infected ~ 2 million cattle in the U.K. BSE is surmised to have arisen as a consequence of feeding cattle meat-and-bone meal made from scrapie-infected sheep (and eventually from BSE-infected cattle). BSE, which has an ~ 5 -year incubation period, was unknown before 1985, most likely because the process for manufacturing meat-and-bone meal was changed in the late 1970s from a way that fully inactivates scrapie prions to one that fails to do so. In 1988, the U.K. banned the feeding of ruminants with ruminant-derived protein (other than milk), so that, following its peak in 1993, the BSE epidemic rapidly abated (a process accelerated by the slaughter of large numbers of cattle at risk of having BSE). However, since humans consumed meat from BSE-infected cattle for over a decade, the question remained, had BSE been transmitted to humans? It

should be noted that 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) is no greater than that in largely vegetarian countries such as India. Nevertheless, in 1994, several cases of CJD in teenagers and young adults were reported in the U.K., although heretofore CJD before the age of 40 was extremely rare (its average age of onset is ~ 64). Individuals with this **new variant CJD (vCJD or nvCJD)**, of which there have been ~ 200 cases yet reported, almost entirely in the U.K., have neurological symptoms and neuropathology that are atypical for sporadic CJD. Moreover, when transmitted to mice expressing bovine PrP^C, vCJD has an incubation time, neurological symptoms, and neuropathology indistinguishable from that caused by BSE. It therefore seems highly likely that vCJD is caused by a prion strain that humans acquired by eating meat products from BSE-infected cattle.

g. Prions Occur in Yeast

Although prions were originally defined to be scrapielike infectious pathogens, it is now evident that this definition must be broadened to include all proteins with stable conformational variants that catalyze their own formation from “wild-type” protein. For example, *Saccharomyces cerevisiae* (baker’s yeast) can harbor a genetic element designated [URE3] that, in sexual reproduction with cells that lack [URE3], is inherited by all progeny rather than according to the rules of Mendelian genetics (Section 1-4B). Yet [URE3] is a chromosomal gene rather than a plasmid-based or mitochondrial gene (which would account for its non-Mendelian inheritance).

[URE3] is identical to the chromosomal gene *URE2*, which specifies a protein, **Ure2**, that in the presence of yeast’s preferred nitrogen sources (ammonia or glutamine) represses the expression of the proteins required to metabolize yeast’s less preferred nitrogen sources (e.g., proline). Yeast that have the [URE3] phenotype (trait) lack this regulation of nitrogen metabolism (nitrogen metabolism is discussed in Chapter 26). However, [URE3] yeast can be “cured” of this condition by treatment with 5 mM guanidinium chloride; that is, they and their progeny then exhibit normal regulation of nitrogen metabolism. Nevertheless, about one yeast cell per million spontaneously reverts to the [URE3] phenotype. This is because Ure2 has a “wild-type” conformational state, which regulates nitrogen metabolism, and a [URE3] form, which catalyzes its own formation from “wild-type” Ure2 to yield amyloid fibers that do not influence nitrogen metabolism. Thus, *Ure2 is a type of prion.*


The yeast genetic element [PSI] encodes a protein, **Sup35**, with similar prionlike properties that participates in transcriptional termination (Section 32-3E). Indeed, the introduction of Sup35 in its [PSI] conformation into the cytoplasm of yeast containing “wild-type” Sup35 induced the formation of the [PSI] phenotype, an experiment that constituted the first direct evidence supporting the prion hypothesis. Moreover, Sup35 can adopt several different fiber conformations *in vitro*, which when introduced into [psi⁻]

cells, produce clearly distinguishable strain variants. Several other fungal proteins that can form prions have also been characterized.

6 STRUCTURAL EVOLUTION

Proteins, as we discussed in Section 7-3, evolve through point mutations and gene duplications. Over eons, through processes of natural selection and/or neutral drift, homologous proteins thereby diverge in character and develop new functions. How these primary structure changes affect function, of course, depends on the protein's three-dimensional structure. In this section, we explore the effects of evolutionary change on protein structures.

A. Structures of Cytochromes *c*

 **See Guided Exploration 10: Protein evolution** The *c*-type cytochromes are small globular proteins that contain a covalently bound heme group (**iron-protoporphyrin IX**; Fig. 9-39). The X-ray structures of the cytochromes *c* from horse (Fig. 8-42), tuna, bonito, rice, and yeast are closely similar and thus permit the structural significance of cytochrome *c*'s amino acid sequences (Section 7-3B) to be assessed. The internal residues of cytochrome *c*, particularly those lining its heme pocket, tend to be invariant or conservatively substituted, whereas surface positions have greater variability. This observation is, in part, an indication of the more exacting packing requirements of a protein's internal regions compared to those of its surface (Section 8-3Bc).

Certain invariant or highly conserved residues (Table 7-4) have specific structural and/or functional roles in cytochrome *c*:

1. The invariant Cys 14, Cys 17, His 18, and Met 80 residues form covalent bonds with the heme group (Fig. 9-39).
2. The nine invariant or highly conserved Gly residues occupy close-fitting positions in which larger side chains would significantly alter the protein's three-dimensional structure.
3. The highly conserved Lys residues 8, 13, 25, 27, 72, 73, 79, 86, and 87 are distributed in a ring around the exposed edge of the otherwise buried heme group. There is considerable evidence that this unusual constellation of positive charges specifically associates with complementary sets of negative charges on the physiological reaction partners of cytochrome *c*, cytochrome *c* reductase, and cytochrome *c* oxidase.

a. Prokaryotic *c*-Type Cytochromes Are Structurally Related to Cytochrome *c*

Although cytochrome *c* occurs only in eukaryotes, similar proteins known as ***c*-type cytochromes** are common in prokaryotes, where they function to transfer electrons at analogous positions in a variety of respiratory and photosynthetic electron-transport chains. Unlike the eukaryotic proteins, however, the prokaryotic *c*-type cytochromes

