1

GENERAL OVERVIEW OF BASIC CONCEPTS IN MOLECULAR BIOPHYSICS
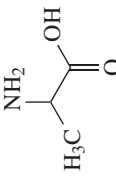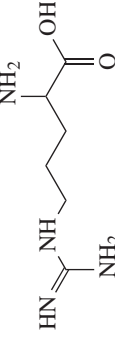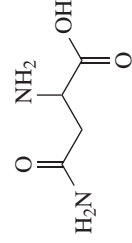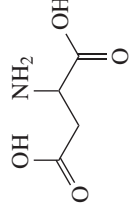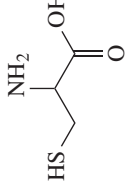
This chapter provides a brief overview of the basic concepts and current questions facing biophysicists in terms of the structural characterization of proteins, protein folding, and protein–ligand interactions. Although this chapter is not meant to provide an exhaustive coverage of the entire field of molecular biophysics, the fundamental concepts are explained in some detail to enable anyone not directly involved with the field to understand the important aspects and terminology.
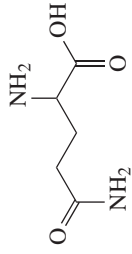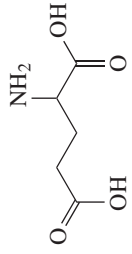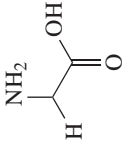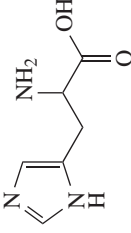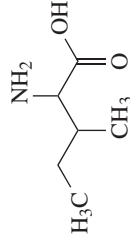
1.1. COVALENT STRUCTURE OF BIOPOLYMERS

Biopolymers are a class of polymeric materials that are manufactured in nature. Depending on the building blocks (or *repeat units* using polymer terminology), biopolymers are usually divided into three large classes. These are polynucleotides (built of nucleotides), peptides and proteins (built of amino acids), and polysaccharides (built of various saccharide units). In this chapter we only consider the general properties of biopolymers using peptides and proteins as examples; questions related to polynucleotides and polysaccharides will be discussed in some detail in Chapter 9.
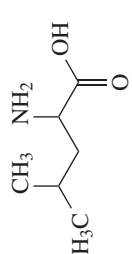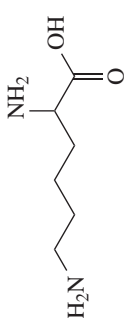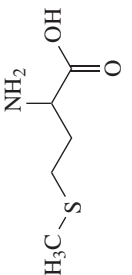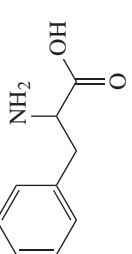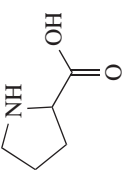
All polypeptides are linear chains built of small organic molecules called *amino acids*. There are 20 amino acids that are commonly considered *canonical* or *natural*. This assignment is based on the fact that these 20 amino acids correspond to 61 (out of a total 64) codons within the triplet code with three remaining codons functioning as terminators of protein synthesis (Table 1.1) (1, 2), although there
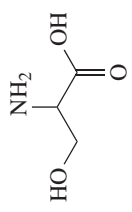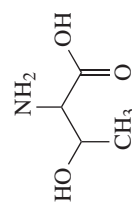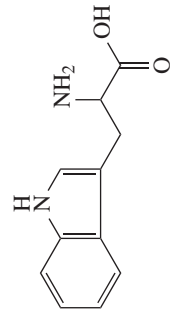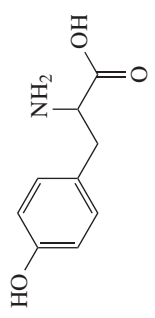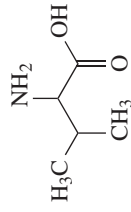
2

TABLE 1.1. Chemical Structure and Masses of Natural (Canonical) Amino Acids

Symbol	Name	Molecular Formula (Residue)	Chemical Structure	Side Chain Character	Monoisotopic Mass[a] (Residue)	Average Mass (Residue)
Ala (A)	Alanine	C_3H_5NO		Nonpolar	71.037	71.079
Arg (R)	Arginine	$C_6H_{12}N_4O$		Basic	156.101	156.188
Asn (N)	Asparagine	$C_4H_6N_2O_2$		Polar	114.043	114.104
Asp (D)	Aspartic acid	$C_4H_5NO_3$		Acidic	115.027	115.089
Cys (C)	Cysteine	C_3H_5NOS		Polar/acidic	103.009	103.145

	Name	Formula	Structure	Property		
Gln (Q)	Glutamine	$C_5H_8N_2O_2$		Polar	128.059	128.131
Glu (E)	Glutamic acid	$C_5H_7NO_3$		Acidic	129.043	129.116
Gly (G)	Glycine	C_2H_3NO		Nonpolar	57.021	57.052
His (H)	Histidine	$C_6H_7N_3O$		Basic	137.059	137.141
Ile (I)	Isoleucine	$C_6H_{11}NO$		Nonpolar	113.084	113.160

TABLE 1.1 (*Continued*)

Symbol	Name	Molecular Formula (Residue)	Chemical Structure	Side Chain Character	Monoisotopic Mass[a] (Residue)	Average Mass (Residue)
Leu (L)	Leucine	$C_6H_{11}NO$		Nonpolar	113.084	113.160
Lys (K)	Lysine	$C_6H_{12}N_2O$		Basic	128.095	128.174
Met (M)	Methionine	C_5H_9NOS		Nonpolar/amphipathic	131.040	131.199
Phe (F)	Phenylalanine	C_9H_9NO		Nonpolar	147.068	147.177
Pro (P)	Proline	C_5H_7NO		Nonpolar	97.053	97.117

			Structure	Category	Monoisotopic	Average
Ser (S)	Serine	$C_3H_5NO_2$		Polar	87.032	87.078
Thr (T)	Threonine	$C_4H_7NO_2$		Polar/amphipathic	101.048	101.105
Trp (W)	Tryptophan	$C_{11}H_{10}N_2O$		Amphipathic	186.079	186.213
Tyr (Y)	Tyrosine	$C_9H_9NO_2$		Amphipathic	163.063	163.176
Val (V)	Valine	C_5H_9NO		Nonpolar	99.068	99.133

[a]See Chapter 3 for a definition of monoisotopic and average masses.

5

are at least as many other amino acids that occur less frequently in living organisms (Table 1.2). Noncanonical amino acids are usually produced by chemical modification of a related canonical amino acid (e.g., oxidation of proline produces hydroxyproline), although at least two of them (selenocysteine and pyrrolysine) should be considered canonical based on the way they are produced and utilized in protein synthesis in vivo by some organisms [the UGA codon that was originally considered as a termination codon is now known to serve also as a Sec (selenocysteine) codon] (3, 4). Furthermore, new components can be added to the protein biosynthetic machinery of both prokaryotes and eukaryotes, which makes it possible to genetically encode unnatural amino acids in vivo (5, 6).

A peculiar structural feature of all canonical (with the exception of glycine) and most noncanonical amino acids is the presence of an asymmetric carbon atom (C^α), which should give rise to two different enantiomeric forms. Remarkably, all canonical amino acids are of the L-type. (D-Forms of amino acids can also be synthesized in vivo and are particularly abundant in fungi; however, these amino acids do not have access to the genetic code.) The rise and persistence of homochirality in the living world throughout the entire evolution of life remains one of the greatest puzzles in biology (7, 8). Examples of homochirality at the molecular level also include almost exclusive occurrence of the D-forms of sugars in the nucleotides. Manifestations of homochirality at the macroscopic level range from specific helical patterns of snail shells to chewing motions of cows.

Unlike most synthetic polymers and structural biopolymers (several examples of which will be presented in Chapter 9), peptides and proteins have a very specific sequence of monomer units. Therefore, even though polypeptides can be considered simply as highly functionalized linear polymers constituting a nylon-2 backbone, these functional groups, or *side chains*, are arranged in a highly specific order. All naturally occurring proteins consist of an exact sequence of amino acid residues linked by peptide bonds (Figure 1.1A), which is usually referred to as the *primary structure*. Some amino acids can be modified after translation, for instance, by phosphorylation or glycosylation. Among these modifications, formation of the covalent bonds between two cysteine residues is particularly interesting, since such *disulfide bridges* can stabilize protein geometry, in which the residues that are distant in the primary structure are held in close proximity to each other in three-dimensional space. A highly specific spatial organization of many (but not all) proteins under certain conditions is often referred to as *higher order structure* and is another point of distinction between them (as well as most biological macromolecules) and the synthetic polymers. Although the disulfide bridges are often important contributors to the stability of the higher order structure, correct protein folding does not necessarily require such covalent "stitches." In fact, cysteine is one of the less abundant amino acids, and many proteins lack it altogether. As it turns out, relatively weak noncovalent interactions between the functional groups of the amino acid side chains and the polypeptide backbone are much more important for the highly specific arrangement of the protein in three-dimensional space. The following section provides a brief overview of such interactions.

TABLE 1.2. Chemical Structure and Masses of Some Less Frequently Occurring Natural (Noncanonical) Amino Acids
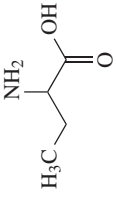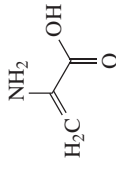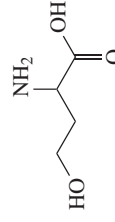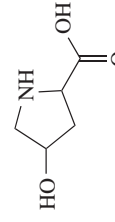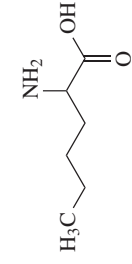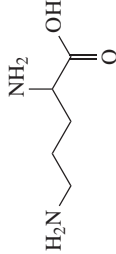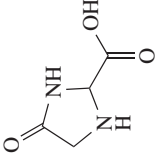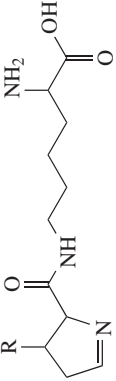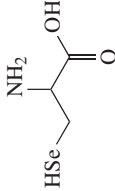
Symbol	Name	Molecular Formula (Residue)	Chemical Structure	Side Chain Character	Monoisotopic Mass (Residue)	Average Mass (Residue)
Abu	2-Aminobutyric acid	C_4H_7NO		Nonpolar	85.053	85.106
Dha	Dehydroalanine	C_3H_3NO		Nonpolar	69.021	69.063
Hse	Homoserine	$C_4H_7NO_2$		Polar	101.048	101.105
Hyp	Hydroxyproline	$C_5H_7NO_2$		Polar	113.048	113.116
Nle	Norleucine	$C_6H_{11}NO$		Nonpolar	113.084	113.160

7

TABLE 1.2 (*Continued*)

Symbol	Name	Molecular Formula (Residue)	Chemical Structure	Side Chain Character	Monoisotopic Mass (Residue)	Average Mass (Residue)
Orn	Ornithine	$C_5H_{10}N_2O$		Basic	114.079	114.147
Pyr	Pyroglutamic acid	$C_5H_5NO_2$		Moderately polar	111.032	111.100
Pyl	Pyrrolysine	$C_{11}H_{16}N_3O_2 +$ R (NH_2, **OH**, or CH_3)		Polar		
Sec	Selenocysteine	C_3H_5NOSe		Polar/acidic	144.960 (150.954[a])	150.039

[a]Most abundant.

```
MTTASTSQVR QNYHQDSEAA INAQINLELY ASYVYLSMSY YFDRDDVALK NFAKYFLHQS
HEEREHAEKL MKLQNQRGGR IFLQDIKKPD CDDWESGLNA MECALHLEKN VNQSLLELHK
LATDKNDPHL CDFIETHYLN EQVKAIKELG DHVTNLRKMG APESGLAEYL FDKHTLGDSD NES
```
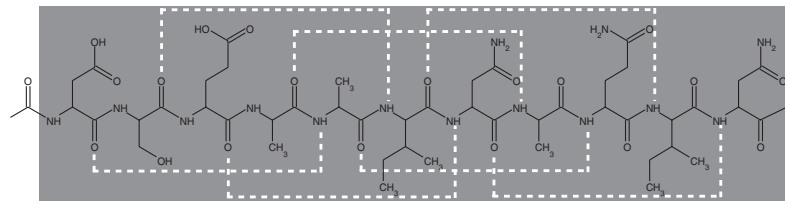

(a)
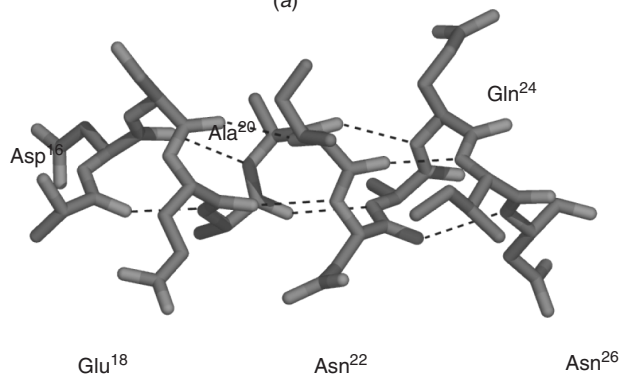
(b)

(c)
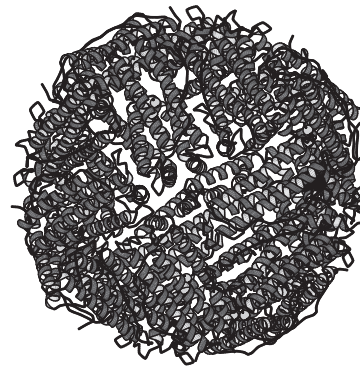
(d)

FIGURE 1.1. Hierarchy of structural organization of a protein (H-form of human fer-ritin). Amino acid sequence determines the primary structure (**A**). Covalent structure of the 11 amino acid residue long segment of the protein (Glu^{16} → Asn^{26}) is shown in the shaded box. A highly organized network of hydrogen bonds along the polypeptide backbone (shown with dotted lines) gives rise to a secondary structure, an α-helix (**B**). A unique spatial arrangement of the elements of the secondary structure gives rise to the tertiary structure, with the shaded box indicating the position of the (Glu^{16} → Asn^{26}) segment (**C**). Specific association of several folded polypeptide chains (24 in the case of ferritin) produces the quaternary structure (**D**).

