Annu. Rev. Biochem. 1997. 66:549–79
Copyright © 1997 by Annual Reviews Inc. All rights reserved

PROTEIN FOLDING: The Endgame

Michael Levitt, Mark Gerstein,[†] Enoch Huang, S. Subbiah,[*] and Jerry Tsai

Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305; [†]Molecular Biophysics and Biochemistry, Yale University, Bass Center, New Haven, Connecticut 06520; [*]Wistar Institute, 3601 Spruce Street, Philadelphia, Pennsylvania 19104, and Bioinformatics Center, National University of Singapore, Kent Ridge, Singapore

KEY WORDS: protein folding, packing, side-chains

ABSTRACT

The last stage of protein folding, the "endgame," involves the ordering of amino acid side-chains into a well defined and closely packed configuration. We review a number of topics related to this process. We first describe how the observed packing in protein crystal structures is measured. Such measurements show that the protein interior is packed exceptionally tightly, more so than the protein surface or surrounding solvent and even more efficiently than crystals of simple organic molecules. In vitro protein folding experiments also show that the protein is close-packed in solution and that the tight packing and intercalation of side-chains is a final and essential step in the folding pathway. These experimental observations, in turn, suggest that a folded protein structure can be described as a kind of three-dimensional jigsaw puzzle and that predicting side-chain packing is possible in the sense of solving this puzzle. The major difficulty that must be overcome in predicting side-chain packing is a combinatorial "explosion" in the number of possible configurations. There has been much recent progress towards overcoming this problem, and we survey a variety of the approaches. These approaches differ principally in whether they use ab initio (physical) or more knowledge-based methods, how they divide up and search conformational space, and how they evaluate candidate configurations (using scoring functions). The accuracy of side-chain prediction depends crucially on the (assumed) positioning of the main-chain. Methods for predicting main-chain conformation are, in a sense, not as developed as that for side-chains. We conclude by surveying these methods. As with side-chain prediction, there are a great variety of approaches, which differ in how they divide up and search space and in how they score candidate conformations.

0066-4154/97/0701-0549$08.00

CONTENTS

INTRODUCTION

The endgame of protein folding refers to the final stage in the folding process. It is believed that at this point in the process the overall fold has already been determined and the side-chains are close to their final positions. The previous steps in the folding process, especially those that determine the shape of the overall fold, are thought to be greatly, if not completely, dictated by hydrophobic interactions (1–3). However, here we argue that the endgame transition to the native structure is governed by somewhat different interactions: tight close-packed contacts between amino acid side-chains. The creation of these contacts has been compared to crystallization (4). Clearly, such tight packing is related to the most important characteristic of native protein structures: their unique and precisely determined yet highly complex three-dimensional shapes. The packing process is likely to be energetically difficult as side-chains prefer to be disordered. The process, therefore, will have a high activation barrier and will be slow.

Packing as a phenomenon is easily visualized and is commonplace in everyday experience. It is dominated by a simple universal energy term, the strong

repulsion between atoms that approach each other too closely. Packing is a short-range phenomenon, which allows a more local treatment considering only surrounding neighbors. Richards was one of the first researchers to emphasize the importance of close-packing in protein structure (5, 6), and his point of view is becoming increasingly accepted.

With its focus on packing, this review on protein folding considers both experimental work and theory, and is by necessity selective, given the large body of theoretical and computational work. Attention is focused on the type of packing observed in proteins and on the prediction of packing for side-chains. We also focus on the more difficult problem of generating main-chain conformations close enough to make side-chain packing predictions possible.

We have deliberately not dealt with certain related topics including molecular dynamics (MD) simulation and homology modeling. Realistic MD simulations of protein folding or unfolding in solution are not reviewed, in spite of the recent work in this field (7–9). Homology modeling is also not reviewed, in spite of the close connection to side-chain packing using a main-chain "borrowed" from a protein with a homologous sequence. The intent here is to concentrate more on basic principles rather than on applications. Furthermore, homology modeling has been recently reviewed (10, 11).

This review is divided into several sections. The first section deals with the observed packing in protein structures as determined by X-ray crystallography and shows that proteins are more tightly packed than almost any other organic matter. The second section extends the review of experimental work to solution studies by relating close-packing to stability and considering how such close-packing arises during the folding process. The third section shows how close-packing has led to the effective solution of the side-chain prediction problem when a sufficiently native-like main-chain conformation is known. The fourth section considers how to generate sufficiently accurate main-chain conformations, primarily by searching large spaces of possible conformations with appropriate energy functions.