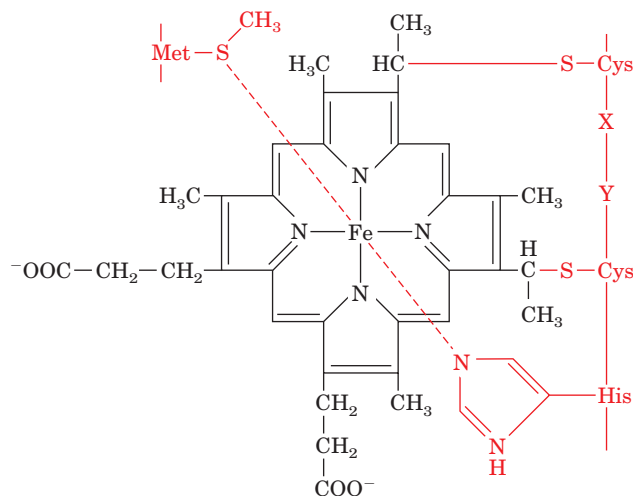


Figure 9-39 Molecular formula of iron-protoporphyrin IX (heme). In *c*-type cytochromes, the heme is covalently bound to the protein (red) by two thioether bonds linking what were the heme vinyl groups to two Cys residues that occur in the sequence Cys-X-Y-Cys-His (residues 14–18 in Table 7-4). Here X and Y symbolize any amino acid residues. A fifth and sixth ligand to the Fe atom, both normal to the heme plane, are formed by a side chain nitrogen of His 18 and the sulfur of Met 80. The iron atom, which is thereby octahedrally liganded, can stably assume either the Fe(II) or the Fe(III) oxidation state. Heme also occurs in myoglobin and hemoglobin but without the thioether bonds or the Met ligand.

exhibit considerable sequence variability among species. For example, the numerous bacterial *c*-type cytochromes whose primary structures are known have from 82 to 134 amino acid residues, whereas eukaryotic cytochromes *c* have a narrower range, from 103 to 112 residues. The primary structures of several representative *c*-type cytochromes have few obvious similarities (Fig. 9-40). Yet their X-ray structures closely resemble each other, particularly in their backbone conformations and side chain packing in the regions surrounding the heme group (Fig. 9-41). Furthermore, most of them have aromatic rings in analogous positions and orientations relative to their heme groups as well as similar distributions of positively charged Lys residues about the perimeters of their heme crevices. The major structural differences among these *c*-type cytochromes stem from various loops of polypeptide chain that are located on their surfaces.

Before the advent of sophisticated sequence alignment algorithms such as BLAST (Section 7-4Bg), the correct alignments of analogous *c*-type cytochrome residues (thin lines in Fig. 9-40) could not have been made on the basis of only their primary structures: These proteins have diverged so far that their three-dimensional structures were essential guides for this task. Three-dimensional structures are evidently more indicative of the similarities among these distantly related proteins than are primary structures. *It is the essential structural and functional elements of proteins, rather than their amino acid residues, that are conserved during evolutionary change.*

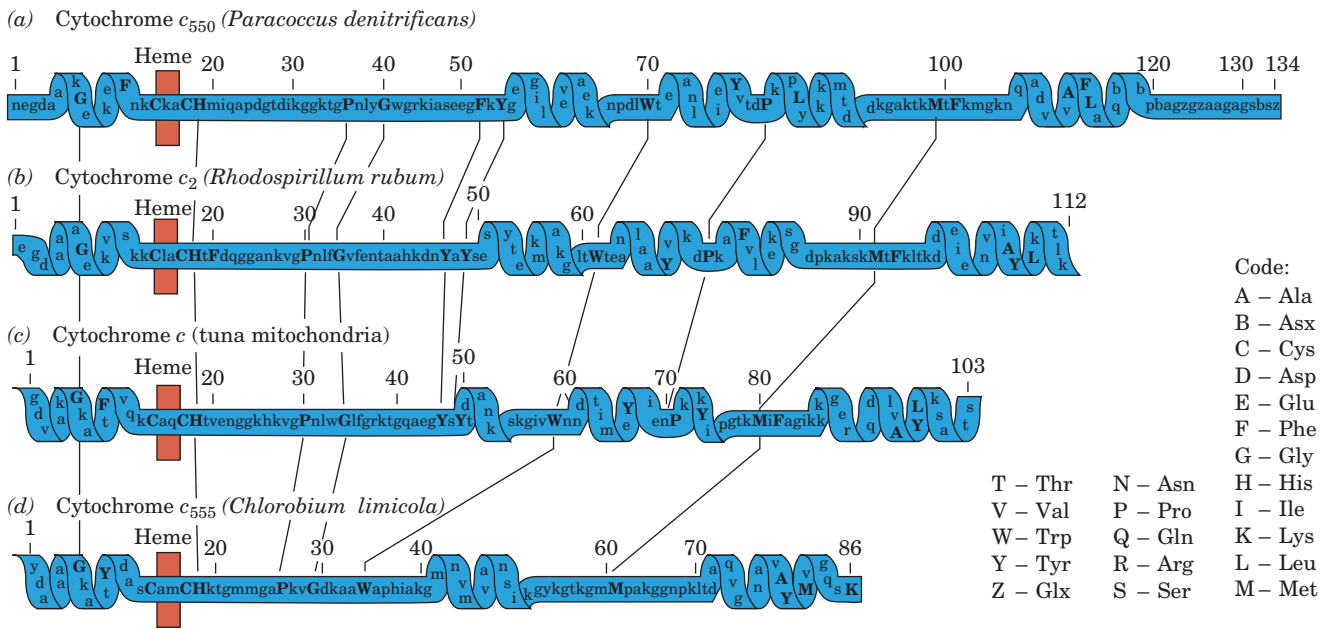


Figure 9-40 Primary structures of some representative c -type cytochromes. (a) Cytochrome c_{550} (the subscript indicates the protein's peak absorption wavelength in visible light, in nm) from *Paracoccus denitrificans*, a respiring bacterium that can use nitrate as an oxidant. (b) Cytochrome c_2 (the subscript has only historical significance) from *Rhodospirillum rubrum*, a purple photosynthetic bacterium. (c) Cytochrome c from tuna

mitochondria. (d) Cytochrome c_{555} from *Chlorobium limicola*, a green photosynthetic bacterium that utilizes H_2S as a hydrogen source. Thin lines connect structurally significant or otherwise invariant residues (*uppercase*). Helical regions are indicated to facilitate structural comparisons with Fig. 9-41. [After Salemme, F.R., *Annu. Rev. Biochem.* **46**, 307 (1977).]