1.2. NONCOVALENT INTERACTIONS AND HIGHER-ORDER STRUCTURE

Just like all chemical forces, the inter- and intramolecular interactions (both covalent and noncovalent) involving biological macromolecules are electrical in nature and can be described generally by the superposition of Coulombic potentials. In practice, however, the noncovalent interactions are subdivided into several categories, each being characterized by a set of unique features.

1.2.1. Electrostatic Interaction

The term *electrostatic interaction* broadly refers to a range of forces exerted among a set of stationary charges and/or dipoles. The interaction between two fixed charges q_1 and q_2 separated by a distance r is given by the Coulomb law:

$$E = \frac{q_1 q_2}{4\pi \varepsilon_0 \varepsilon r}, \qquad (1\text{-}2\text{-}1)$$

where ε_0 (defined in SI to have the numerical value of $8.85 \cdot 10^{-12}$ C^2/N·m) is the absolute permittivity of vacuum and ε is the dielectric constant of the medium. Although the numerical values of the dielectric constants of most homogeneous media are readily available, the use of this concept at the microscopic level is not very straightforward (9, 10). The dielectric constant is a measure of the screening of the electrostatic interaction due to the polarization of the medium, hence the difficulty in defining a single constant for a protein, where such screening depends on the exact location of the charges, their environment, and so on. Although in some cases the values of the "effective" dielectric constants for specific protein systems can be estimated based on the experimental measurements of the electrostatic interactions, such an approach has been disfavored by many for a long time (11). In this book we will follow the example set by Daune (12) and will write all expressions with $\varepsilon = 1$.

Interaction between a charge q and a permanent dipole p separated by distance r is given by

$$E = -\frac{qp \cdot \cos \theta}{4\pi \varepsilon_0 r^2}, \qquad (1\text{-}2\text{-}2)$$

where θ is the angle between the direction of the dipole and the vector connecting it with the charge q. If the dipole is not fixed directionally, it will align itself to minimize the energy (1-2-2); that is, $\theta = 0$. However, if such energy is small compared to thermal energy, Brownian motion will result in the averaging of all values of θ with only a small preference for those that minimize the electrostatic energy, resulting in a much weaker overall interaction:

$$E = -\frac{q^2 p^2}{(4\pi \varepsilon_0)^2 \cdot 3k_{\mathrm{B}} T r^4}, \qquad (1\text{-}2\text{-}3)$$

Interaction between two dipoles p_1 and p_2 separated by a distance r in this approximation is given by

$$E = -\frac{2p_1^2 p_2^2}{(4\pi\varepsilon_0)^2 \cdot 3k_{\mathrm{B}}Tr^6},\qquad(1\text{-}2\text{-}4)$$

while the interaction between the two fixed dipoles will be significantly stronger ($\sim 1/r^3$).

Polarization of a molecule can also be viewed in terms of electrostatic interaction using a concept of *induced dipoles* (12). Such interaction is, of course, always an attractive force, which is inversely proportional to r^4 (for a charge–induced dipole interaction) or r^6 (for a permanent dipole–induced dipole interaction). Finally, interaction between two polarizable molecules can be described in terms of a weak induced dipole–induced dipole interaction.

1.2.2. Hydrogen Bonding

The electrostatic interactions considered in the preceding sections can be treated using classical physics. *Hydrogen bonding* is an example of a specific noncovalent interaction that cannot be treated within the framework of classical electrostatics. It refers to an interaction occurring between a proton donor group (e.g., —OH, —NH$_3{}^+$) and a proton acceptor atom that has an unshared pair of electrons. Although hydrogen bond formation (e.g., R=Ö: ••• H—NR$_2$) may look like a simple electrostatic attraction of the permanent dipole–induced dipole type, the actual interaction is more complex and involves charge transfer within the proton donor–acceptor complex. The accurate description of such exchange interaction requires the use of sophisticated apparatus of quantum mechanics.

The importance of hydrogen bonding as a major determinant and a stabilizing factor for the higher-order structure of proteins was recognized nearly seventy years ago by Mirsky and Pauling, who wrote in 1936: "the [native protein] molecule consists of one polypeptide chain which continues without interruption throughout the molecule... this chain is folded into a uniquely defined configuration, in which it is held by hydrogen bonds between the peptide nitrogen and oxygen atoms..." (13). Considerations of the spatial arrangements that maximize the amount of hydrogen bonds within a polypeptide chain later led Pauling to the prediction of the existence of the α-helix, one of the most commonly occurring local motifs of higher order structure in proteins (14). Hydrogen bonds can be formed not only within the macromolecule itself, but also between biopolymers and water molecules (the latter act as both proton donors and acceptors). Hydrogen bonding is also central for understanding the physical properties of water, as well as other protic solvents.

1.2.3. Steric Clashes and Allowed Conformations of the Peptide Backbone: Secondary Structure

Both electrostatic and hydrogen bonding interactions within a flexible macro-molecule would favor three-dimensional arrangements of its atoms that minimize
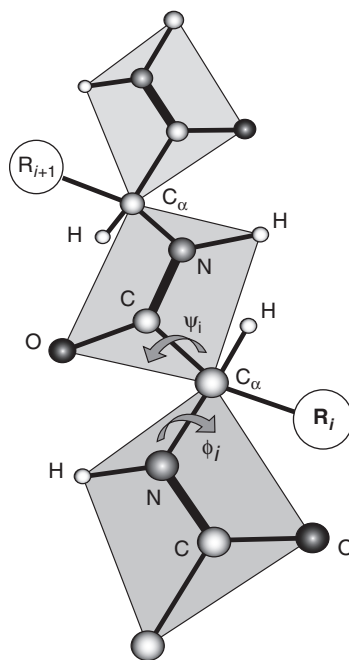
FIGURE 1.2. Peptide bond and the degrees of freedom determining the polypeptide backbone conformation.

the overall potential energy. However, there are two fundamental restrictions that limit the conformational freedom of the macromolecule. One is, of course, the limitation imposed by covalent bonding. The second limitation is steric hindrance, which also restricts the volume of conformational space available to the biopolymer. In this section we consider the limits imposed by steric clashes on the conformational freedom of the polypeptide backbone.

The peptide amide bond is represented in Figure 1.1A as a single bond (i.e., C—N); however, it actually has a partial double bond character in a polypeptide chain. The double bond character of the C—N linkage, as well as the strong preference for the *trans* configuration of the amide hydrogen and carbonyl oxygen atoms,* result in four atoms lying in one plane. A slight deformation of this configuration does occur in many cases, but it is rather insignificant. Figure 1.2 shows two successive planes linked by a C^α atom of the ith amino acid residue. The two degrees of freedom at this junction are usually referred to as ϕ_i and ψ_i angles and the backbone conformation of the polypeptide composed of n amino acid residues can be described using $n - 1$ parameters (pairs of ϕ_i and ψ_i). The

---

*The exception to this rule is offered by proline, which, as an imino acid, has its side chain also bonded to the nitrogen atom. Thus, the cis- and trans-forms are almost isoenergetic, leading to the possibility of *cis*-Xaa-Pro bonds (where Xaa is any amino acid residue) in folded proteins, and statistically at the level of 5–30% in unstructured polypeptides.
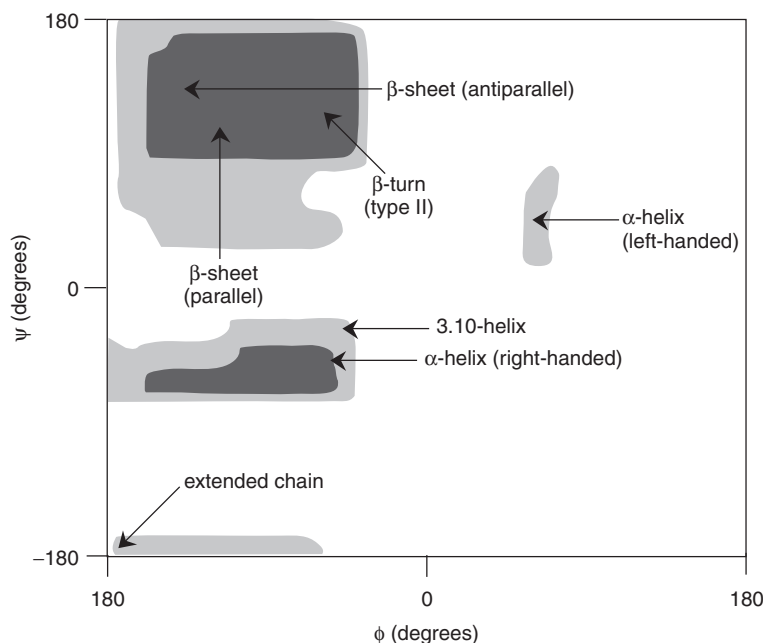
FIGURE 1.3. A schematic representation of the Ramachandran plot.

steric restrictions limit the conformational volume accessible to polypeptides, which is usually represented graphically on the (ϕ, ψ) plane using *conformational maps* or *Ramachandran plots* (15). An example of such a diagram shown in Figure 1.3 clearly indicates that only a very limited number of configurations of the polypeptide backbone are allowed sterically.

Several regions within the accessible conformational volume are of particular interest, since they represent the structures that are stabilized by highly organized networks of hydrogen bonds. The α-helix is one such structure, where the carbonyl oxygen atom of the ith residue is hydrogen bonded to the amide of the $(i + 4)$th residue (Figure 1.1B). This local motif, or spatial arrangement of a segment of the polypeptide backbone, is an example of a *secondary structure*, which is considered the first stage of macromolecular organization to form higher-order structure. Another commonly occurring element of the secondary structure is located within a larger "island" of sterically allowed conformations on the Ramachandran plot. Such conformations [upper left corner on the (ϕ, ψ) plane in Figure 1.3] are rather close to the fully extended configuration of the chain and, therefore, cannot be stabilized by local hydrogen bonds. Nevertheless, formation of strong stabilizing networks of hydrogen bonds becomes possible if two strands are placed parallel or antiparallel to each other, forming the so-called *β-pleated sheets*.

The third important local structural motif is the *turn*, which causes a change in the chain direction within a folded protein. Whereas loops are generally flexible

sections of chain, turn structures tend to be more rigid and are stabilized by hydrogen bonding or specific side chain interactions. These turn structures can be highly important particularly in antiparallel β-sheet structures, where a complete reversal of the chain is required to enable packing of adjacent strands. Other less frequently occurring elements of secondary structure (such as 3.10 or π helices) can also be identified on the Ramachandran plot.

So far, we have largely ignored the contributions of the amino acid side chains to the protein conformation. One obvious consequence of the existence of a variety of different side chains is the dependence of the Ramachandran plots for each particular (ϕ_i, ψ_i) pair on the identity of the ith amino acid residue. For example, a significantly larger conformational volume is available to glycine, as compared to amino acid residues with bulky side chains. Furthermore, different side chains placed at "strategic" locations may exert a significant influence on the stability of the secondary structural elements. We will illustrate this point using the α-helix as an example. All hydrogen bonds in a helix are nearly parallel to each other (and to the axis of the helix). This highly ordered pattern of hydrogen bonding results in a noticeable dipole moment, with the N-terminal end of the helix being a positive pole. Obviously, the presence of a positively charged residue at or near the N-terminal end of the helix will destabilize it due to the unfavorable charge–permanent dipole interaction (1-2-2). On the other hand, the presence of a negatively charged residue will be energetically favorable and increase the stability of the helix. Likewise, the presence of charged residues at or near the C-terminal end of the helix will also have a significant influence on the stability of this element of secondary structure.

It is important to note, however, that uncharged side chains may also be very important determinants of the higher-order structure of proteins and polypeptides due to the *hydrophobic interactions*. These will be considered in the following section.