WHAT CAN WE LEARN FROM X-RAY STRUCTURES?

The best source of information on packing in protein molecules comes from the hundreds of highly refined high-resolution protein structures that have been determined over the past three decades. These structures show a high degree of order in all the residues, except occasionally those on the surface of the protein.

How Is Packing Characterized?

The packing efficiency of a given atom is defined as the ratio of the volume of its van der Waals (VDW) envelope to the amount of space it actually occupies

(5, 12, 13). This simple definition masks considerable complexity. First of all, how does one determine the volume of the VDW envelope (14)? This obviously requires knowledge of what the VDW radii of atoms are, a subject on which there is no universal agreement (12, 15), particularly for water molecules and polar atoms (16, 17). Second, how does one determine how much space an atom occupies? Or, equivalently, how much additional "cavity" volume should be associated with a particular atom in addition to its envelope volume? These latter questions can be addressed by various geometric constructions, discussed in the following section.

The absolute packing efficiency of an atom is most useful in a comparative sense, e.g. when comparing equivalent atoms in different parts of a protein structure. In calculating the ratio of packing efficiencies, the VDW envelope volume remains the same and cancels. One is left with just the ratio of space an atom occupies in one environment to the space it occupies in another.

VORONOI CONSTRUCTION Voronoi volume calculations are geometrically rigorous methods that determine how much space an atom occupies. These calculations were originally developed by Voronoi (18). They were first applied to molecular systems by Bernal & Finney (19) and to proteins by Richards (5). Since then they have been used successfully in the calculation of standard volumes of protein residues, in characterizing protein-protein interactions, in understanding protein motions, and in analyzing cavities in protein structure (6, 12, 20–25). They have also been used in the analysis of liquids (26, 27), and the faces of Voronoi polyhedra have been used to characterize protein accessibility and to assess the fit of docked substrates in enzymes (28, 29).

The Voronoi procedure allocates all space amongst a collection of atoms. Each atom is surrounded by a polyhedron and allocated the space within it. The faces of Voronoi polyhedra are formed by constructing dividing planes perpendicular to the interatomic vectors between atoms, and the edges of the polyhedra result from the intersection of these planes.

The Voronoi procedure requires the location of all neighboring atoms. This is possible in the protein core, but on the protein surface many of the neighbors of a protein atom are water molecules, which are often not well localized in crystal structures. A variety of approaches have been developed to deal with this difficulty. The simplest is to surround the protein with a shell of water molecules generated on a regular grid (5). It is also possible to use predefined boundary shapes (such as the snub cube) to truncate the "open" polyhedra at the protein surface (23). This sort of truncation can be smoothly and rigorously achieved by using a particular generalization of the Voronoi construction called the alpha-shape (30, 31). In MD simulations employing periodic boundary conditions, all atoms are completely surrounded by solvent, circumventing this problem (17, 27).

OTHER CONSTRUCTIONS A number of methods for measuring volumes and packing are not based on Voronoi polyhedra (6). Connolly developed a method for the determination of volumes based on the direct integration of the space inside of the molecular surface envelope (32–34). Gregoret & Cohen (35) developed a simplified way of evaluating the packing in a structure at a residue level, rather than at the atomic level.

All the other approaches have concentrated on the explicit identification and measurement of cavities in protein structures (36–42). The advantage of cavity identification algorithms is that the exact location of cavities is often of great interest. However, because the association between a particular cavity and a particular protein atom is somewhat arbitrary, one cannot directly calculate packing efficiencies for individual atoms as with the Voronoi procedure. Another difficulty with cavity identification algorithms is that many of these algorithms model cavities in terms of idealized spherical shapes. Such modeling does not allow a complete partition of space; after the volumes of the spherical cavities and the atoms' VDW envelopes are accounted for, there is still leftover space.

How Tightly Packed Is the Protein Core?

Packing calculations on protein structure were done first by Richards more than two decades ago (5) and then soon after by others (20, 21). These initial calculations revealed some important facts about protein structure. First, in the protein core, atoms and residues of a given type have a roughly constant (or invariant) volume because the atoms inside proteins are packed together tightly, with the interior of the protein better resembling a close-packed solid than a liquid or gas. This high packing efficiency ratio of internal protein atoms is roughly what is expected for the close-packing of hard spheres (0.74).