B. Gene Duplication

Gene duplication may promote the evolution of new functions through structural evolution (Section 7-3C). In over half of the multidomain proteins of known structure, two or more of the domains are structurally quite similar. Consider, for example, the four domains of yeast protein disulfide isomerase (PDI; Fig. 9-17). It seems highly unlikely that these complex but topologically similar domains could have independently evolved their present structures—a process

known as **convergent evolution**. Almost certainly, they arose through duplications of the gene specifying an ancestral domain accompanied by the fusion of the resulting four genes to yield a single gene specifying a polypeptide that folds into four similar domains. The differences between the four domains are therefore due to their **divergent evolution**.

Structurally similar domains often occur in proteins whose other domains bear no resemblance to one another.

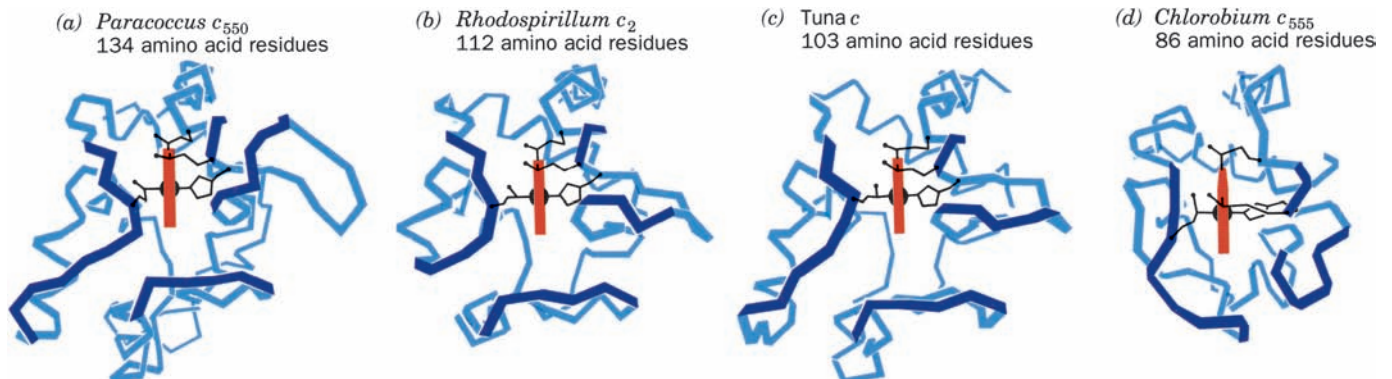



Figure 9-41 Three-dimensional structures of the c -type cytochromes whose primary structures are displayed in Fig. 9-40. The polypeptide backbones (*blue*) are shown in analogous orientations such that their heme groups (*red*) are viewed edge on. The Cys, Met, and His side chains that covalently link the heme to

the protein are also shown. (a) Cytochrome c_{550} from *P. denitrificans*. (b) Cytochrome c_2 from *Rs. rubrum*. (c) Tuna cytochrome c . (d) Cytochrome c_{555} from *C. limicola*. [Illustration, Irving Geis. Image from the Irving Geis Collection, Howard Hughes Medical Institute. Reprinted with permission.]  See Kinemage Exercise 5-1

The redox enzymes known as **dehydrogenases**, for example, each consist of two domains: a domain that binds redox-active dinucleotides such as NAD^+ and which is structurally similar in all the dehydrogenases, and a dissimilar substrate-binding domain that determines the specificity and mode of action of each enzyme. Indeed, in some dehydrogenases, such as glyceraldehyde-3-phosphate dehydrogenase (Fig. 8-45), the dinucleotide-binding domain occurs at the N-terminal end of the polypeptide chain, whereas in others it occurs at the C-terminal end. Each of these dehydrogenases must have arisen by the fusion of the

gene specifying an ancestral dinucleotide-binding domain with a gene encoding a proto-substrate-binding domain. This must have happened very early in evolutionary history, perhaps in the precellular stage (Section 1-5Ca), because there are no significant sequence similarities among these dinucleotide-binding domains. Evidently, *a domain is as much a unit of evolution as it is a unit of structure. By genetically combining these structural modules in various ways, nature can develop new functions far more rapidly than it can do so by the evolution of completely new structures through point mutations.*

CHAPTER SUMMARY

1 Protein Folding: Theory and Experiment Under renaturing conditions, many proteins fold to their native structures in a matter of seconds. Helices and sheets, which together constitute ~60% of the average protein, are so common because they efficiently fill space. Proteins are hierarchically organized, that is, they consist of domains, which consist of subdomains, etc. They are highly tolerant of sequence changes, to which they adapt by local rather than global structural alterations. Some proteins are natively unfolded, although they assume stable structures when binding to their target molecules.

The rapidity with which proteins renature indicates that they fold in an ordered manner rather than via a random search of all their possible conformations. Thus the study of protein folding requires rapid mixing and observational techniques such as stopped-flow devices, circular dichroism (CD), pulsed H/D exchange followed by NMR, and fluorescence resonance energy transfer (FRET). The folding of small single-domain proteins is initiated by a hydrophobic collapse to yield a molten globule, which appears within ~5 ms. This is followed by the stabilization of secondary structure and then the formation of tertiary structure to yield the native protein in a matter of several seconds. Folding is thought to follow landscape theory, which postulates that a polypeptide folds via a folding funnel and hence can take any of a great variety of pathways to reach its native state. This is consistent with the finding that proteins fold in a hierarchical manner. The sequence of a protein appears to specify its folding pathway as well as its native structure.

2 Folding Accessory Proteins Even though it is clear that a protein's primary structure dictates its three-dimensional structure, many proteins require the assistance of accessory proteins such as protein disulfide isomerase (PDI), peptidyl prolyl cis–trans isomerases, and molecular chaperones to fold/assemble to their native structures. PDI consists of four thioredoxinlike domains, two of which contain exposed Cys residues that form disulfide bonds, either internally or with another protein in a disulfide interchange reaction. Two families of peptidyl prolyl cis–trans isomerases have been characterized, the cyclophilins, which bind cyclosporin A, and FK506 binding protein, which binds FK506.