1.2.4. Solvent–Solute Interactions, Hydrophobic Effect, Side Chain Packing, and Tertiary Structure

The term *hydrophobic effect* (16–19) refers to a tendency of nonpolar compounds (such as apolar amino acid side chains, Table 1.1) to be sequestered from the polar solutions (such as aqueous solution) to an organic phase. Such behavior is ubiquitous in nature and has been observed and described at least two millennia ago, although the term hydrophobic was coined only in 1915 (18). The initial view of the hydrophobic interaction was rather simplistic and implied attraction between the like media (e.g., oil–oil attraction). A very different view (which is now commonly accepted) was proposed in the mid-1930s by Hartley, who suggested that the nonpolar species are excluded from polar solvent because of their inability to compete with the strong interaction between the polar molecules themselves (20). In Tanford's words, "antipathy between hydrocarbon and water rests on the strong attraction of water for itself" (21). An intriguing aspect of the hydrophobic interaction is that the placement of a hydrocarbon molecule in

water may be enthalpically favorable. This fact was the basis for a widespread skepticism over the concept of hydrophobic interactions, although such views did not prevail (22). It is now understood that the solvent—solute affinity is determined by the free energy (not the enthalpy alone), and it is the unfavorable free energy that leads to the observed disaffinity of water and nonpolar solutes.

Various microscopic explanations of the hydrophobic effect are usually based on the *frozen water patches* or *microscopic iceberg* model proposed originally by Frank and Evans (23), who suggested that placing a nonpolar solute in water creates a loose "cage" of first-shell water molecules around it. The creation of such a cage has a significant entropic price (due to the "forced" water ordering), hence the overall unfavorable free energy (despite favorable enthalpic contribution). A reader interested in a more detailed account of the physics of hydrophobicity and related phenomena is referred to a recent tutorial by Southall, Dill, and Haymet (18).

Although the initial work on the hydrophobic effect was focused on hydrocarbons, its main results and conclusions can easily be extended to nonpolar side chains of polypeptides and proteins, which are buried in a *hydrophobic core* of a folded or collapsed protein molecule in order to eliminate or at least minimize any contacts with the polar solvent. A very interesting historical account on the elucidation of the nature of hydrophobic interaction and its role in protein folding can be found in an excellent review by Tanford (24). While the hydrophobic side chains are generally more stable if sequestered away from solvent in the central core of the protein,* the hydrophilic residues usually decorate the solvent-exposed surface of the protein. This is achieved by combining the elements of secondary structure (α-helices, β-sheets, and turns) in a unique three-dimensional arrangement, or *tertiary structure*. It is the tertiary structure that affords proteins their unique biological function, whether it be purely structural, the precise spatial organization of side chains to effect catalysis of a reaction, presentation of a surface or loop for signaling or inhibition, creating a cavity or groove to bind ligand, or any of the other vast range of functions that proteins can perform.

Hydrophobic interaction is, of course, not the only driving force giving rise to a unique tertiary structure. Additional stabilization is afforded by close proximity of acidic and basic residues, which is frequently observed in the folded structure, enabling the formation of salt bridges. These can be viewed as charge–charge interactions (1-2-1). We have already mentioned that certain elements of secondary structure have intrinsic (permanent) dipole moments. Favorable arrangement of such dipoles with respect to one another (e.g., in the so-called helical bundles) may also become a stabilizing factor (1-2-2) in addition to the hydrophobic interaction. It is probably worth mentioning that in the vast majority of proteins the interactions stabilizing the tertiary structure are cooperative. In other words, significant enthalpic gains are achieved only if several segments of the protein are in close proximity and interact with each other. All such factors have been evolutionarily optimized for each protein but the important thing to realize is

---

*Proteins tend to be very well-packed molecules so the side chain atoms sequestered from the solvent must come into close contact with each other, hence the term *hydrophobic packing*.

(a)   (b)

(c)   (d)
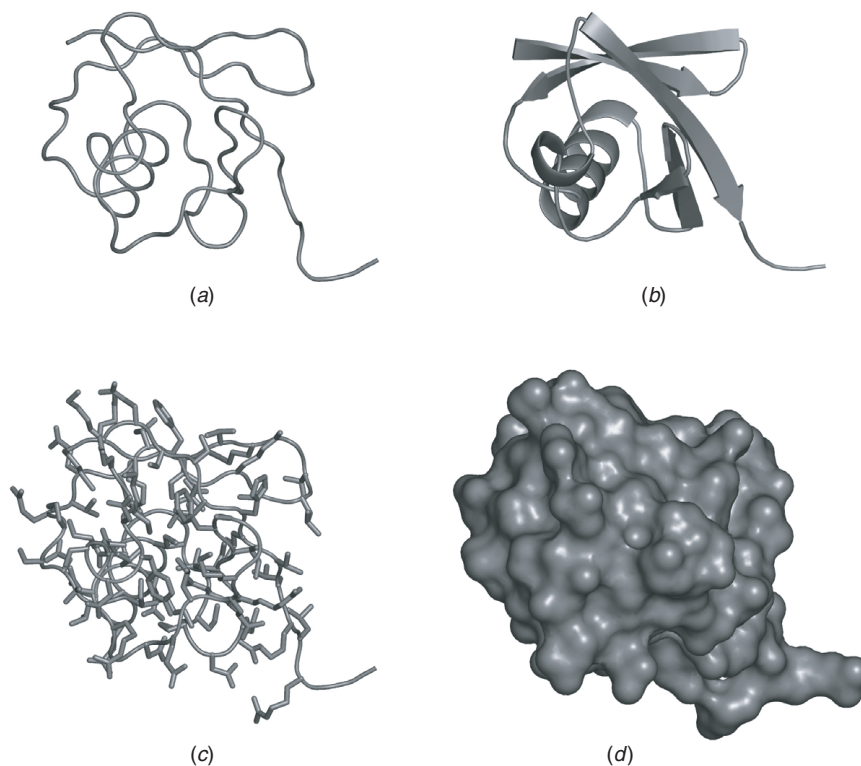
FIGURE 1.4. Different representations of the higher order structure of natively folded proteins.

that any one natural protein sequence has only a single most stable conformation, and the genetically encoded primary sequence alone is necessary and sufficient to define the final folded structure of the protein (Figure 1.4).

Many proteins adopt similar common structural motifs resulting from combinations of secondary structure elements such as the alternating $\beta\alpha\beta$ structure, 4-helix bundles, or β-barrels. As more and more protein structures are solved,[*] the number of protein architectures increases, although it has been predicted there are only a limited number of fold motifs (25–29). Such conclusion is based on the observations that (i) topological arrangements of the elements of secondary structure are highly skewed by favoring very few common connectivities and (ii) folds can accommodate unrelated sequences [as a general rule, structure is more robust than sequence (30, 31)]. Therefore, the fold universe appears to be dominated by a relatively small number of giant attractors,[†] each accommodating

---

[*]So far, structures have been solved for only about 1% of approximately 1,000,000 proteins of known sequences.
[†]The total number of folds is estimated to be less than 2000, of which 500 have already been characterized.
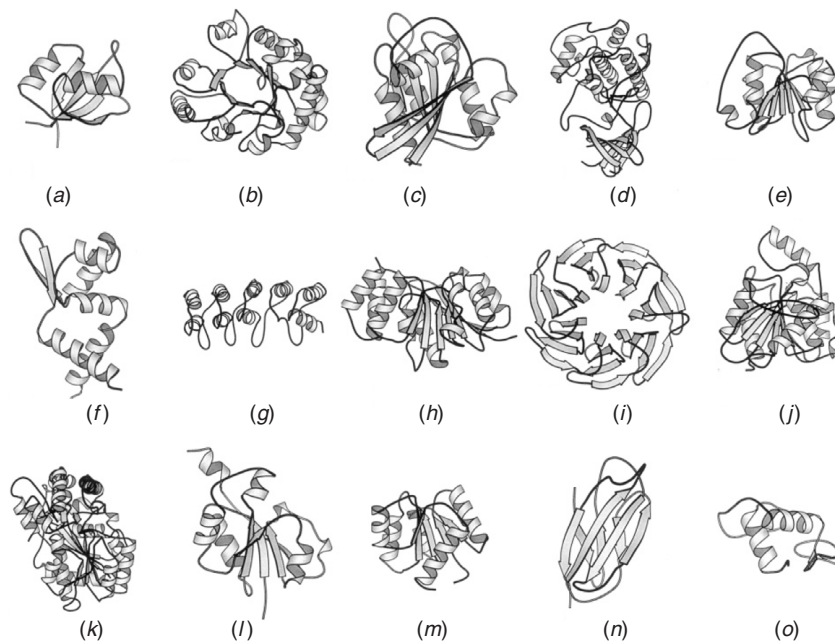
FIGURE 1.5. The 15 most populated folds selected on the basis of a structural annotation of proteins from completely sequenced genomes of 20 bacteria, five Archaea, and three eukaryotes. From left to right and top to bottom: ferredoxin-like (4.45%) (**A**), TIM-barrel (3.94%) (**B**), P-loop containing nucleotide triphosphate hydrolase (3.71%) (**C**), protein kinases (PK) catalytic domain (3.14%) (**D**), NAD(P)-binding Rossmann-fold domains (2.80%) (**E**), DNA:RNA-binding 3-helical bundle (2.60%) (**F**), α-α superhelix (1.95%) (**G**), *S*-adenosyl-L-methionine-dependent methyltransferase (1.92%) (**H**), 7-bladed β-propeller (1.85%) (**I**), α/β-hydrolases (1.84%) (**J**), PLP-dependent transferase (1.61%) (**K**), adenine nucleotide α-hydrolase (1.59%) (**L**), flavodoxin-like (1.49%) (**M**), immunoglobulin-like β-sandwich (1.38%) (**N**), and glucocorticoid receptor-like (0.97%) (**O**). The values in parentheses are the percentages of annotated proteins adopting the respective folds. Reproduced with permission from (32). © 2001 Springer-Verlag.

a large number of unrelated sequences. Figure 1.5 represents the 15 most populated folds selected on the basis of a structural annotation of proteins from completely sequenced genomes of 20 bacteria, five Archaea, and three eukaryotes (32).

The existence of a "finite set of natural forms" in the protein world has inspired some to invoke the notion of Platonic forms that are "determined by natural law" (33), a suggestion that seems more poetic than explanatory. What has become clear though is that very similar tertiary structures can be adopted by quite dissimilar primary sequences (32). Protein primary sequences can be aligned and regions identified that are identical or homologous (meaning the chemical nature of the amino acid side chain is similar—e.g., polar, nonpolar, acidic, basic). However, even sequences with quite low homology can have a

very similar overall fold, depending on the tertiary interactions that stabilize them. Although tertiary structure is sometimes viewed as the highest level of spatial organization of the single-chain (i.e., monomeric) proteins, an even higher level of organization is often seen in larger proteins (generally more than 150 amino acid residues). Such proteins form clearly recognizable *domains*, which tend to be contiguous in primary structure and often enjoy certain autonomy from one another.

1.2.5. Intermolecular Interactions and Association: Quaternary Structure

Above and beyond the folding of monomeric chains, many protein chains can also assemble to form multisubunit complexes, ranging from relatively simple homodimers (such as hemoglobin molecules of primitive vertebrates, e.g. lamprey, hagfish) to large homo-oligomers (such as an iron-storage protein ferritin, assembled of 24 identical subunits) to indeed assemblies of different proteins (such as ribosomes). Such assemblies are usually considered to be the highest level of molecular organization at the microscopic level, which is usually referred to as *quaternary structure*. Although covalent links are sometimes formed between the monomeric constituents of a multimeric protein assembly (e.g., in a form of the disulfide bonds), the noncovalent interactions (discussed in the preceding sections) are usually much more important players.

The archetype of quaternary structure is the mammalian hemoglobin, which is a noncovalent tetramer ($\alpha_2\beta_2$) consisting of two pairs of similar monomeric chains (α- and β-globins). The arrangement of the monomers in the tetramer (which is in fact a dimer composed of two heterodimers) is crucial for the function of hemoglobin as an oxygen transporter. A tetramer composed of four identical globins (β_4) can also be formed and is indeed present in the blood of people suffering from some forms of thalassemia. However, this homotetramer (termed hemoglobin H, or Hb H), lacks the most important characteristic of the "normal" hemoglobin (Hb A), namely, high cooperativity of oxygen binding. We will revisit quaternary structure formation and functional characteristics in the closing chapter of this book.

1.3. THE PROTEIN FOLDING PROBLEM

1.3.1. What Is Protein Folding?

Polymers can adopt different conformations in solution depending on functionality and the interaction with neighboring chains, other parts of the same chain, and the bulk solvent. However, almost all synthetic copolymers (polymers consisting of more than one type of repeat units) consist of a range of different length chains and, in many cases, a nonspecific arrangement of monomer groups. On the other hand, the primary structure of a given protein is always the same, creating a homogeneous and highly monodisperse copolymer. Protein sequences are generally optimized to prevent nonspecific intermolecular interactions and individual

molecules will fold to adopt a unique stable conformation governed solely by the primary sequence of amino acids. The ability of proteins to attain a unique higher-order structure sets them apart from most random copolymers. Most proteins can fold reversibly in vitro, without being aided by any sophisticated cellular machinery,* suggesting that the folding mechanism is solely determined by the primary structure of the protein, as well as the nature of the solvent. Folded proteins may remain stable indefinitely in most cases, suggesting that the native structures represent the global free energy minima among all kinetically accessible states (34).