More recent calculations measuring the packing in proteins (25) have shown that the packing inside proteins is somewhat tighter than observed initially ($\sim$4%) and that the overall packing efficiency of atoms in the protein core is greater than in crystals of organic molecules. When molecules are packed this tightly, small changes in packing efficiency are quite significant. In this regime, the limitation on close-packing is hard-core repulsion, so even a small change is quite substantial energetically. Furthermore, Richards & Lim (13) pointed out that the number of allowable configurations that a collection of atoms can adopt without hard-core overlap drops off very quickly as these atoms approach the close-packed limit.

The exceptionally tight packing in the protein core seems to require a precise jigsaw puzzle–like fitting together of the residues inside proteins. This appears to be true for the majority of atoms inside proteins (34). However, there are exceptions, and some studies have focused on these, showing how the packing inside proteins is punctuated by defects or cavities (39, 42, 43). If these defects are large enough, they can accommodate buried water molecules (44–46).

Researchers using highly simplified two-dimensional lattice models to study protein structure have pointed out that tight packing in the protein core may drive or force the formation of secondary structures (2, 47, 48). This conjecture has been tested on somewhat more realistic off-lattice models of protein structure (49, 50). The results have been mixed in the sense that these models do observe high packing density driving the formation of secondary structure but to a much lesser degree than in the lattice models.

How Tightly Packed Are Other Parts of Proteins?

THE SURFACE Measuring the packing efficiency inside of the protein core provides a good standard, and a number of other studies have compared this efficiency to that in other parts of the protein. The most obvious thing to compare with the protein inside is the protein outside, or surface. This comparison is particularly interesting from a packing perspective because the protein surface is covered by water, which is known to be packed much less tightly than protein and in a distinctly different fashion [the tetrahedral packing geometry of water molecules gives a packing efficiency ratio of $\sim$0.34, less than half that of hexagonal close-packed solids (51)].

Calculations based on crystal structures and simulations have shown that the protein surface has an intermediate packing, being packed less tightly than the core but not as loosely as liquid water (15, 17). One can understand the packing being looser at the surface than in the core in terms of a simple trade-off between hydrogen bonding and close-packing. In the absence of interactions other than van der Waals attractions and repulsions, liquids (and solids) tend to pack closely, and the geometry of their interaction can be described simply in terms of a simple hard-sphere (i.e. billiard-ball) model (52). However, if there are also highly directional interactions, such as the hydrogen bond in water, the situation is more complicated. Often the close-packing has to be explicitly traded off to maintain hydrogen bonding. This trade-off can be visualized in simulations of the packing in simple toy systems (53–55).

An important aspect of the looser packing at the protein surface is how this packing is expected to change when the protein surface binds to another molecule, particularly another protein. Calculations measuring the packing in protein-protein interfaces have been done, such as those in antibody-antigen and protease-inhibitor complexes (56, 57). These calculations have shown that the packing at protein-protein interfaces is roughly comparable to that in the protein interior and is tighter than the packing usually observed at the surface. Thus, the formation of a close-packed interface may be a driving force in docking. Simple shape complementarity (in the sense of a close-packed jigsaw puzzle) is an integral part of many docking programs (58–61).

INTERNAL INTERFACES A comparison of the packing at various internal interfaces inside of proteins, particularly at domain-domain interfaces, is also interesting. Such comparisons are often closely coupled with analysis of protein flexibility.

It has been argued that motion is possible across a close-packed interface such that the close-packing is maintained throughout the motion. To prevent the atoms from bumping into one another, the motion has to be fairly small and parallel to the plane of the interface. There cannot be large torsion angle changes, so side-chains maintain the same rotamer configuration (62). A large motion is achieved by concatenating many of these small motions at many different interfaces. This sort of small, sliding motion has been dubbed "shear motion" (63, 64), and it has been carefully documented in numerous cases (65, 66; see 64 for a list). Moreover, physical studies have shown that a folded protein does not have a single perfectly defined conformation (67). Rather, it has some intrinsic flexibility and can readily jump among many nearly energetically identical micro-states without significantly changing its packing. This sort of small-scale flexibility is what makes shear motions possible.

Following a somewhat different line of reasoning, it has also been proposed that certain interfaces may be particularly mobile, precisely because they contain defects and are not close-packed. This idea was suggested in the 1970s (22). Since then a number of workers have noted that there are relatively more cavities at interdomain interfaces (36, 68) than elsewhere on protein interiors. Hubbard & Argos (68), in particular, claim that these cavities have a functional role in the mechanisms of protein movements.