The chaperonins, such as GroEL and GroES, stimulate the proper folding of certain misfolded proteins through a cyclic sequence of concerted conformational changes that is driven by the binding and hydrolysis of ATP. GroES is a cap-shaped heptamer

and GroEL is a 14-mer arranged in two apposed heptameric rings that form two unconnected end-to-end hollow barrels. Together, GroEL and GroES form a bullet-shaped complex that contains a closed cavity (an Anfinsen cage) in which misfolded proteins can fold without interference by aggregation with other misfolded proteins. In doing so, GroEL/ES partially unfolds a misfolded and conformationally trapped protein of up to ~70 kD and releases it so as to enable it to travel down its folding funnel via a new route. Such proteins undergo an average of 14 cycles of binding and release before achieving their native folds. Many of the ~85 *E. coli* proteins that stringently require the GroEL/ES system for proper folding contain α/β domains, whose structural complexity is largely responsible for their misfolding. Eukaryotic Group II chaperonins have built in lids that appear to function analogously to GroES.

3 Protein Structure Prediction and Design The prediction of protein secondary structures from only amino acid sequences has been reasonably successful using empirical techniques such as the Chou–Fasman method. However, sophisticated computational techniques yield more reliable predictions. Comparative (homology) modeling can provide accurate tertiary structures for polypeptides with >30% identity to a protein of known structure. Fold recognition (threading) techniques have only been marginally successful for determining the structures of proteins that have no apparent homology with proteins of known structure. (Modern) *de novo* structure determination methods yield correct folding topologies with a success rate of ~20% and occasionally reasonably accurate atomic models. The reverse process, computationally based protein design, has been more successful, in part, because one can “overengineer” a protein to take up a desired conformation.

4 Protein Dynamics Proteins are flexible and fluctuating molecules whose group motions have characteristic periods ranging from 10^{-15} to over 10^3 s. X-ray analysis, which reveals the average atomic mobilities in a protein, indicates that proteins tend to be more mobile at their peripheries than in their interiors. Molecular dynamics simulations indicate that native protein structures each consist of a large number of closely related and rapidly interconverting conformational substates of nearly equal stabilities. Without this flexibility, enzymes would be nonfunctional. The rates of aromatic ring flipping, as revealed by NMR measurements, indicate that internal group mobilities within proteins vary both with the protein and with

the position within the protein. The exchange of a protein's internal protons with solvent requires its transient local unfolding. Hydrogen exchange studies therefore demonstrate that proteins have a great variety of infrequently occurring internal motions.

5 Conformational Diseases: Amyloids and Prions A number of often fatal human diseases are associated with the deposition of amyloid in the brain and other organs. Although the various amyloidogenic proteins are unrelated in both sequence and native structure, all form similar amyloid fibrils that consist mainly of β sheets whose planes extend along the fibril axis. The two known human lysozyme variants that have amyloidogenic properties are conformationally much looser than wild-type lysozyme. In Alzheimer's disease, a neurodegenerative disease of mainly the elderly, the proteolysis of A β -precursor protein (β PP) in brain tissue yields the 40- to 42-residue amyloid- β protein (A β), which forms the amyloid fibrils that kill neurons.

Humans and other mammals are subject to infectious neurodegenerative diseases such as scrapie, which are caused by prions. Prions appear to consist of only a single species of protein named PrP. PrP exists in two forms: the normal cellular form, PrP^C, a conserved membrane-anchored cell-surface protein on neurons; and PrP^{Sc}, which although chemically identical to PrP^C, has a different conformation. PrP^{Sc} autocatalytically converts PrP^C to PrP^{Sc} thereby accounting for the infectious properties of

PrP^{Sc} and the observation that *Prn-p*^{0/0} mice are resistant to scrapie. PrP^{Sc} is proteolytically degraded in the cell to form a protease-resistant core, PrP 27–30, that aggregates to form the neurotoxic amyloid fibrils thought to be responsible for the symptoms of prion diseases. A given species of PrP may take up several different self-propagating fibrous conformations to yield different prion strains. Fungi, such as yeast, also have proteins with prionlike properties.

6 Structural Evolution The X-ray structures of eukaryotic cytochromes *c* demonstrate that internal residues and those having specific structural and functional roles tend to be conserved during evolution. Prokaryotic *c*-type cytochromes from a variety of organisms structurally resemble each other and those of eukaryotes even though they have little sequence similarity. This indicates that the three-dimensional structures of proteins rather than their amino acid sequences are conserved during evolutionary change. The structural similarities between the domains in many multidomain proteins indicate that these proteins arose through the duplication of the genes specifying the ancestral domains followed by their fusion. For example, the structural resemblance between the dinucleotide-binding domains of dehydrogenases suggests that these proteins arose by duplication of a primordial dinucleotide-binding domain followed by its fusion with a gene specifying a proto-substrate-binding domain. In this manner, proteins with new functions can evolve much faster than by a series of point mutations.