Two classic puzzles are usually considered in connection with protein folding: (i) the *Blind Watchmaker's paradox* and (ii) the *Levinthal paradox*. The former is named after a classic book by Dawkins (35), an outspoken critic of the *intelligent design* concept (36). It states that biological (function-competent) proteins could not have originated from random sequences. The Levinthal paradox states that the folded state of a protein cannot be found by a random search (37). Both paradoxes have been historically framed in terms of a random search through vast spaces (sequence space in the *Blind Watchmaker's paradox* and conformational space in the *Levinthal paradox*), and the vastness of the searched spaces is equated with physical impossibility. Both paradoxes are elegantly solved within the framework of the *energy landscape* description of the folding process by invoking the notion of a guided search (38). The concept of protein energy landscapes and its relevance to the protein folding problem will be considered in some detail in the following sections of this chapter.

1.3.2. Why Is Protein Folding So Important

In the postgenomic era, structure determination has become of paramount importance since it leads to a three-dimensional picture of each gene product and, in many cases, gives hints as to the function of the protein. However, the static structure only represents the end point of the chemical reaction of protein folding. Polypeptide chains are translated as extended structures from RNA on the ribosome of cells, but how does this unstructured sequence fold into its final biologically active structure? Are specific local structures present in the newly translated chain? Is there a specific pathway or reaction coordinate of protein folding? The principles that govern the transitions of biopolymers from totally unstructured to highly ordered states (which often include several subunits assembled in a highly organized fashion) remain one of the greatest mysteries in structural biology (39, 40). Deciphering this code is key to understanding a variety of biological processes at the molecular level (e.g., recognition, transport, signaling, and biosynthesis), since the specificity of biological activity in proteins (as well as other biomolecules) is dictated by their higher-order structure.

---

*In the cell a number of helper proteins, or chaperones, assist efficient folding in the crowded cytosolic environment. However, it is generally agreed that chaperones do not in themselves direct folding, instead acting as gatekeepers to prevent misfolding and aggregation.

Aside from the obvious academic interest to biophysicists in discovering exactly how these biological machines work, there are many more practical implications. Only if we understand all of the processes that are involved in producing a biologically active protein can we hope to harness this power by designing proteins with specific functions. It may already be possible computationally to model an ideal binding site or even optimal arrangement of side chains to catalyze a chemical reaction, but without a thorough knowledge of how this site can be placed into an intact protein molecule, we cannot take advantage of the cellular machinery for the design of therapeutic protein drugs, or even molecules that can catalyze otherwise difficult chemical reactions. For instance, there are many enzymes in nature that catalyze reactions with extremely high specificity and efficiency, whereas chemists lag far behind. Hydrogenase enzymes, for example, catalyze the reduction of protons to produce diatomic hydrogen, a reaction that in a laboratory environment requires application of harsh reactants at elevated temperature or pressure, but which within the catalytic center of the protein occurs at physiological temperatures and with remarkably small energy requirements. Obviously, biological organisms have had a much longer time to optimize these processes relative to the chemical industry. If one can understand in detail the roles of each residue in a protein chain for both the folding and dynamics of the molecule, then the possibilities for protein engineering are boundless. Interestingly, artificial sequences quite often lead to proteins that either do not fold at all or are only marginally stable. This clearly demonstrates the extremely fine balance of forces present, which can be destroyed by just a single amino acid residue substitution, deletion, or insertion.

Another important aspect of understanding protein folding is to find ways of preventing the process from going awry (41–44). An ever-increasing number of pathological conditions that result from misfolding of proteins in the cell are being identified (45–49). Amyloid plaques actually result from the undesirable formation of quaternary structure when a normally monomeric peptide folds incorrectly and self-assembles to form long proteinaceous fibers. Similarly, other proteins that are not correctly folded may not present the correct binding surface for interaction with their physiological partners. Thus, not only correct folding but also the correct assembly of proteins are key to their correct biological function. Even relatively few mutations within a protein sequence may prevent folding to the native structure and hence prove pathological. In other cases, mutation can reduce the efficiency of folding or favor an alternative mode of folding that leads to aggregation and deposition of insoluble amyloid plaques within cells. We will consider the issues related to misfolding and aggregation in the following sections of this chapter.

Finally, one more fundamental problem related to protein folding that has become a focal point of extensive research efforts is the prediction of the native structure and function of a protein based on its primary structure. Since the sequence of each natural protein effectively encodes a single tertiary structure, prediction of the latter is, in essence, a global optimization problem, which is similar to one encountered in crystallography and the physics of clusters (50). The complication that arises when such a global optimization methodology is applied

to determine the position of the global energy minimum for a protein is the vastness of the system that precludes calculations based on the first principles. So far, the most successful methods of structure prediction rely on the identification of a *template protein* of known structure, whose sequence is highly homologous to that of the protein in question. If no template structure can be identified, de novo prediction methods can be used, although it remains to be seen if such methods can predict structures to a resolution useful for biochemical applications (51). Prediction of protein function based on its sequence and structure is an even more challenging task, since homologous proteins often have different functions (52).

1.3.3. What Is the Natively Folded Protein and How Do We Define a Protein Conformation?

Before proceeding further with a description of protein folding, it would be useful to define some terms commonly used in the field in order to avoid confusion. The *native state* of a protein is defined as the fully folded biologically active form of the molecule. This has generally been considered as a single state with a well-defined tertiary structure, as determined by crystallography or NMR spectroscopy. More recently, researchers have come to appreciate the importance of dynamics within the protein structure. Even the native state is not a static single structure but may in fact, depending on the protein, have small or even large degrees of flexibility that are important for its physiological function.

Unfortunately, the use of the term *protein conformation* in literature has become rather inconsistent and often results in confusion. Historically, protein conformation referred to a specific "three-dimensional arrangement of its constituent atoms" (53). This definition, however, is rather narrow, since it does not reflect adequately the dynamic nature of proteins. One particularly annoying complication that arises when conformation is defined using only microscopic terms (e.g., atomic coordinates) is due to the fact that a majority of proteins have segments lacking any stable structure even under native conditions. These could be either the terminal segments that are often invisible in the X-ray structures or flexible loops whose conformational freedom is often required for a variety of functions ranging from recognition to catalysis. In general, it is more than likely that any two randomly selected natively folded protein molecules will *not* have identical sets of atomic coordinates and, as a result, will not be assigned to one conformation if the geometry-based definition is strictly applied. Therefore, it seems that the thermodynamics-based definition of a protein conformation is a better choice. Throughout this book, we will refer to the protein conformation not as a specific microstate, but as a *macrostate*, which can be envisioned as a collection of microstates separated from each other by low-energy barriers ($\leq k_B T$). In other words, if one microstate is accessible from another at room temperature, we will consider them as belonging to one conformation, even if there is a substantial difference in their configurations. According to this view, a protein conformation is a continuous subset of the conformational space (i.e., a continuum of well-defined configurations) that is accessible to a protein confined to a certain local

minimum. Although such definition is not without its own problems,* its utility becomes obvious when we consider non-native protein conformations.

1.3.4. What Are Non-native Protein Conformations? Random Coils, Molten Globules, and Folding Intermediates

In the case of unfolded proteins, which are assumed to be completely nonrigid polypeptide chains, the *random coil* (54), we must consider the ensemble of molecules displaying an impressive variety of configurations (Figure 1.6). In a truly random coil, as might be the case for a synthetic polymer with identical monomer units in a good solvent, there may well be no conformational preferences for the chain. However, proteins are decorated with side chains of different chemical nature along their length, such that in water or even a chemical denaturant one might expect there to be local preferences due to hydrophilic or hydrophobic interactions, and indeed steric effects. Thus, for a number of proteins studied in solution, some persistent local and nonlocal conformational effects have been detected, indicating that an unfolded protein is not in fact a truly random coil. On the other hand, the enthalpy of these interactions is very small in comparison to the entropy of the flexible chain, so the overall free energy of each of these conformers will be very similar. On a free energy surface, these would be represented as shallow wells in the generally flat surface of unfolded state free energy.

The relative position of a local energy minimum with respect to the native state gives rise to a further set of descriptions of intermediate states. As a protein folds it may sample stabilizing conformations that contain persistent structure constituting a local free energy minimum. At the earliest stages of folding there may only be a few interactions that may be very transient—these are termed *early intermediates*. By contrast, species may accumulate further along in the folding process that contain a large although incomplete number of native-like contacts. These are referred to as *late intermediates*, implying that they should form toward the end of the kinetic folding process. There is also the possibility that these local minima arise from stabilizing contacts that are not present in the native protein, and in fact need to be broken apart before the molecule can productively fold. These *off-pathway intermediates* may also arise from intermolecular interactions between folding chains and can lead to nonproductive aggregation that prevents further folding.

The above intermediate states form during folding in the "forward" direction from the unfolded to the native state and, since they are only partially stable, generally do not accumulate sufficiently to be detected other than transiently. It

*For example, this definition is temperature dependent. Indeed, if any two local minima are separated by a high-energy barrier ($>k_{\mathrm{B}}T$), the interconversion between these two states does not occur readily at room temperature T, and these two states should be viewed as two different conformations. However, raising the temperature significantly above the room temperature T will eventually make the passage over this barrier possible, leading to a merger of the two microstates to a single conformation.
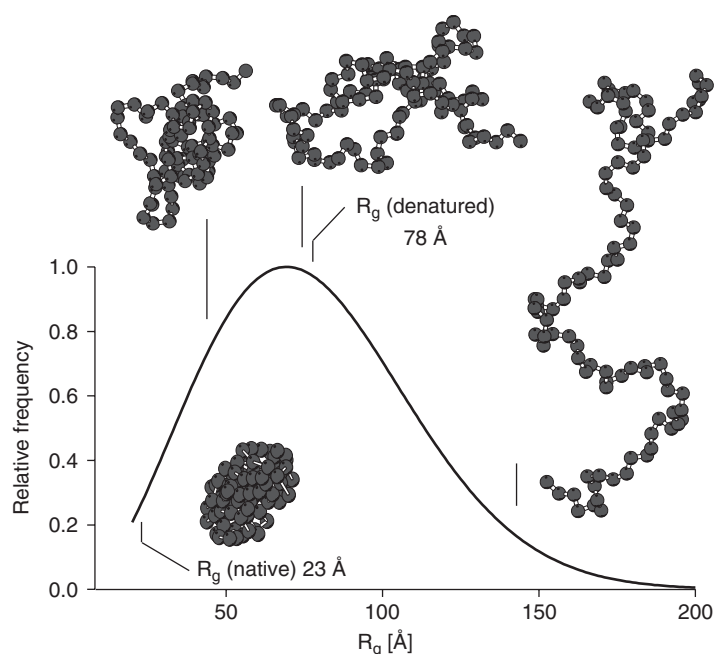
FIGURE 1.6. Representative configurations of a random coil (a freely joined chain of 100 hard spheres) and the distribution of its radius of gyration R_g. The R_g values of a model protein phosphoglycerate kinase are indicated for comparison. Adapted with permission from (54). © 1996 Elsevier.

is also possible that such intermediates may form during the reverse process (i.e. protein unfolding), allowing them to be studied by other methods. Unfortunately, the conditions for unfolding (e.g., chemical denaturant, low pH, high tempera-ture) are generally so harsh that once the stabilizing interactions in the native state have been removed, the unfolding process occurs with high cooperativity and without accumulation of intermediates. However, under mildly denaturing conditions, partially folded states have been detected at equilibrium for a num-ber of proteins, and these have been termed *molten globules* (55). The original definition of the molten globule state was quite specific: a structural state that has significant secondary structure but with no fixed tertiary interactions. There are various biophysical tests for this, such as the ability of the protein to bind hydrophobic dyes, consistent with a significant amount of exposed hydrophobic surface area, as would be expected for a partially folded state. The definition has become somewhat relaxed to include many other partially folded ensembles observed, kinetically or at equilibrium, which almost fit the definition. What is clear is that the molten globule itself is a much more dynamic structure than pre-viously thought. Several new concepts have recently been introduced to reflect the structural diversity and the dynamic character of the molten globule state, such as "a precursor of the molten globule" and "a highly structured molten globule" (56).

One common question that arises is whether the equilibrium molten globule intermediate is actually the same species as that detected in the folding pathway of proteins. Thermodynamically, there is nothing to suggest they should be, since the equilibrium by definition is independent of the pathway (57, 58). However, comparisons of the characteristics of transient intermediates with the corresponding equilibrium partially folded state have concluded that the similarities are very close, at least for the proteins studied (59–62). Also, a number of states transiently populated by the native state ensemble under mildly destabilizing conditions have been shown to have similarities to folding intermediates. Thus, it seems likely that, at least in the later stages of folding, there is indeed some kind of folding/unfolding pathway with specific intermediate states visited in both the folding and unfolding directions.

1.3.5. Protein Folding Pathways

In the preceding section we began to use the term *folding pathway*, which is understood to be a series of structural changes leading from the fully denatured state of the protein to its native conformation. Introduction of the concept of a folding pathway resolves the Levinthal paradox mentioned earlier by suggesting that the folding process is a directed process involving conformational biases, rather than a merely random conformational search. Despite the vast number of degrees of freedom in macromolecules, the number of folding pathways was initially believed to be rather limited (63). A general scheme of protein folding within this paradigm is presented in terms of rapid equilibration of unfolded protein molecules between different conformations prior to complete refolding. Such equilibrium favors certain compact conformations that have lower free energies than other unfolded conformations, and some of these favored conformations are important for efficient folding. The rate-limiting step is thought to occur late in the pathway and to involve a high-energy, distorted form of the native conformation. The latter is a single transition state through which essentially all molecules refold (64).