Packing is also expected to be important in protein motions involving hinges. Numerous studies have emphasized how critical the packing at the base of the hinge is [in the same sense that the "packing" at the base of a door hinge determines how easily the door can close (69–73)]. Hinge motions often involve creating a new protein-protein interface (e.g. a new domain-domain interface is formed during hinged domain closure). Calculations have shown that these interfaces are close-packed in the same manner as the interfaces involved in protein-protein recognition (72). This conclusion suggests that the formation of a new close-packed interface may be a driving force for hinge motions.

WHAT DOES EXPERIMENT HAVE TO SAY?

Clearly proteins are close-packed in the crystal state. Such close-packing is also seen in protein structures determined in solution by nuclear magnetic resonance (NMR), but these proteins are generally rather small ($<$100 residues) and do not always have a large core region. This section further considers proteins in solution. We first examine whether close-packing stabilizes proteins in solution

and then review experimental work on how proteins fold to achieve such close packing.

Proteins Are Well Packed in Solution

In solution, volumetric studies of both amino acids (74, 75) and whole proteins (76–81) have been common. The most recent study (82) is quite comprehensive, covering 15 proteins within a temperature range of 18–25°C. The results show that studying whole proteins is more accurate than measuring the individual amino acid volumes in solution. By studying whole proteins, the authors derive some useful relationships on the basis of the molecular weight of a protein without prior knowledge of the crystallographic data. As rough estimates, the van der Waals volume V_w, the molecular volume V_m, and the accessible surface area S_a can all be related to the molecular weight M_r as follows (in Å^3): $V_w = [100(\pm 300)] + [0.77(\pm 0.01)]M_r$; $V_m = [1200(\pm 500)] + [1.04(\pm 0.02)]M_r$; and $S_a = -[1200(\pm 200)] + [14.5(\pm 0.25)]M_r^{2/3}$. The authors also show that packing efficiencies are relatively constant between 0.72 and 0.78. This range is very similar to the previously mentioned packing efficiencies computed from protein structures solved by X-ray crystallography (6, 20, 25). The fact that packing efficiencies are not limited to some finite value suggests that the packing in individual proteins is not so rigid as the jigsaw model would have us believe (82a). With extensive studies of T4 lysozyme packing mutants, Matthews and coworkers (for a review see 82b) have shown that the protein's backbone accommodates changes to the size of the protein core. While losing some stability, these lysozyme mutants are still chemically active. Therefore, proteins possess a well-packed, plastic interior, meaning that the core can tolerate a certain amount of variation in packing density.

Good Packing Leads to Greater Stability

Improving the packing of the protein interior has recently become a method for increasing stability (13, 83, 84). Nature uses this principle in the design of thermostable proteins (85), and several groups have successfully applied it to protein design. Thus far, researchers have been able to create more stable proteins by intentionally increasing the packing efficiency for ribonuclease H1 (86), T4 lysozyme (87), and λ-repressor (88). Most recently, Munson et al (89) have re-engineered the internal packing of the four-helix-bundle protein, Rop. Their results further support the idea that increasing the core packing efficiency can increase stability; however, it has also been found that sometimes the increased stability caused a decrease in function. In a related experiment, Ramachandran & Udgaonkar (90) added significant nonpolar volume to the core of the protein barstar by chemically modifying its two free cysteines. They showed that the change caused an increase in protein stability without a decrease

in activity or major alteration in structure as measured by circular dichroism (CD). Until the crystal structure of the altered barstar is solved, they reason that this extra stability might be attributed to increased core packing efficiency.

How Does Good Packing Arise?

From an unfolded conformation, proteins must somehow establish their high degree of side-chain packing. Two descriptive models of protein folding, initially proposed in the early 1970s, provide insight into this process. The nucleation model (91, 92) argued that protein folding begins with a kernel of residues making specific native-like contacts. Once the protein forms this rate-limiting configuration, the remaining structure quickly folds into place. Alternatively, in the hydrophobic collapse model (93), the protein first aggregates its nonpolar groups to form a structure with a loose hydrophobic core. Then secondary structural elements develop around this core, hypothesized to be similar to molten globule, which finally folds in a slow step to form the tightly packed native structure. In the framework model (94), a slightly different formulation of the hydrophobic collapse model, the secondary structure forms first, and then the hydrophobic groups aggregate. Therefore, in the nucleation model, the tight packing forms rapidly with no intermediates, whereas for both collapse models, the tight packing occurs only after the formation of a molten globule–like state.