REFERENCES

Protein Folding

- Anfinsen, C.B., Principles that govern the folding of protein chains, *Science* **181**, 223–230 (1973). [A Nobel laureate explains how he got his prize.]
- Aurora, R. and Rose, G.D., Helix capping, *Protein Sci.* **7**, 21–38 (1998). [Summarizes the evidence that helix capping interactions stabilize helices.]
- Baldwin, R.L., Pulsed H/D-exchange studies of folding intermediates, *Curr. Opin. Struct. Biol.* **3**, 84–91 (1993).
- Baldwin, R.L., Protein folding from 1961 to 1982, *Nature Struct. Biol.* **6**, 814–817 (1999). [An intellectual history.]
- Baldwin, R.L. and Rose, G.D., Is protein folding hierarchic? I. Local structure and peptide folding; and II. Folding intermediates and transition states, *Trends Biochem. Sci.* **24**, 26–33; and 77–83 (1999).
- Behe, M., Lattman, E.E., and Rose, G.D., The protein folding problem: The native fold determines the packing but does packing determine the native fold? *Proc. Natl. Acad. Sci.* **88**, 4195–4199 (1991).
- Betts, S. and King, J., There's a right way and a wrong way: *in vivo* and *in vitro* folding, misfolding and subunit assembly of the P22 tailspike, *Structure* **7**, R131–R139 (1999).
- Buchner, J. and Kiefhaber, T., *Protein Folding Handbook*, Wiley-VCH (2005). [An authoritative 5-volume work on most aspects of protein folding.]
- Dalal, S., Balasubramanian, S., and Regan, L., Protein alchemy: Changing β -sheet into α -helix, *Nature Struct. Biol.* **4**, 548–552 (1997). [Reports the sequence changes in protein GB1 that cause it to assume the fold of Rop protein.]
- Dill, K.A. and Chan, H.S., From Levinthal to pathways to funnels, *Nature Struct. Biol.* **4**, 10–19 (1997). [Reviews the landscape theory of protein folding.]
- Dill, K.A., Ozkan, S.B., Shell, M.S., and Weikl, T.R., The protein folding problem, *Annu. Rev. Biophys.* **37**, 289–316 (2008).
- Dunker, A.K., Silman, I., Uversky, V.N., and Sussman, J.L., Function and structure of inherently disordered proteins, *Curr. Opin. Struct. Biol.* **18**, 756–764 (2008).
- Dyson, H.J. and Wright, P.E., Intrinsically unstructured proteins and their functions, *Nature. Rev. Mol. Cell. Biol.* **6**, 197–208 (2005).
- Englander, S.W., Protein folding intermediates and pathways studied by hydrogen exchange, *Annu. Rev. Biophys. Biomol. Struct.* **29**, 213–238 (2000).
- Englander, W.S., Mayne, L., and Krishna, M.M.G., Protein folding and misfolding: mechanism and principles, *Q. Rev. Biophys.* **40**, 287–326 (2007).
- Fersht, A., *Structure and Mechanism in Protein Science*, Chapters 17–19, Freeman (1999).
- Fink, A.L., Natively unfolded proteins, *Curr. Opin. Struct. Biol.* **15**, 35–41 (2005).
- Fitzkee, N.C., Fleming, P.J., Gong, H., Panasik, N., Jr., Street, T.O., and Rose, G.D., Are proteins made from a limited parts list? *Trends Biochem. Sci.* **30**, 73–80 (2005).
- Gillespie, B. and Plaxco, K.W., Using protein folding rates to test protein folding theories, *Annu. Rev. Biochem.* **73**, 837–859 (2004).
- Kubelka, J., Hofrichter, J., and Eaton, W.A., The protein folding 'speed limit,' *Curr. Opin. Struct. Biol.* **14**, 76–88 (2004).
- Matthews, B.W., Studies on protein stability with T4 lysozyme, *Adv. Prot. Chem.* **46**, 249–278 (1995).
- Meyers, R.A., *Proteins. From Analytics to Structural Genomics*, Vol. 1, Chapters 1, 2, and 4, Wiley-VCH (2007). [Discusses X-ray crystallography, NMR spectroscopy, and circular dichroism of proteins.]