The classic folding pathway paradigm specifically states that "proteins are not assembled via a large number of independent pathways, nor is folding initiated by a nucleation event in the unfolded protein followed by rapid growth of the folded structure" (64). Nevertheless, a large body of experimental evidence now suggests the existence of large number of folding routes. Furthermore, over the past several years it has become clear that the length of the polypeptide chain is an important factor in determining mechanistic details. Smaller proteins (<100 residues) appear to prefer a nucleation-type mechanism that does not involve any specific metastable intermediate species (65). On the other hand, for a number of larger proteins, intermediate states with specific regions of the early-formed stable structure have been established. If an intermediate is detected, then this argues strongly for a specific protein folding pathway, but in the protein lysozyme, for instance, parallel folding pathways were found, suggesting multiple possible trajectories. In smaller proteins no observable intermediates accumulate, which

might indicate a less directed folding mechanism, or indeed that partially formed structures are too labile to be detected. Based on these observations, a modified view of a folding pathway invokes several common stages of folding, consistent with the progressive development of structure and stability through an ever-slowing set of reactions (66).

Three major models have arisen to attempt to explain this narrowing of the conformational search process (67). The simple *framework* model suggests that secondary structure elements would first form based on their sequence-intrinsic propensities, followed by collision of these preformed structures to form tertiary interactions. Alternatively, *nucleation* of short regions of sequence to form transient secondary structure could act as a template on which adjacent parts of the chain would condense and propagate structure. The third mechanism, *hydrophobic collapse*, calls for the hydrophobic residues to conglomerate nonspecifically to minimize solvent exposure, followed by rearrangement to the final native structure. The actual mechanism of protein folding probably involves some or all of these processes and there is evidence for each mechanism from folding studies of different proteins.

The controversy with regard to whether protein folding follows a specific pathway or whether each molecule follows a completely different trajectory to achieve its final folded state has been elegantly resolved in the *new view of protein folding* (68–70). The centerpiece of this theory is the concept of a protein energy landscape in the conformational space, which is discussed in the following section.

1.4. PROTEIN ENERGY LANDSCAPES AND THE FOLDING PROBLEM

1.4.1. Protein Conformational Ensembles and Energy Landscapes: Enthalpic and Entropic Considerations

Classically, a simple chemical reaction is considered to proceed along a reaction coordinate, in which chemical bonds are formed or broken in a well-defined manner, with a transition state at the highest point on the energy profile where bond breaking and formation are occurring, and possibly detectable metastable intermediate structures at local free energy minima. Transition state theory can be applied to relate reaction rates to the heights of the various free energy barriers along the reaction coordinate. Macromolecules are much more complex, however, and the folding of a protein involves the formation of a large number of interactions that may either be local in nature or indeed involve regions that are quite distant in the polypeptide sequence. Nevertheless, for many years the protein folding reaction was assumed to occur via a similar sequential pathway, perhaps involving a number of intermediate species along the way, but for each unfolded molecule the mechanism of folding to the native state was identical.

Significant advances in theory during the past decade have changed our understanding of the basic principles that govern the protein folding process and have

offered an elegant way to resolve the Levinthal paradox (69, 71–75). In the case of small organic molecules in a reaction, there are only a small number of conformations available to the reactant species, but for an unfolded protein the conformational space sampled by the unstructured chain is vast by comparison, even for a relatively small protein. Thus, it might seem difficult to imagine the chain becoming oriented in such a way as to proceed to fold via a single pathway, and the process would surely be extremely inefficient. Realization that folding may proceed through multiple parallel pathways, rather than a single route, has led to the introduction of the concepts of protein energy landscapes (or folding funnels), a cornerstone of the "new view" of protein folding.

Protein folding can be viewed much like any other chemical reaction, which may be represented in three dimensions by a conformational energy surface: the trajectories on these surfaces lead from reactants (unstructured states) to products (the native state). Because entropy plays a much more significant role in protein folding reactions, it is necessary to consider the free energy, rather than simply potential energy (76). The enthalpic gain of forming hydrogen bonds and making favorable hydrophobic or hydrophilic contacts is compensated by a significant loss of entropy as the chain becomes more and more conformationally restricted. For simple molecules the entropic term is generally far less significant, but in the case of a folding protein the overall free energy of stabilization in the folded protein may be only a few kcal/mole, being the very small difference between large ΔH (formation of stabilizing interactions) and $T \Delta S$ (loss of entropy upon folding to the native state) terms. The conformational entropy loss for a protein, which continues to adopt a better defined three-dimensional structure, is often defined on the "per residue" basis. It can be estimated by using only the backbone entropy [the entropy loss due to side chain packing is significantly less (74)]:

$$\Delta s = s_{\mathrm{u}} - s_{\mathrm{f}} \approx k_{\mathrm{B}} \cdot \ln \left(\frac{\Omega_{\mathrm{u}}}{\Omega_{\mathrm{f}}} \right), \qquad (1\text{-}4\text{-}1)$$

where $s_{\mathrm{u,f}}$ and $\Omega_{\mathrm{u,f}}$ represent the entropies and the numbers of microstates per residue in the unfolded and natively folded forms of the protein, respectively. Estimations of Δs for small proteins at room temperature give an entropy loss on the order of tens of J/(mol·K·residue) (74). To ensure fast folding despite the unfavorable entropy change, the corresponding free energy surface (or energy landscape) must have the form of a multidimensional "funnel" (Figure 1.7), where the vertical axis (the depth of the funnel) represents the number of native contacts made (Q) or the relative free energy of the conformational space. The horizontal axis corresponds to the conformational entropy of the system. In this representation it becomes clear that as the folding chain makes more native contacts, the chain entropy is reduced along with the overall free energy. The native state resides at the bottom of the potential well, characterized by low entropy and a global free energy minimum. A funneled energy landscape is robust to both environmental changes and sequence mutations, since most potentially competing low-energy states are similar in structure (74), as represented by the multiple local minima at the bottom of the funnel in Figure 1.7.
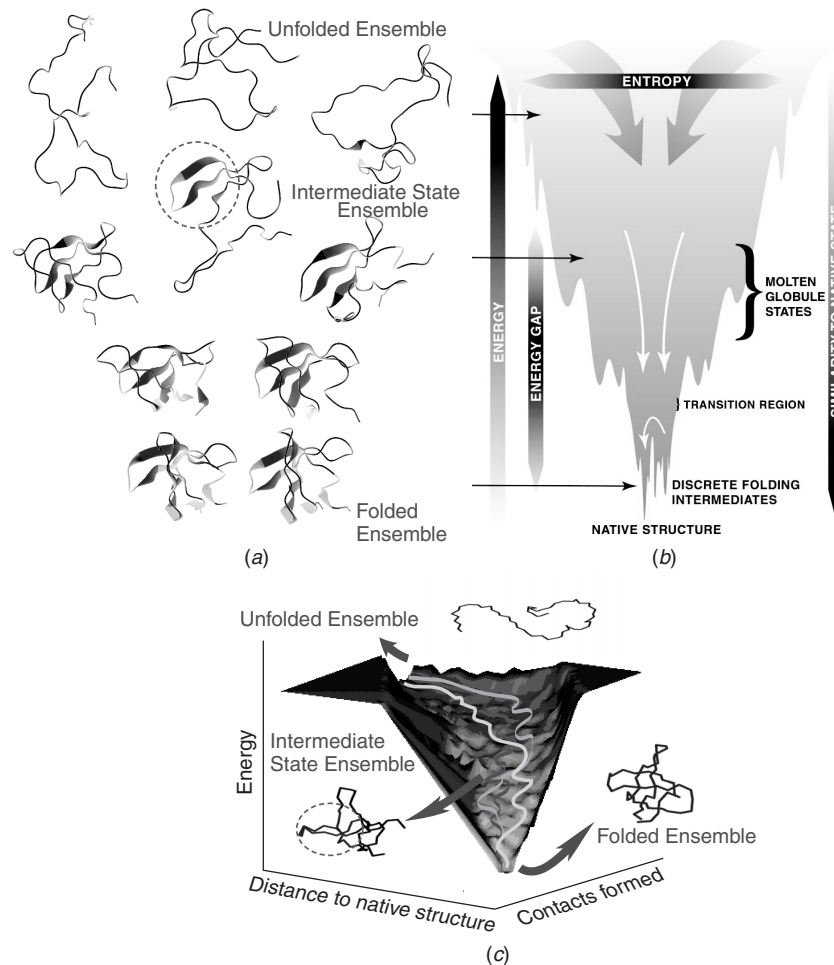
FIGURE 1.7. Schematic representation of a protein folding funnel. As the large ensemble of structures of an unfolded polypeptide compacts, forming native-like contacts through intermediate states and finally to the native state (**A**), the energy surface can be represented schematically as seen in panel (**B**). Multiplicity of folding routes is shown with different folded trajectories on the energy surface (**C**); however, asymmetry of the surface biases the trajectories toward the preferred route, which can be considered a folding pathway. Reproduced with permission from (120). © 2001 AAAS.

An important feature of the folding funnel is that its slopes are not always monotones, hence the competition between the "downhill slide" toward the native fold and the possibility of equally favorable excursions into local free energy minima, depending on the ruggedness of the energy surface. These local minima may represent transient formation of partially folded species, accumulation of intermediates, or misfolded forms, depending on the trajectory taken by the chain

along the energy surface. This general model allows for multiple pathways with no specific order of structure formation, but subtle changes in the energy surface would lead to a far more directed approach by energetically favoring particular regions of conformational space. It also provides at least one possible solution for Levinthal's problem of folding time scale. If a favored conformation is visited on the energy landscape, then it would be sufficiently stabilized to reside there for a longer time and hence restrict the conformational search for the most stable native structure.

As has been stated already, one of the most important features that distinguishes these folding funnel models from more traditional reaction coordinates is the key feature of conformational entropy. Thus, we are forced to consider each stage along a folding reaction not as a single structure but instead as a conformational ensemble. The unfolded state—the nearly flat plane at the top of the funnel—demonstrates the large number of microstates in the unfolded polypeptide chain, whereas the deep energy well of the native state may be a single structure or indeed a small ensemble of similar structures closely related in energy. Along the folding trajectory as the chain becomes more ordered, these ensembles become smaller and smaller, but throughout there is always some conformational flexibility that must be considered.

It must be stressed that these folding landscapes are purely theoretical. Based on experimental data it has proved possible to style different landscapes more as a representation than a true physical picture, showing local free energy wells for intermediate species, and alternate possible routes for parallel pathways. Even with the most powerful modern computing facilities, however, it is not yet possible to predict the folding pathway(s) of a protein, although some trajectories in a computational ensemble may appear to fit well with experimental data (77). Only with a complete understanding of the energetics involved in driving a protein to fold will we be able to computationally predict how a given protein will fold.

Semiquantitative models and experiments are revealing how the folding free energy surface is sculpted by the protein sequence and its environment. Although any downhill path from the unfolded state will eventually lead to a native conformation at the bottom of the funnel, the asymmetry (energetic heterogeneity) of the surface can bias the choice of folding routes (78). The existence of such "preferred" folding routes can be observed experimentally and interpreted in terms of folding pathways (vide supra). The sometimes conflicting demands of folding, structure, and function determine which folding pathways, if any, dominate (75).

1.4.2. Equilibrium and Kinetic Intermediates on the Energy Landscape

Under native conditions, folding of small proteins usually appears to be a highly cooperative process (79, 80). Using standard biophysical techniques, only the native and unfolded states are generally sufficiently populated to be detectable. For instance, a titration of chemical denaturant studied by circular dichroism for most proteins will generate a series of spectra that can be deconvoluted to contributions solely from the two endpoints of the unfolding reaction, namely, the

native and denatured states. This led to the belief that the folding of proteins was not only cooperative but also two-state, that is, without populated intermediate states.

In contrast, under certain conditions, significant accumulation of an intermediate conformational ensemble may occur if the free energy barrier is sufficiently high, referred to as an *equilibrium intermediate*. These species can be studied in great detail since the rate of conversion is low, allowing significant structural information to be obtained (81). Classic examples of these include the partially folded state of the apo-form of myoglobin (82), acid- and alcohol-induced *A*-state of ubiquitin (83), or the acid-induced molten globule of α-lactalbumin (84). As we have already mentioned in the preceding sections of this chapter, these equilibrium intermediate states in some cases may represent important conformations visited along the folding trajectory. Therefore, structural information about these states can give valuable clues as to the nature of the conformational search process. Transient formation or indeed accumulation of certain intermediates is usually induced in vitro by simply changing the protein's environment (e.g., by varying the solution pH, temperature, presence of chaotropes) (60, 85). This can result in significant alterations of the energy surface, decreasing the free energy of the intermediate states, and thus increasing their equilibrium population. One needs to be aware, however, that such equilibrium intermediates may differ significantly from the kinetically observed species. As pointed out recently by Fersht, an equilibrium intermediate by definition need not necessarily be on the preferred kinetic folding pathway since thermodynamic (equilibrium) parameters are independent of mechanism (58). The goal here though is to make these elusive states more amenable to study using a variety of biophysical techniques, in order to determine not only the conformational preferences of a partially folded protein, but also the possible conformations transiently visited by the native state that may be vital to its in vivo function.