Folding Pathways

A current topic of debate is whether the molten globule is an intermediate on or off the folding pathway (for a review see 95). Studying the kinetics of intermediate formation can distinguish between these possibilities. Put simply, if the molten globule is part of the folding pathway, its accumulation speeds up the formation of the native conformation (the folding rate is proportional to the fractional concentration of the intermediate). For off-pathway molten globules, formation of these structures inhibits the formation of the native conformation because the protein must fold back through the unfolded state to reach the native one (or the folding rate is proportional to 1 minus the intermediate's fractional concentration). Alternative or parallel pathways (96) show a certain fraction of the unfolded species fold quickly into the native state, while the remaining molecules follow a slower on-pathway model. The same researchers have shown that the molecules on the slower pathway form an intermediate with helical secondary structure that is just slightly more energetically stable than the unfolded state, and this minor increase in stability retards the folding reaction (G Wildegger & T Kiefhaber, submitted).

Furthermore, in almost all the equilibrium and kinetic studies, the authors assumed a sequential pathway for protein folding. This view assumes that folding proceeds similarly to a chemical reaction (98, 99). The intermediates along

this path help guide the protein to its native state (100). More recent theoretical developments suggest that folding follows an energy landscape (for a review see 101, 102). In this model, the intermediates arise because of kinetic traps where the protein is actually slightly misfolded. To continue, the protein needs to unfold only somewhat. The model is able to explain the behavior of small, fast-folding proteins (usw. <80 residues), which fold on the order of milliseconds instead of the usual seconds and without distinguishable intermediates (103–110). Because these proteins are too small to form stable intermediates, they avoid the kinetic traps and therefore fold directly to the native state. Another way to make sense of the rapid folding of small proteins is that the combinatorial search for correct side-chain packing in a small protein is much simpler and faster than in a large one. Baldwin (101) notes that this model could be thought of as an extension of the jigsaw puzzle folding model (82a). Here, the initial starting state is not fixed, and energetics coupled to a certain amount of randomness determine the folding pathway.

Equilibrium Experiments—The Molten-Globule State

The molten globule (112) has yielded a great deal of experimental information regarding the structure of intermediates during protein folding. This conformational state, an equilibrium folding intermediate induced under mild denaturing conditions, consists of the following characteristics: (*a*) It is less compact than the native state. (*b*) It is more compact than the unfolded state. (*c*) It contains extensive secondary structure. (*d*) It has loose tertiary contacts without tight side-chain packing. Recently, increasing evidence supports the idea that the molten globule may possess defined tertiary contacts (for a review see 113). It has been argued that the molten globule state contains water molecules or is "wet" (114), but an experiment by Kiefhaber et al (115) found that an unfolding intermediate with molten globule attributes is dry. Strong support for either case has yet to be found. Beyond these similarities, the molten globule conformations are very diverse among proteins and even among different molten globules induced from the same protein (116, 117). For this reason, we discuss each molten globule system individually.

CARBONIC ANHYDRASE The low pH form of carbonic anhydrase shows characteristics of a molten globule (118). Like others, this molten globule resembles a kinetic folding intermediate (119). Besides the molten globule, carbonic anhydrase provides evidence for an interesting second equilibrium intermediate (120). Because this state occurs at higher concentrations of denaturant and is less compact than the molten globule, the authors believe that it represents a premolten globule. They also show that this intermediate still contains considerable secondary structure and liken it to the burst intermediate seen in kinetic studies (121, 122).

α-LACTALBUMIN The protein α-lactalbumin can produce two forms of the molten globule under different conditions, both of which have been well characterized (123, 124); the acid form is produced at low pH and the apo form at neutral pH in the absence of calcium. Dissecting the protein to study only the alpha helical domain, Peng & Kim (125) showed that at low pH this domain contains enough of a tertiary fold that native disulfides could be found when they oxidized a reduced species in the molten globule state. On the basis of these results, along with CD and NMR data, the authors believe that the molten globule is an expanded native state with no specific side-chain interactions. Further investigation by the same group showed that the beta sheet domain is largely unstructured in the low pH molten globule (126). Such a bipartite structure is interesting because small-angle solution X-ray scattering showed a unimodal distribution, which implies that the molten globule is roughly spherical in solution (127). Using Raman optical activity measurements and studying both α-lactalbumin molten globules, Wilson et al (117) also found that both molten globules are native-like but that the apo form is less sensitive to temperature denaturation since it is more ordered.