- Minor, D.L., Jr. and Kim, P.S., Context-dependent secondary structure formation of a designed protein sequence, *Nature* **380**, 730–734 (1996). [Describes the position-dependent conformation of the chameleon sequence in protein GB1.]
- Oliveberg, A. and Wolynes, P.G., The experimental survey of protein-folding energy landscapes, *Q. Rev. Biophys.* **36**, 245–288 (2006).
- Onuchic, J.N. and Wolynes, P.G., Theory of protein folding, *Curr. Opin. Struct. Biol.* **14**, 70–75 (2004).
- Pain, R.H. (Ed.), *Mechanisms of Protein Folding* (2nd ed.), Oxford University Press (2000).
- Piston, D.W. and Kremers, G.-J., Fluorescent protein FRET: the good, the bad and the ugly, *Trends Biochem. Sci.* **32**, 407–414 (2007).
- Roder, H. and Shastry, M.C.R., Methods for exploring early events in protein folding, *Curr. Opin. Struct. Biol.* **9**, 620–626 (1999).
- Udgaonkar, J.B., Multiple routes and structural heterogeneity in protein folding, *Annu. Rev. Biophys.* **37**, 489–510 (2008).
- Wang, C.C. and Tsou, C.L., The insulin A and B chains contain sufficient structural information to form the native molecule, *Trends Biochem. Sci.* **16**, 279–281 (1991).
- Folding Accessory Proteins**
- Booth, C.R., Meyer, A.S., Cong, Y., Topf, M., Sali, A., Ludtke, S.J., Chiu, W., and Frydman, J., Mechanism of lid closure in the eukaryotic chaperonin TRiC/CCT, *Nature Struct. Biol.* **15**, 746–753 (2008).
- Chen, L. and Sigler, P.B., The crystal structure of a GroEL/peptide complex: Plasticity as a basis for substrate diversity, *Cell* **99**, 757–768 (1999).
- Clark, P.L., Protein folding in the cell: reshaping the folding funnel, *Trends Biochem. Sci.* **29**, 527–534 (2004).
- Ellis, R.J., Macromolecular crowding: Obvious but underappreciated, *Trends Biochem. Sci.* **26**, 597–604 (2001).
- Ellis, R.J., Molecular chaperones: assisting assembly in addition to folding, *Trends Biochem. Sci.* **31**, 395–401 (2006).
- Frydman, J., Folding of newly translated proteins *in vivo*: The role of molecular chaperones, *Annu. Rev. Biochem.* **70**, 603–649 (2001).
- Gruber, C.W., Cemazar, M., Heras, B., Martin, J.L., and Craik, D.J., Protein disulfide isomerase: the structure of oxidative folding, *Trends Biochem. Sci.* **31**, 455–464 (2006).
- Hartl, F.U., and Hayer-Hartl, M., Molecular chaperones in the cytosol: From nascent chain to unfolded protein, *Science* **295**, 1852–1858 (2002).
- Horst, R., Bertelson, E.B., Fiaux, J., Wider, G., Horwich, A.L., and Wüthrich, K., Direct NMR observation of a substrate protein bound to the chaperonin GroEL, *Proc. Natl. Acad. Sci.* **102**, 12748–12753 (2005); and Horst, R., Fenton, W.A., Englander, S.W., Wüthrich, K., and Horwich, A.L., Folding trajectories of human dihydrofolate reductase inside the GroEL–GroES chaperonin cavity and free in solution, *Proc. Natl. Acad. Sci.* **104**, 20788–20792 (2007).
- Horwich, A.R. (Ed.), Protein Folding in the Cell, *Adv. Prot. Chem.* **59** (2002). [Contains authoritative articles on a variety of folding accessory proteins.]
- Horwich, A.R., Farr, G.W., and Fenton, W.A., GroEL–GroES-mediated protein folding, *Chem. Rev.* **106**, 1917–1930 (2006); and Horwich, A.R., Fenton, W.A., Chapman, E., and Farr, G.W., Two families of chaperonin: Physiology and mechanism, *Annu. Rev. Cell Dev. Biol.* **23**, 115–145 (2007).
- Kerner, M.J., et al., Proteome-wide analysis of chaperone-dependent protein folding in *Escherichia coli*, *Cell* **122**, 209–220 (2005).
- Lin, Z. and Rye, H.S., GroEL-mediated protein folding: Making the impossible, possible, *Crit. Rev. Biochem. Mol. Biol.* **41**, 211–239 (2006).
- Mamathambika, B.S., and Bardwell, J.C., Disulfide-linked protein folding pathways, *Annu. Rev. Cell Dev. Biol.* **24**, 211–235 (2008).
- Morano, K.A., New tricks for an old dog. The evolving world of Hsp70, *Ann. N.Y. Acad. Sci.* **1113**, 1–14 (2007).
- Ransom, N.A., Farr, G.W., Roseman, A.M., Gowen, B., Fenton, W.A., Horwich, A.L., and Saibil, H.R., ATP-bound states of GroEL captured by cryo-electron microscopy, *Cell* **107**, 869–879 (2001).
- Saibil, H.R., Chaperone machines in action, *Curr. Opin. Struct. Biol.* **18**, 35–42 (2008).
- Schiene, C. and Fischer, G., Enzymes that catalyse the restructuring of proteins, *Curr. Opin. Struct. Biol.* **10**, 40–45 (2000). [Discuss protein disulfide isomerases and peptidyl prolyl cis–trans isomerases.]
- Schreiber, S.L., Chemistry and biology of immunophilins and their immunosuppressive ligands, *Science* **251**, 238–287 (1991).
- Sharma, S., Chakraborty, K., Müller, B.K., Astola, N., Tang, Y.-C., Lamb, D.C., Hayer-Hartl, M., and Hartl, F.U., Monitoring protein conformation along the pathway of chaperon-assisted folding, *Cell* **133**, 142–153 (2008).
- Shtilerman, M., Lorimer, G.H., and Englander, S.W., Chaperonin function: Folding by forced unfolding, *Science* **284**, 822–825 (1999).
- Spiess, C., Meyer, A.S., Reissmann, S., and Frydman, J., Mechanism of the eukaryotic chaperonin: protein folding in the chamber of secrets, *Trends Cell Biol.* **14**, 598–604 (2004).
- Stan, G., Brooks, B.R., Lorimer, G.H., and Thirumalai, D., Residues in substrate proteins that interact with GroEL in the capture process are buried in the native state, *Proc. Natl. Acad. Sci.* **103**, 4433–4438 (2006).
- Thirumalai, D. and Lorimer, G.H., Chaperone-mediated protein folding, *Annu. Rev. Biophys. Biomol. Struct.* **30**, 245–269 (2001).
- Tian, G., Xiang, S., Noiva, R., Lennarz, W.J., and Schindelin, H., The crystal structure of yeast protein disulfide isomerase suggests cooperativity between its active sites, *Cell* **124**, 61–73 (2006).
- Wandinger, S.K., Richter, K., and Buchner, J., The Hsp90 chaperone machinery, *J. Biol. Chem.* **283**, 18473–18477 (2008); and Pearl, L.H. and Prodromou, C., Structure and mechanism of the Hsp90 molecular chaperone machinery, *Annu. Rev. Biochem.* **75**, 271–294 (2006).
- Xu, Z., Horwich, A.L., and Sigler, P.B., The crystal structure of the asymmetric GroEL–GroES–(ADP)₇ chaperonin complex, *Nature* **388**, 741–750 (1997).
- Zhao, Y. and Ke, H., Crystal structure implies that cyclophilin predominantly catalyzes the *trans* to *cis* isomerization, *Biochemistry* **35**, 7356–7361 (1996).
- Protein Structure Prediction and Design**
- Baxeavanis, A.D. and Ouellette, B.F.F. (Eds.), *Bioinformatics. A Practical Guide to the Analysis of Genes and Proteins* (3rd ed.), Chapters 8 and 9, Wiley-Interscience (2005).
- Blaber, M., Zhang, X., and Matthews, B.W., Structural basis of amino acid α helix propensity, *Science* **260**, 1637–1640 (1993).
- Bujnicki, J.M. (Ed.), *Prediction of Protein Structures, Functions, and Interactions*, Wiley (2009).
- Chou, P.Y. and Fasman, G.D., Empirical predictions of protein structure, *Annu. Rev. Biochem.* **47**, 251–276 (1978). [Exposition of a particularly simple method of protein secondary structure prediction.]