Refolding of large proteins often does not conform to the simple two-state model discussed in the beginning of this section. Nevertheless, such proteins may be refolded spontaneously by rapid dilution from a chemically denatured state into native conditions. Kinetic studies of these processes have enabled detection of transient *kinetic intermediates*, which serve as "resting points" in a protein folding process. Since the energy difference between the global minimum and any of the surrounding local minima is usually quite high, the Boltzmann weight of the states that correspond to the local minima is very low. Under native conditions, kinetic intermediates become populated only transiently during refolding experiments. These species have been the focus of close experimental scrutiny, since their structure and behavior may reveal many intimate details of the protein folding process. In order to be detected these species must have characteristics different from either the native or unfolded states. For instance, the aromatic residues may experience an environment that makes them hyperfluorescent, or secondary structure may form in advance of tertiary interactions, making the intermediate detectable using stopped-flow optical techniques such as circular dichroism. Alternatively, the secondary structure may protect certain amide protons against

exchange with bulk solvent, which can be detected by NMR or mass spectrometry, as we shall see in Chapters 2 and 5. The experimental identification and characterization of kinetic intermediates has been the focus of a great deal of research over the past decade, as researchers attempt to glean the determinants of protein folding. Due to their transient nature, however, kinetic intermediates often cannot be observed directly and their properties can only be inferred from indirect measurements.

If the energy barrier separating a kinetic intermediate from the native conformation is high, a significant accumulation of such intermediates may occur. These are referred to as *kinetically trapped* or *metastable* intermediate states. Under certain conditions, excessive accumulation of metastable kinetic intermediates in the course of the folding process may trigger nonspecific interactions among them and, in extreme cases, aggregation. Likewise, aggregation can also be caused by incorrect folding (e.g., due to a sequence mutation). Aggregation processes in vivo are prevented by chaperones, a special class of proteins that bind and sequester the misfolded and partially folded polypeptides (86–89). It is important to note, however, that the chaperones only assist, but do not direct, protein folding.

1.5. PROTEIN DYNAMICS AND FUNCTION

1.5.1. Limitations of the Structure–Function Paradigm

Proteins carry out their functions by interacting with other proteins, as well as other molecules ranging from giant biopolymers (such as DNA) to small organic molecules and monatomic ligands. In all cases, protein–ligand binding is the first stage of the interaction, which can be followed by a variety of processes ranging from sophisticated chemical transformation of the ligand (e.g., enzyme catalysis) to simple release of the ligand in the presence of other cofactors (e.g., transport proteins). In this section, we only consider the main characteristics of the protein–ligand binding process.

Binding has traditionally been considered within a framework of the structure–function paradigm, a cornerstone of molecular biology for many years. It was over a hundred years ago that Fischer coined the term *lock-and-key* to emphasize the requirement for a stereochemical fit between an enzyme and its substrate in order for binding to occur (90). The limits of this view of the binding process became obvious in the middle of the last century, when a large body of newly acquired information on enzyme kinetics appeared to be in conflict with the notion that "the enzyme was a rather rigid negative of the substrate and that the substrate had to fit into this negative to react" (91). The revision of the lock-and-key theory by Koshland led to a rise of the so-called *induced fit* theory, whose major premise was that the "reaction between the enzyme and substrate can occur only after a change in protein structure induced by the substrate itself" (91). Conformational changes occurring during enzyme catalysis are relatively small scale and affect mostly the catalytic site (92). Similarly, the conformational changes occurring in
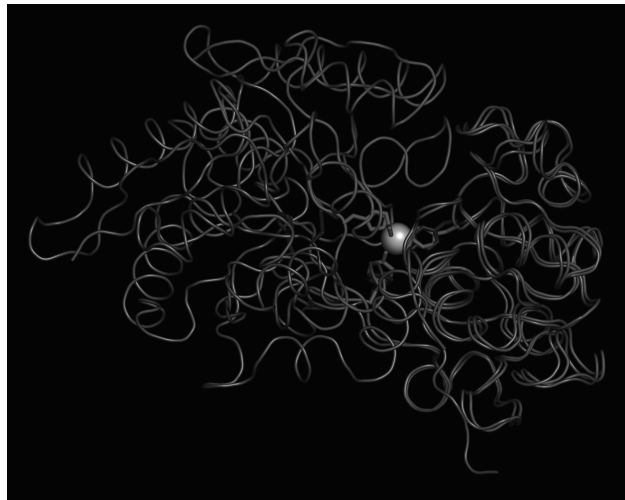
FIGURE 1.8. Superimposed crystal structures of the apo- and holo-forms of the N-lobe of human serum transferrin. Created with PyMol. © 2004 Delano Scientific LLC. See insert for color representation.

other proteins as a result of the induced-fit type binding usually affect a limited fraction of the protein structure. An example of that is ferric ion binding by the iron-transport protein transferrin, an event that results in the repositioning of two protein domains within each lobe of the protein (Figure 1.8). Although the overall effect of such repositioning is quite significant (and results in closing the cleft between the two domains), the number of affected amino acid residues does not exceed a dozen (93). More recently, numerous examples of large-scale conformational changes induced by ligand binding have been reported. The most extreme case is represented by the *intrinsically disordered* proteins, which actually lack stable structure under native conditions in the absence of the ligand (94).

The above considerations strongly suggest that structure is not the sole determinant of protein function. As elegantly put by Onuchic and Wolynes, "the twentieth century's fixation on structure catapulted folding to center stage in molecular biology. The lessons learned about folding may, in the future, increase our understanding of many functional motions and large-scale assembly processes" (78).

In the following section we will consider various aspects of protein dynamics under native conditions that may be important modulators or even determinants of function.

1.5.2. Protein Dynamics Under Native Conditions

With very few exceptions the protein structure under native conditions is not a rigid crystalline state, but undergoes local breathing motions, involving anything from side chain rotation to rearrangement of secondary structure elements relative

to each other. Although the existence of such motions within the native state of the protein can be detected with a variety of experimental techniques, their exact nature remains a subject of discussion in the literature. The commonly accepted models of local dynamics within natively folded proteins invoke the notions of a *structural fluctuation* (a localized transient unfolding affecting only few atoms within the protein) (95) or a *mobile defect* (which considers not only the emergence and dissipation of local disorder, but also the possibility of its propagation through the protein structure). Although the latter model has not enjoyed as much attention as the former, thorough theoretical considerations suggest that local perturbations of the secondary structure may in fact propagate through certain elements of the secondary structure (e.g., α-helices) in the form of a soliton (96). Alternatively, the local dynamics can be described with a *solvent penetration* model (slow diffusion of the solvent molecules into and out of the protein interior) (97). Such description is actually very similar to the mobile defect model, as applied to the integral solute–solvent system, instead of the protein molecule alone.

Above and beyond local structural fluctuations, the dynamics of proteins under native conditions is exemplified by transiently sampling alternative (higher-energy or "activated") conformations. Such activated (non-native) states are often functionally important despite their low Boltzmann weight (98, 99). An example of such behavior can be seen in cellular retinoic acid binding protein I (CRABP I), which sequesters and transports insoluble all-*trans* retinoic acid (RA) in the cytosol. The structures of the *apo-* and the *holo*-forms of this protein are very similar, consistent with the lock-and-key type of binding. However, the native structure of CRABP I provides no clue as to how the ligand gets access to the internal protein cavity, which is its binding site (Figure 1.9). Obviously, in order to provide entrance into the cavity, a fraction of the native structure has to be lost transiently, an event consistent with the notion of sampling an activated protein state. Realization of the importance of transient non-native protein structures for their function has not only greatly advanced our understanding of processes as diverse as recognition, signaling, and transport, but also has had profound practical implications, particularly for the design of drugs targeting specific proteins (100).

Many proteins use dynamics as a means of communication between different domains. This process, by which a signal such as a binding event in one domain triggers a conformational change in another domain, is known as *allostery*. The paradigm for this effect is hemoglobin, a tetrameric protein mentioned earlier. Binding a molecule of oxygen at the heme site of the α-chain induces a change in the oxygen affinity of the β-chain binding site by rearrangement of the inter-domain interactions (101, 102). Another example is the chaperone protein DnaK, which assists in preventing the misfolding of nascent chains as they emerge from the ribosome. This 70 kDa protein consists of an ATPase domain joined via a short linker region to a peptide-binding domain. Binding of ATP causes a conformational change in the peptide-binding domain that increases its affinity for substrate. Subsequent hydrolysis of nucleotide in the ATPase domain signals a

FIGURE 1.9. Crystal structures of the apo- and holo-forms of cellular retinoic acid binding protein I. Created with PyMol. © 2004 Delano Scientific LLC. See insert for color representation.

conformational change in the adjacent domain that releases the unfolded polypeptide and allows it to begin to refold. The exact mechanism by which this allosteric communication occurs is still poorly understood. It is clear, however, that it must involve dynamic events at the interdomain interface that transmits the signal between the two binding sites (103).

Another class of recently discovered proteins relies on dynamics even more heavily for function. *Intrinsically disordered* proteins are remarkable in that they appear to have very little stable folded structure in isolation, contrary to our classical view of proteins as folded species. Several hundred proteins have now been identified that contain large segments of disorder even under native conditions, and these seem to serve a wide variety of functions in vivo (104–106). A number of these are involved in activation or inhibition of transcription or translation, and while these proteins appear unstructured under native conditions, they undergo a structural transition in the presence of their cognate substrate, whether this be another protein or a recognition site on a molecule of DNA or RNA (94, 107). One important aspect of intrinsic disorder may be the necessity for this class of proteins to recognize and bind to multiple sites. Whereas a highly structured protein may only have a limited and very specific binding site available to it, one which has a highly dynamic structure should be able to adapt to a variety of different structural motifs. This intrinsic disorder phenomenon

also seems contrary to the commonly accepted view that an unstructured protein should be targeted for proteolysis, or else degradation by the proteasome. It seems this class of proteins manages to avoid such scenarios either by having regions that are sterically inaccessible or else by not containing residues that are sensitive to proteases. Indeed, one other observation is that many intrinsically disordered proteins have a relatively short lifetime in the cell: they are expressed as needed in response to a signal and then rapidly removed by degradation. This would provide an efficient mechanism to switch on or off a cellular process for only a short period of time. A larger number of proteins in the eukaryotic genome have been predicted to have disordered regions compared to prokaryotes, perhaps indicating the need for higher organisms to adjust more rapidly to environmental changes.

1.5.3. Biomolecular Dynamics and Binding from the Energy Landscape Perspective

The development of the folding funnel concept has also far-reaching consequences for our understanding of how proteins interact with each other and with other ligands. A recently introduced general scheme of protein folding and binding implies that the only difference between the two processes is chain connectivity (108, 109), namely, that monomeric protein folding represents an energy funnel for a single chain, whereas protein–protein association and peptide binding is a similar landscape but with discontinuous backbone connections. In the more general case, however, the concept can be extended to encompass the chemical nature of the ligand and the energetics of the binding process, whether it be noncovalent interaction or a chemical process as in the case of enzymatic catalysis. The energy funnel concept can be applied theoretically to describe the process by which a protein recognizes and binds another molecule. Rigid proteins that bind ligands via a lock-and-key type mechanism presumably do not require significant dynamic events so will have few local minima similar in energy to the native state. One interesting exception (CRABP) will be considered later in this section. In contrast, those proteins that utilize an induced fit binding mechanism may have a rugged energy surface characterized by a number of local minimum conformational states at the bottom of the folding/binding funnel (108).

The idea of a binding funnel has also been demonstrated computationally by Zhang et al. to explain the fast protein–ligand association rates exhibited by many proteins (110). In this model, the initial collision event is accompanied by favorable interactions to form a long-lived encounter complex that significantly limits the search process to the ligand-bound conformation. The funnel energy landscape of protein binding may be a common feature in protein–protein associations (111). Knowledge of the relative energies and structural features of the local conformational minima available to proteins is clearly key to the understanding of what makes proteins efficient in binding their physiological ligands. Likewise, an extension of the protein folding problem is to understand how protein monomers assemble as functional multimers or other macromolecular

complexes. In the cellular environment this must be an efficient process aided by dynamics and specific recognition events, all of which may be described in terms of the energetics of accessible conformational space.

All visibly different modes of binding (lock-and-key, induced fit, and binding of intrinsically disordered proteins) appear to have only quantitative differences within the framework provided by the binding funnel concept. We illustrate this point in Figures 1.10–1.12, which represent hypothetical folding funnels for proteins of each class in the absence and in the presence of their respective ligands. A protein whose ligand-binding behavior conforms to the lock-and-key type interaction (such as CRABP I considered earlier) is suggested to have an activated state whose structural features increase the rate of ligand entry into the binding site (Figure 1.10). The protein molecules sample this activated state relatively frequently due to its relatively low energy. Once the ligand enters the binding site, its interaction with the protein (multiple hydrophobic contacts in the case of CRABP I–RA interaction) increases the stability of the native conformation. Although the "visitations" to the activated state are still possible, they do not occur as frequently due to the increased energy difference between the two states. As a result, the protein can acquire the ligand relatively easily but does not release it unless the energy landscape is altered again, for example, by a competing receptor of the ligand.