- Cuff, J.A. and Barton, G.J., Evaluation and improvement of multiple sequence methods for protein secondary structure prediction, *Proteins* **34**, 508–519 (1999). [The principles behind Jpred3.]
- Das, R. and Baker, D., Macromolecular modeling with Rosetta, *Annu. Rev. Biochem.* **77**, 363–382 (2008).
- DeGrado W.F., Summa, S.M., Pavone, V., Nastro, F., and Lombardi, A., De novo design and structural characterization of proteins and metalloproteins, *Annu. Rev. Biochem.* **68**, 779–819 (1999).
- Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., and Baker, D., Design of a novel protein fold with atomic-level accuracy, *Science* **302**, 1364–1368 (2003). [The design of Top7.]
- Lesk, A.M., *Introduction to Bioinformatics* (3rd ed.), pp. 333–358, Oxford University Press (2008).
- Mirny, L. and Shakhnovitch, E., Protein folding theory: From lattice to all-atom models, *Annu. Rev. Biophys. Biomol. Struct.* **30**, 361–396 (2001).
- Moult, J., Fidelis, K., Kryshchovych, A., Rost, B., and Tramontano, A., Critical assessment of methods of protein structure prediction—Round VIII, *Proteins* **77** (Issue S9), 1–4 (2009). [The summary article of the issue of *Proteins: Structure, Function, and Bioinformatics* that reports the results of CASP8.]
- Rose, G.D., Prediction of chain turns in globular proteins on a hydrophobic basis, *Nature* **272**, 586–590 (1978).
- Tramontano, A., *Protein Structure Prediction. Concepts and Applications*, Wiley-VCH (2006).
- Zaki, M.J. and Bystroff, C. (Eds.), *Protein Structure Prediction* (2nd ed.), Humana Press (2008).
- Protein Dynamics**
- Henzler-Wildman, K. and Kern, D., Dynamic personalities of proteins, *Nature* **450**, 964–972 (2007).
- Karplus, M. and McCammon, A., Molecular dynamics simulations of biomolecules, *Nature Struct. Biol.* **9**, 646–651 (2002).
- Palmer, A.G., III, Probing molecular motion by NMR, *Curr. Opin. Struct. Biol.* **7**, 732–737 (1997).
- Protein dynamics, *Science* **324**, 197–215 (2009). [A special section containing four articles.]
- Rasmussen, B.F., Stock, A.M., Ringe, D., and Petsko, G.A., Crystalline ribonuclease A loses function below the dynamical transition at 220 K, *Nature* **357**, 423–424 (1992).
- Ringe, D. and Petsko, G.A., Mapping protein dynamics by X-ray diffraction, *Prog. Biophys. Mol. Biol.* **45**, 197–235 (1985).
- Scheraga, H.A., Khalili, M., and Liwo, A., Protein-folding dynamics: Overview of molecular simulation techniques, *Annu. Rev. Phys. Chem.* **58**, 57–83 (2007).
- Conformational Diseases**
- Booth, D.R., et al., Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis, *Nature* **385**, 787–793 (1997); and Funahashi, J., Takano, K., Ogashira, K., Yamagata, Y., and Yutani, K., The structure, stability, and folding process of amyloidogenic mutant lysozyme, *J. Biochem.* **120**, 1216–1223 (1996).
- Büeler, H., Aguzzi, A., Sailer, A., Greiner, R.A., Autenreid, P., Aguet, M., and Weissmann, C., Mice devoid of PrP are resistant to scrapie, *Cell* **73**, 1339–1347 (1993); and Büeler, H., Fischer, M., Lang, Y., Bluethmann, H., Lipp, H.-P., DeArmond, S.J., Prusiner, S.B., Aguet, M., and Weissmann, C., Normal development and behaviour of mice lacking the neuronal cell-surface PrP protein, *Nature* **356**, 577–582 (1992).
- Buxbaum, J.N. and Tagoe, C.E., The genetics of amyloidoses, *Annu. Rev. Med.* **51**, 543–569 (2000).
- Caughey, B., Baron, G.S., Chesebro, B., and Jeffrey, M., Getting a grip on prions: oligomers, amyloids, and pathological membrane interactions, *Annu. Rev. Biochem.* **78**, 177–204 (2009).
- Chien, P., Weissman, J.S., and DePace, J.H., Emerging principles of conformation-based inheritance, *Annu. Rev. Biochem.* **73**, 617–656 (2004).
- Chiti, F. and Dobson, C.M., Protein misfolding, functional amyloid, and human disease, *Annu. Rev. Biochem.* **75**, 333–366 (2006).
- Collinge, J. and Clarke, A.R., A general model of prion strains and their pathogenicity, *Science* **318**, 930–936 (2007).
- Deleault, N.R., Harris, B.T., Rees, J.R., and Supattapone, S., Formation of native prions from minimal components *in vitro*, *Proc. Natl. Acad. Sci.* **104**, 9741–9746 (2007).
- Geula, C., Wu, C.-K., Saroff, D., Lorenzo, A., Yuan, M., and Yankner, B.A., Aging renders the brain vulnerable to amyloid β -protein neurotoxicity, *Nature Med.* **4**, 827–831 (1998).
- Goedert, M. and Spillantini, M.G., A century of Alzheimer's disease, *Science* **314**, 777–781 (2006).
- Gregersen, N., Bross, P., Vang, S., and Christensen, J.H., Protein misfolding and human disease, *Annu. Rev. Genomics Hum. Genet.* **7**, 103–124 (2006).
- Hardy, J. and Selkoe, D.J., The amyloid hypothesis of Alzheimer's disease: Progress and problems on the road to therapeutics, *Science* **297**, 353–356 (2002).
- Jackson, G.S. and Clarke, A.R., Mammalian prion proteins, *Curr. Opin. Struct. Biol.* **10**, 69–74 (2000).
- Kajava, A., Squire, J.M., and Parry, D.A.D. (Eds.), *Fibrous Proteins: Amyloids, Prions and Beta Proteins*, *Adv. Prot. Chem.* **73** (2006). [The last four chapters are on various aspects of amyloids and prions.]
- Moore, R.A., Taubner, L.M., and Priola, S.A., Prion misfolding and disease, *Curr. Opin. Struct. Biol.* **19**, 14–22 (2009).
- Pan, K.M., Baldwin, M., Nguyen, J., Gasset, M., Serban, A., Groth, D., Mehlhorn, I., Huang, Z., Fletterick, R.J., Cohen, F.E., and Prusiner, S.B., Conversion of α -helices into β -sheet features in the formation of the scrapie prion proteins, *Proc. Natl. Acad. Sci.* **90**, 10962–10966 (1993).
- Prusiner, S.B. (Ed.), *Prion Biology and Diseases* (2nd ed.), Cold Spring Harbor Laboratory Press (2004); and Prion diseases, *in* Valle, D. (Ed.), *The Online Metabolic & Molecular Bases of Inherited Disease*, <http://www.ommbid.com/>, Chap. 224.
- Rochet, J.C. and Lansbury, P.T., Jr., Amyloid fibrillogenesis: Themes and variations, *Curr. Opin. Struct. Biol.* **10**, 60–68 (2000).
- Sawaya, M.R., et al., Atomic structures of amyloid cross- β spines reveal varied steric zippers, *Nature* **447**, 453–457 (2007); and Nelson, R., Sawaya, M.R., Balbirnie, M., Madsen, A.Ø., Riek, C., Grothe, R., and Eisenberg, D., Structure of the cross- β spine of amyloid-like fibrils, *Nature* **435**, 773–778 (2005).
- Selkoe, D.J., Cell biology of protein misfolding: the examples of Alzheimer's and Parkinson's diseases, *Nature Cell Biol.* **6**, 1054–1061 (2004).
- Soto, C., Estrada, L., and Castilla, J., Amyloids, prions and the inherent infectious nature of misfolded proteins, *Trends Biochem. Sci.* **31**, 150–155 (2006).
- Sparrer, H.E., Santoso, A., Szoka, F.C., Jr., and Weissman, J.S., Evidence for the prion hypothesis: Induction of the yeast [PSI⁺] factor by *in vitro*-converted Sup35 protein, *Science* **289**, 595–599 (2000).
- Tuite, M.F., Yeast prions and their prion-forming domain, *Cell* **100**, 289–292 (2000).
- Weissmann, C., The state of the prion, *Nature Rev. Microbiol.* **2**, 861–862 (2004).
- Wiltzius, J.J.W., Sievers, S.A., Sawaya, M.R., Cascio, D., Popov, D., Riek, C., and Eisenberg, D., Atomic structure of the cross- β

spine of islet amyloid polypeptide (amylin), *Prot. Sci.* **17**, 1467–1474 (2008).

Zahn, R., Liu, A., Lührs, T., Riek, R., von Schroetter, C., Garcia, F.L., Billeter, M., Calzolari, L., Wider, G., and Wüthrich, K., NMR solution structure of the human prion protein, *Proc. Natl. Acad. Sci.* **97**, 145–150 (2000); and Liu, H., Farr-Jones, S., Ulyanov, N.B., Llinas, M., Marqusee, S., Groth, D., Cohen, F.E., Prusiner, S.B., and James, T.L., Solution structure of Syrian hamster prion protein rPrP(90–231), *Biochemistry* **38**, 5362–5377 (1999).

Structural Evolution

Bajaj, M. and Blundell, T., Evolution and the tertiary structure of proteins, *Annu. Rev. Biophys. Bioeng.* **13**, 453–492 (1983).

Dickerson, R.E., Timkovitch, R., and Almasy, R.J., The cytochrome fold and the evolution of bacterial energy metabolism, *J. Mol. Biol.* **100**, 473–491 (1976).

Eventhoff, W. and Rossmann, M., The structures of dehydrogenases, *Trends Biochem. Sci.* **1**, 227–230 (1976).

Lesk, A.M., NAD-binding domains of dehydrogenases, *Curr. Opin. Struct. Biol.* **5**, 775–783 (1995).

Moore, A.D., Björklund, Å.K., Ekman, D., Bornberg-Baur, E., and Elofsson, A., Arrangements in the modular evolution of proteins, *Trends Biochem. Sci.* **33**, 444–451 (2008).

Scott, R.A. and Mauk, A.G. (Eds.), *Cytochrome c. A Multidisciplinary Approach*, University Science Books (1996).

PROBLEMS

1. How long will it take the polypeptide backbone of a 6-residue folding nucleus to explore all its possible conformations? Repeat the calculation for 10-, 15-, and 20-residue folding nuclei. Why, in the classic view of protein folding, are folding nuclei thought to be no larger than 15 residues?

***2.** Consider a protein with 10 Cys residues. On air oxidation, what fraction of the denatured and reduced protein will randomly reform the native set of disulfide bonds if: (a) The native protein has five disulfide bonds? (b) The native protein has three disulfide bonds?

3. Why are β sheets more commonly found in the hydrophobic interiors of proteins than on their surfaces?

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

5. Explain how landscape theory is consistent with the observation that many small proteins appear to fold to their native conformations without detectable intermediates, that is, via two-state mechanisms.

6. Explain why Pro residues can occupy the N-terminal turn of an α helix.

7. Explain why β sheets are less likely to form than α helices during the earliest stages of protein folding.

8. Molten globules are thought to be predominantly stabilized by hydrophobic forces. Why aren't hydrogen bonding forces implicated in doing so?

***9.** The GroEL/ES cycle diagrammed in Fig. 9-25 only circulates in the clockwise direction. Explain the basis for this irreversibility in terms of the sequence of structural and binding changes in the GroEL/ES system.

***10.** Predict the secondary structure of the C peptide of proinsulin (Fig. 9-4) using the methods of Chou–Fasman and Rose.

11. As Mother Nature's chief engineer, now certified as a master helix builder, you are asked to repeat Problem 8-8 with the stipulation that the α helix really be helical. Use Table 9-1.

***12.** Predict the secondary structure of the N-terminal domain of yeast protein disulfide isomerase using Jpred3 (<http://www.compbio.dundee.ac.uk/www-jpred/>). How does this prediction compare with the observed structure of this domain (PDBid 2B5E; the domain in Fig. 9-17)? [To enter the sequence of this domain into Jpred3, first point your browser at the PDB (<http://www.rcsb.org/pdb>), enter the PDBid 2B5E, click on the "Sequence Details" tab at the top of the resulting page, and determine the sequence range of the N-terminal domain. Then click on the UniProt reference (P17967), scroll down to the sequence, copy the relevant portion to the Jpred3 input box, edit out everything but the sequence, and click on the "Make Prediction" button. On the page that comes up indicating that sequence matches were found in the PDB, click on the "continue" button. When the Results page appears (you may have to wait some time for it), click on "View Simple" to see the Jpred3 prediction. The observed secondary structure of the N-terminal domain is diagrammed on the foregoing "Sequence Details" page.]

13. Indicate the probable effects of the following mutational changes on the structure of a protein. Explain your reasoning. (a) Changing a Leu to a Phe, (b) changing a Lys to a Glu, (c) changing a Val to a Thr, (d) changing a Gly to an Ala, and (e) changing a Met to a Pro.

14. Explain why Trp rings are usually completely immobile in proteins that have rapidly flipping Phe and Tyr rings.

15. Explain why *Prn-p*^{0/0} mice are resistant to scrapie. What might be the susceptibility of heterozygous *Prn-p*^{+/0} mice to scrapie?

***16.** Discuss the merits of the hypothesis that the dinucleotide-binding domains of the dehydrogenases arose by convergent evolution.