A similar analysis can be carried out for proteins conforming to the induced fit type behavior (we will use the N-lobe of human serum transferrin as an example). Although the X-ray data suggest the existence of two distinct conformations of the protein depending on the presence of the ligand (open conformation for the apo-form and closed for the holo-form of the protein), there is experimental

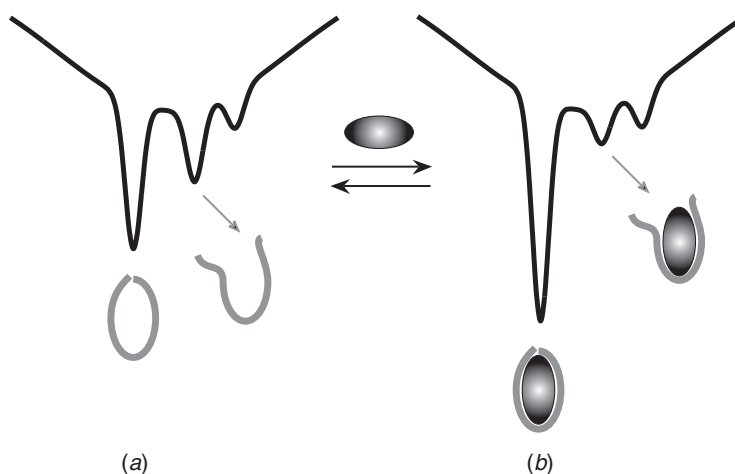
(a)   (b)

FIGURE 1.10. Schematic representations of the energy landscapes for the apo- (**A**) and holo- (**B**) forms of a protein whose ligand-binding behavior conforms to the lock-and-key type interaction.
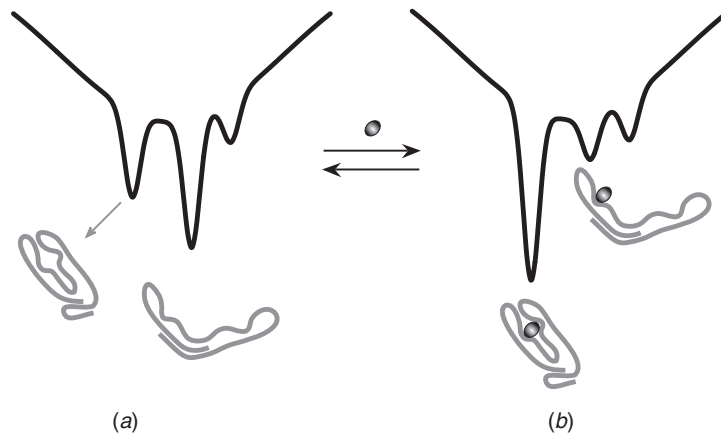
(a)   (b)

FIGURE 1.11. Schematic representations of the energy landscapes for the apo- (**A**) and holo- (**B**) forms of a protein whose ligand-binding behavior conforms to the induced fit type interaction.

(a)   (b)
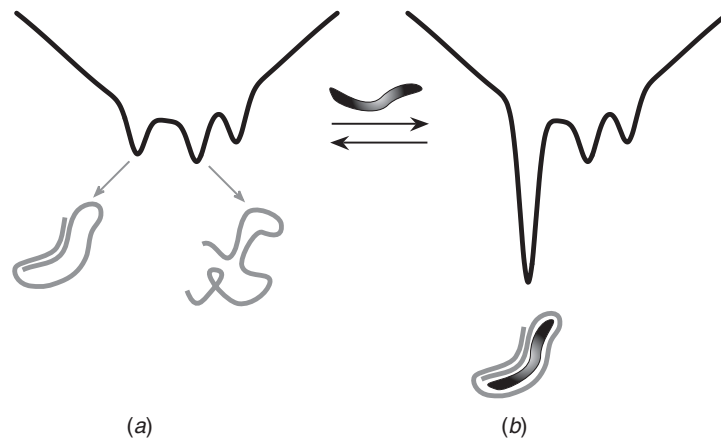
FIGURE 1.12. Schematic representations of the energy landscapes for the apo- (**A**) and holo- (**B**) forms of an intrinsically unstructured protein whose folding is induced by the ligand.

evidence suggesting that both conformations coexist in solution under equilibrium (112). The open conformation is, of course, favored in the absence of the ligand, while iron binding shifts the equilibrium toward the closed state (Figure 1.11). Such a shift is qualitatively similar to the one considered for the lock-and-key interaction. The only difference is that the protein state corresponding to the global energy minimum in the absence of the ligand becomes "downgraded" to the status of an activated state (local energy minimum) as a result of the ligand binding.

Finally, folding of an intrinsically unstructured protein in the process of ligand binding can be viewed as a preferential stabilization of one particular conformation among many available to the protein in the ligand-free form (Figure 1.12). An example of such behavior is presented by the β-chains of mammalian hemoglobins, which populate at least four different states (only one of them appears to be very close to the compact natively folded conformation) in solution in the absence of its binding partner, α-globin (113). Again, the general features of binding in this scheme appear to be very similar to those seen in the previous two examples (Figures 1.10 and 1.11), the major distinct feature being the absence of the preferred conformation in the absence of the ligand. We will revisit the physiological significance of such interactions in Chapter 11.

1.5.4. Energy Landscapes Within a Broader Context of Nonlinear Dynamics: Information Flow and Fitness Landscapes

It becomes increasingly clear that the significance of the concept of energy landscapes extends well beyond the fields of protein folding and even molecular biophysics. Recently, Huang and Ingber questioned the validity of the commonly accepted paradigm of cell regulation as a collection of pathways that link receptors with genes by asking: "Can identification of all these signaling proteins and their assignment into distinct functional pathways lead to the full understanding of developmental control and cell fate regulation?" (114). The existence of distributed information within cellular signaling, as well as the fact that a single signaling molecule may activate several genes and even produce opposite effects depending on its microenvironment led to a suggestion that the concept of *linear signaling pathways* is inappropriate. The suggestion that the signaling and regulatory pathways are not simple linear connections between receptors and genes is similar to the earlier realization of the limits of the classic concept of protein folding pathways. The paradox of signaling nonlinearity is resolved by the introduction of a concept of *cellular states*, and the switches between these states are viewed as *biological phase transitions* (114). In this new view, the cell fates are viewed as common end programs or attractors within the entire regulatory network, which can be visualized as a *potential landscape* with multiple minima. Each minimum corresponds to a certain fate of the cell (e.g., differentiation, proliferation, apoptosis). A similar idea was recently proposed as the basis of a quantitative model of carcinogenesis, in which normal cells in vivo occupy a ridge-shaped maximum in a well-defined *tissue fitness landscape*, a configuration that allows cooperative coexistence of multiple cellular populations (115).

Finally, *dynamic fitness landscapes* have proved to be a valuable concept in evolutionary biology, molecular evolution (116, 117), combinatorial optimization, and the physics of disordered systems (118). In evolutionary biology this concept is often used to visualize the relationship between genotypes (or phenotypes) and replication success, an idea initially put forward by Wright (119). An evolving population typically climbs uphill in the fitness landscape, until it reaches a local optimum (Figure 1.13), where it then remains, unless a rare mutation opens a path to a new fitness peak.
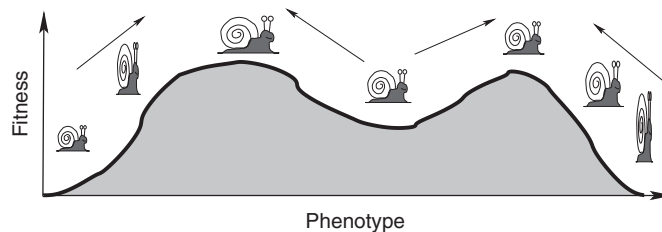
FIGURE 1.13. Schematic representation of a fitness landscape, where the fitness of a snail species depends on its shape. Mutations continuously produce variants that are selected if their fitness is larger than the fitness of the current "wild-type" snail. As a consequence, the shape of the snails changes over time until it reaches a maximum of the fitness landscape. Reproduced with permission from (118).

REFERENCES

1. Thieffry D., Sarkar S. (1998). Forty years under the central dogma. *Trends Biochem. Sci.* **23**: 312–316.

2. Nirenberg M. (2004). Historical review: deciphering the genetic code—a personal account. *Trends Biochem. Sci.* **29**: 46–54.

3. Hatfield D. L., Gladyshev V. N. (2002). How selenium has altered our understanding of the genetic code. *Mol. Cell. Biol.* **22**: 3565–3576.

4. Ibba M., Soll D. (2002). Genetic code: introducing pyrrolysine. *Curr. Biol.* **12**: R464–R466.

5. Budisa N., Minks C., Alefelder S., Wenger W., Dong F., Moroder L., Huber R. (1999). Toward the experimental codon reassignment in vivo: protein building with an expanded amino acid repertoire. *FASEB J.* **13**: 41–51.

6. Chin J. W., Cropp T. A., Anderson J. C., Mukherji M., Zhang Z., Schultz P. G. (2003). An expanded eukaryotic genetic code. *Science* **301**: 964–967.

7. Avetisov V., Goldanskii V. (1996). Mirror symmetry breaking at the molecular level. *Proc. Natl. Acad. Sci. U.S.A.* **93**: 11435–11442.

8. Podlech J. (2001). Origin of organic molecules and biomolecular homochirality. *Cell. Mol. Life Sci.* **58**: 44–60.

9. Warshel A., Papazyan A. (1998). Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Curr. Opin. Struct. Biol.* **8**: 211–217.

10. Schutz C. N., Warshel A. (2001). What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins* **44**: 400–417.

11. Harvey S. C. (1989). Treatment of electrostatic effects in macromolecular modeling. *Proteins* **5**: 78–92.

12. Daune M. (1999). *Molecular Biophysics: Structures in Motion*. New York: Oxford University Press.

13. Mirsky A. E., Pauling L. (1936). On the structure of native, denatured, and coagulated proteins. *Proc. Natl. Acad. Sci. U.S.A.* **22**: 439–447.

14. Pauling L., Corey R. B., Branson H. R. (1951). The structure of proteins—2 hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* **37**: 205–211.

15. Ramachandran G. N., Ramakrishnan C., Sasisekharan V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**: 95–99.

16. Tanford C. (1978). Hydrophobic effect and organization of living matter. *Science* **200**: 1012–1018.

17. Scheraga H. A. (1998). Theory of hydrophobic interactions. *J. Biomol. Struct. Dyn.* **16**: 447–460.

18. Southall N. T., Dill K. A., Haymet A. D. J. (2002). A view of the hydrophobic effect. *J. Phys. Chem. B* **106**: 521–533.

19. Widom B., Bhimalapuram P., Koga K. (2003). The hydrophobic effect. *Phys. Chem. Chem. Phys.* **5**: 3085–3093.

20. Hartley G. S. (1936). *Aqueous Solutions of Paraffin-Chain Salts*. Paris: Hermann & Cie.

21. Tanford C. (1979). Interfacial free-energy and the hydrophobic effect. *Proc. Natl. Acad. Sci. U.S.A.* **76**: 4175–4176.

22. Hildebrand J. H. (1968). A criticism of term hydrophobic bond. *J. Phys. Chem.* **72**: 1841–1842; Nemethy G., Scheraga H. A., Kauzmann W. *Ibid.*: 1842.

23. Frank H. S., Evans M. W. (1945). Free volume and entropy in condensed systems. III. Entropy in binary liquid mixtures; partial molar entropy in dilute solutions; structure and thermodynamics in aqueous electrolytes. *J. Chem. Phys.* **13**: 507–532.

24. Tanford C. (1997). How protein chemists learned about the hydrophobic factor. *Protein Sci.* **6**: 1358–1366.

25. Chothia C., Hubbard T., Brenner S., Barns H., Murzin A. (1997). Protein folds in all-α and all-β classes. *Annu. Rev. Biophys. Biomol. Struct.* **26**: 597–627.

26. Zhang C., DeLisi C. (1998). Estimating the number of protein folds. *J. Mol. Biol.* **284**: 1301–1305.

27. Govindarajan S., Recabarren R., Goldstein R. A. (1999). Estimating the total number of protein folds. *Proteins* **35**: 408–414.

28. Hou J., Sims G. E., Zhang C., Kim S. -H. (2003). A global representation of the protein fold space. *Proc. Natl. Acad. Sci. U.S.A.* **100**: 2386–2390.

29. Caetano-Anolles G., Caetano-Anolles D. (2003). An evolutionarily structured universe of protein architecture. *Genome Res.* **13**: 1563–1571.

30. Rost B. (1997). Protein structures sustain evolutionary drift. *Fold. Des.* **2**: S19–S24.

31. Wood T. C., Pearson W. R. (1999). Evolution of protein sequences and structures. *J. Mol. Biol.* **291**: 977–995.

32. Zhang C., DeLisi C. (2001). Protein folds: molecular systematics in three dimensions. *Cell. Mol. Life Sci.* **58**: 72–79.

33. Denton M. J., Marshall C. J., Legge M. (2002). The protein folds as platonic forms: new support for the pre-Darwinian conception of evolution by natural law. *J. Theor. Biol.* **219**: 325–342.

34. Epstein C. J., Goldberger R. F., Anfinsen C. B. (1963). Genetic control of tertiary protein structure—studies with model systems. *Cold Spring Harbor Symp. Quant. Biol.* **28**: 439–449.

35. Dawkins R. (1996). *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe Without Design*. New York: Norton.

36. Pennock R. T. (2001). *Intelligent Design Creationism and Its Critics: Philosophical, Theological, and Scientific Perspectives*. Cambridge, MA: MIT Press.

37. Levinthal C. (1968). Are there pathways for protein folding? *J. Chim. Phys. Phys. Chim. Biol.* **65**: 44–45.

38. Dill K. (1999). Polymer principles and protein folding. *Protein Sci.* **8**: 1166–1180.

39. Radford S. E., Dobson C. M. (1999). From computer simulations to human disease: emerging themes in protein folding. *Cell* **97**: 291–298.

40. Jaenicke R. (1998). Protein self-organization in vitro and in vivo: partitioning between physical biochemistry and cell biology. *Biol. Chem.* **379**: 237–243.

41. Jaenicke R. (1995). Folding and association versus misfolding and aggregation of proteins. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **348**: 97–105.

42. Rochet J. -C., Lansbury P. T. (2000). Amyloid fibrillogenesis: themes and variations. *Curr. Opin. Struct. Biol.* **10**: 60–68.

43. Uversky V. N. (2003). Protein folding revisited. A polypeptide chain at the folding–misfolding–nonfolding cross-roads: which way to go? *Cell. Mol. Life Sci.* **60**: 1852–1871.

44. Dobson C. M. (2004). Principles of protein folding, misfolding and aggregation. *Semin. Cell Dev. Biol.* **15**: 3–16.

45. Carrell R. W., Lomas D. A. (1997). Conformational disease. *Lancet* **350**: 134–138.

46. Bellotti V., Mangione P., Stoppini M. (1999). Biological activity and pathological implications of misfolded proteins. *Cell. Mol. Life Sci.* **55**: 977–991.

47. Thompson A. J., Barrow C. J. (2002). Protein conformational misfolding and amyloid formation: characteristics of a new class of disorders that include Alzheimer's and Prion diseases. *Curr. Med. Chem.* **9**: 1751–1762.

48. Bucciantini M., Giannoni E., Chiti F., Baroni F., Formigli L., Zurdo J., Taddei N., Ramponi G., Dobson C. M., Stefani M. (2002). Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* **416**: 507–511.

49. Caughey B., Lansbury P. T. (2003). Protofibrils, pores, fibrils, and neurodegeneration: separating the responsible protein aggregates from the innocent bystanders. *Annu. Rev. Neurosci.* **26**: 267–298.

50. Wales D. J., Scheraga H. A. (1999). Global optimization of clusters, crystals, and biomolecules. *Science* **285**: 1368–1372.

51. Schonbrun J., Wedemeyer W. J., Baker D. (2002). Protein structure prediction in 2002. *Curr. Opin. Struct. Biol.* **12**: 348–354.

52. Whisstock J. C., Lesk A. M. (2003). Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* **36**: 307–340.

53. Price N. C. (2000). Conformational issues in the characterization of proteins. *Biotechnol. Appl. Biochem.* **31**: 29–40.

54. Smith L. J., Fiebig K. M., Schwalbe H., Dobson C. M. (1996). The concept of a random coil. Residual structure in peptides and denatured proteins. *Fold. Des.* **1**: R95–R106.

55. Ptitsyn O. B. (1995). Molten globule and protein folding. *Adv. Protein Chem.* **47**: 83–229.

56. Uversky V. N. (1998). How many molten globules states exist? *Biofizika* **43**: 416–421.

57. Clarke J., Fersht A. R. (1996). An evaluation of the use of hydrogen exchange at equilibrium to probe intermediates on the protein folding pathway. *Fold. Des.* **1**: 243–254.

58. Clarke J., Itzhaki L. S., Fersht A. R. (1997). Hydrogen exchange at equilibrium: a short cut for analysing protein-folding pathways? *Trends Biochem. Sci.* **22**: 284–287.

59. Jennings P. A., Wright P. E. (1993). Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Science* **262**: 892–896.

60. Bai Y., Sosnick T. R., Mayne L., Englander S. W. (1995). Protein folding intermediates: native state hydrogen exchange. *Science* **269**: 192–196.

61. Bai Y. (1999). Equilibrium amide hydrogen exchange and protein folding kinetics. *J. Biomol. NMR* **15**: 65–70.

62. Parker M. J., Marqusee S. (2001). A kinetic folding intermediate probed by native state hydrogen exchange. *J. Mol. Biol.* **305**: 593–602.

63. Creighton T. E. (1984). Pathways and mechanisms of protein folding. *Adv. Biophys.* **18**: 1–20.

64. Creighton T. E. (1988). Toward a better understanding of protein folding pathways. *Proc. Natl. Acad. Sci. U.S.A.* **85**: 5082–5086.

65. Gunasekaran K., Eyles S. J., Hagler A. T., Gierasch L. M. (2001). Keeping it in the family: folding studies of related proteins. *Curr. Opin. Struct. Biol.* **11**: 83–93.

66. Matthews C. R. (1993). Pathways of protein folding. *Annu. Rev. Biochem.* **62**: 653–683.

67. Daggett V., Fersht A. R. (2003). Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.* **28**: 18–25.

68. Baldwin R. L. (1995). The nature of protein folding pathways: the classical versus the new view. *J. Biomol. NMR* **5**: 103–109.

69. Dill K. A., Chan H. S. (1997). From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**: 10–19.

70. Pande V. S., Grosberg A., Tanaka T., Rokhsar D. S. (1998). Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.* **8**: 68–79.

71. Bryngelson J. D., Onuchic J. N., Socci N. D., Wolynes P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**: 167–195.

72. Onuchic J. N., Luthey-Schulten Z., Wolynes P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**: 545–600.

73. Brooks C. L. III, Gruebele M., Onuchic J. N., Wolynes P. G. (1998). Chemical physics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **95**: 11037–11038.

74. Plotkin S. S., Onuchic J. N. (2002). Understanding protein folding with energy landscape theory. Part I: Basic concepts. *Q. Rev. Biophys.* **35**: 111–167.

75. Gruebele M. (2002). Protein folding: the free energy surface. *Curr. Opin. Struct. Biol.* **12**: 161–168.

76. Dinner A. R., Sali A., Smith L. J., Dobson C. M., Karplus M. (2000). Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* **25**: 331–339.

77. Fersht A. R., Daggett V. (2002). Protein folding and unfolding at atomic resolution. *Cell* **108**: 573–582.

78. Onuchic J. N., Wolynes P. G. (2004). Theory of protein folding. *Curr. Opin. Struct. Biol.* **14**: 70–75.

79. Jackson S. E. (1998). How do small single-domain proteins fold? *Fold. Des.* **3**: R81–R91.

80. Myers J. K., Oas T. G. (2002). Mechanisms of fast protein folding. *Annu. Rev. Biochem.* **71**: 783–815.

81. Ptitsyn O. B. (1994). Kinetic and equilibrium intermediates in protein folding. *Protein Eng.* **7**: 593–596.

82. Eliezer D., Jennings P. A., Dyson H. J., Wright P. E. (1997). Populating the equilibrium molten globule state of apomyoglobin under conditions suitable for structural characterization by NMR. *FEBS Lett.* **417**: 92–96.

83. Brutscher B., Bruschweiler R., Ernst R. R. (1997). Backbone dynamics and structural characterization of the partially folded A state of ubiquitin by ^{1}H, ^{13}C, and ^{15}N nuclear magnetic resonance spectroscopy. *Biochemistry* **36**: 13043–13053.

84. Koshiba T., Yao M., Kobashigawa Y., Demura M., Nakagawa A., Tanaka I., Kuwajima K., Nitta K. (2000). Structure and thermodynamics of the extraordinarily stable molten globule state of canine milk lysozyme. *Biochemistry* **39**: 3248–3257.

85. Bai Y., Englander J. J., Mayne L., Milne J. S., Englander S. W. (1995). Thermodynamic parameters from hydrogen exchange measurements. *Methods Enzymol.* **259**: 344–356.

86. Fink A. L. (1999). Chaperone-mediated protein folding. *Physiol. Rev.* **79**: 425–449.

87. Hartl F. U., Hayer-Hartl M. (2002). Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **295**: 1852–1858.

88. Trombetta E. S., Parodi A. J. (2003). Quality control and protein folding in the secretory pathway. *Annu. Rev. Cell Dev. Biol.* **19**: 649–676.

89. Mogk A., Bukau B. (2004). Molecular chaperones: structure of a protein disaggregase. *Curr. Biol.* **14**: R78–R80.

90. Fischer E. (1894). Einfluss der configuration auf die wirkung derenzyme. *Ber. Dtsch. Chem. Ges.* **27**: 2985–2993.

91. Koshland D. E. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **44**: 98–104.

92. Koshland D. E. (1998). Conformational changes: how small is big enough? *Nat. Med.* **4**: 1112–1114.

93. Jeffrey P. D., Bewley M. C., MacGillivray R. T., Mason A. B., Woodworth R. C., Baker E. N. (1998). Ligand-induced conformational change in transferrins: crystal structure of the open form of the N-terminal half-molecule of human transferrin. *Biochemistry* **37**: 13978–13986.

94. Dunker A. K., Brown C. J., Lawson J. D., Iakoucheva L. M., Obradovic Z. (2002). Intrinsic disorder and protein function. *Biochemistry* **41**: 6573–6582.

95. Maity H., Lim W. K., Rumbley J. N., Englander S. W. (2003). Protein hydrogen exchange mechanism: local fluctuations. *Protein Sci* **12**: 153–160.

96. d'Ovidio F., Bohr H. G., Lindgard P. A. (2003). Solitons on H bonds in proteins. *J. Phys. Condens. Mat.* **15**: S1699–S1707.

97. Miller D. W., Dill K. A. (1995). A statistical mechanical model for hydrogen exchange in globular proteins. *Protein Sci.* **4**: 1860–1873.

98. Tsai C. D., Ma B., Kumar S., Wolfson H., Nussinov R. (2001). Protein folding: binding of conformationally fluctuating building blocks via population selection. *Crit. Rev. Biochem. Mol. Biol.* **36**: 399–433.

99. Papoian G. A., Wolynes P. G. (2003). The physics and bioinformatics of binding and folding — an energy landscape perspective. *Biopolymers* **68**: 333–349.

100. Carlson H. A. (2002). Protein flexibility and drug design: how to hit a moving target. *Curr. Opin. Chem. Biol.* **6**: 447–452.

101. Perutz M. F. (1972). Stereochemical mechanism of cooperative effects in haemoglobin. *Biochimie* **54**: 587–588.

102. Perutz M. F., Wilkinson A. J., Paoli M., Dodson G. G. (1998). The stereochemical mechanism of the cooperative effects in hemoglobin revisited. *Annu. Rev. Biophys. Biomol. Struct.* **27**: 1–34.

103. Pellecchia M., Montgomery D. L., Stevens S. Y., Vander Kooi C. W., Feng H. P., Gierasch L. M., Zuiderweg E. R. (2000). Structural insights into substrate binding by the molecular chaperone DnaK. *Nat. Struct. Biol.* **7**: 298–303.

104. Churchill M. E., Travers A. A. (1991). Protein motifs that recognize structural features of DNA. *Trends Biochem. Sci.* **16**: 92–97.

105. Sigler P. B. (1988). Transcriptional activation. Acid blobs and negative noodles. *Nature* **333**: 210–212.

106. Frankel A. D., Kim P. S. (1991). Modular structure of transcription factors: implications for gene regulation. *Cell* **65**: 717–719.

107. Iakoucheva L. M., Brown C. J., Lawson J. D., Obradovic Z., Dunker A. K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**: 573–584.

108. Tsai C. -J., Xu D., Nussinov R. (1998). Protein folding via binding and vice versa. *Folding and Design* **3**: R71–R80.

109. Tsai C. -J., Kumar S., Ma B., Nussinov R. (1999). Folding funnels, binding funnels, and protein function. *Protein Sci.* **8**: 1181–1190.

110. Zhang C., Chen J., DeLisi C. (1999). Protein–protein recognition: exploring the energy funnels near the binding sites. *Proteins* **34**: 255–267.

111. Tovchigrechko A., Vakser I. A. (2001). How common is the funnel-like energy landscape in protein–protein interactions? *Protein Sci.* **10**: 1572–1583.

112. Baker H. M., Anderson B. F., Baker E. N. (2003). Dealing with iron: common structural principles in proteins that transport iron and heme. *Proc. Natl. Acad. Sci. U.S.A.* **100**: 3579–3583.

113. Griffith W. P., Kaltashov I. A. (2003). Highly asymmetric interactions between globin chains during hemoglobin assembly revealed by electrospray ionization mass spectrometry. *Biochemistry* **42**: 10024–10033.

114. Huang S., Ingber D. E. (2000). Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks. *Exp. Cell Res.* **261**: 91–103.

115. Gatenby R. A., Vincent T. L. (2003). An evolutionary model of carcinogenesis. *Cancer Res.* **63**: 6212–6220.

116. Drossel B. (2001). Biological evolution and statistical physics. *Adv. Phys.* **50**: 209–295.

117. Wilke C. O., Ronnewinkel C., Martinetz T. (2001). Dynamic fitness landscapes in molecular evolution. *Phys. Rep.* **349**: 395–446.

118. Reidys C. M., Stadler P. F. (2002). Combinatorial landscapes. *SIAM Rev.* **44**: 3–54.

119. Wright S. (1967). "Surfaces" of selective value. *Proc. Natl. Acad. Sci. U.S.A.* **58**: 165–172.

120. Brooks C. L., Onuchic J. N., Wales D. J. (2001). Statistical thermodynamics. Taking a walk on a landscape. *Science* **293**: 612–613.