CYTOCHROME *C* Cytochrome *c* requires low pH and addition of salt to form a molten globule (128). The salt screens repulsive electrostatic interactions caused by the acidic conditions and allows the protein to collapse. This state has been characterized as possessing an increased volume (129) and increased compressibility (130). Jeng et al (131) have shown that the N- and C-terminal helices are responsible for most of the molten globule's secondary structure. These two helices form during the early stages of folding (132) and contact each other in the native structure (133). Two groups (134, 135) have shown that packing interactions between these terminal helices are just as important to the stability of the molten globule as they are to the native state. They mutated residues important to the interaction of the N- and C-terminal helices and found destabilization of both the native and molten globule states. This result implies that the molten globule of cytochrome *c* uses some native packing contacts for stability. As an overall picture, results from small-angle X-ray scattering (136) suggest that the cytochrome *c*'s molten globule best fits a structure containing a compact core with random coils extending from it.

MYOGLOBIN Depending on its environment, myoglobin in its apo form can fold into a number of molten globular states. Like cytochrome *c*, apomyoglobin collapses from a largely unfolded conformation at pH 2 into a molten globular form upon addition of salt (137). This form of the molten globule is assumed to be similar to the one at pH 4.2 in the absence of salt (138) and has been characterized by Hughson et al (139). Their NMR analysis showed that the A, G, and H helices arrange themselves in a native-like conformation. These helices also

form during the initial stages of apomyoglobin refolding (140). In the folded state, these three helices pack against each other with large hydrophobic contact areas (141, 142), while independently they have very little helical content (143–145). At pH 2 with sodium trichloroacetate, apomyoglobin forms another molten globule state with more helical structure (146). This form is considered to be further along in the folding pathway (140). Studying both molten globular forms, Nishii et al (138) found cold and heat denaturation of the two forms, indicating that hydrophobicity contributes to the molten globules' stability. Using small-angle X-ray scattering to measure radius of gyration, they also showed that the molten globules were less compact. Hughson et al (143) mutated residues important to the packing between the A, G, and H helices of the pH 4.2 molten globule and found no perturbation of stability from acid denaturation. In fact, overpacking the interface caused an increase in stability. Approaching the problem from a different angle, Kiefhaber & Baldwin (147) created mutations that increased the helical structure of the pH 4.2 molten globule. This mutant required higher concentrations of urea to become denatured from a molten globule state, showing that increasing the secondary structure stabilizes the molten globule.

So far, these studies suggest that myoglobin folds according to the hydrophobic collapse model, but work published this past year supports an alternate view. The same lab that performed mutational studies on the pH 4.2 molten globule repeated these experiments (148) using urea, instead of acid, to denature the protein. They found that the mutations at the A, G, and H helical interfaces destabilized the molten globule as well as the native conformation. From their measurements the investigators computed that packing interactions in the molten globule are about half as strong as in the native state. Kataoka and coworkers (149) presented solution X-ray data that suggest the pH 2 trichloroacetate-stabilized molten globule consists of a single hydrophobic core surrounded by a disordered polypeptide chain. The evidence comes from the calculation of a distance distribution function. The trichloroacetate-stabilized molten globule at pH 2 showed a bimodal distribution, which is indicative of two different domains in this molten globule. Since this apomyoglobin contains only a single folding center, the authors attributed the second mode in the distribution to the unfolded portions of the chain. Native holomyoglobin and apomyoglobin, as well as other molten globules [cytochrome *c* (149) and α-lactalbumin (127)], possess unimodal distance distribution functions characteristic of a globular protein with a generally spherical shape in solution. Altogether, these experimental results lend support to the nucleation model.

Kinetic Experiments

While the previous studies looked at stable, equilibrium intermediates, the following experiments analyzed transient, kinetic intermediates found during

refolding or unfolding of the protein. Using methods such as CD or NMR coupled to stop-flow techniques to monitor the folded state of the protein, these experiments usually find a quick burst phase of folding during which intermediates cannot be detected (121, 122). After this initial burst, there is a slow phase while the molecule searches for its native state.

As discussed above, an early kinetic intermediate of both cytochrome *c* and apomyoglobin has been found that contains characteristics similar to its related molten globule (131, 140). Investigators have found the same in other systems. For ribonuclease A, Yamaguchi et al (150) found a negative change in volume as the protein went from a folded to an unfolded state by measuring the Gibbs free energy difference during pressure denaturation. Refolding of the solvent-denatured protein produces two identifiable intermediates: The near-native intermediate requires a conformational change due to a proline isomerization to reach a completely folded conformation (151–153). The other intermediate occurs early in refolding and resembles a molten globule state (154). Studies of the volume change upon refolding (155) and unfolding (156) of ribonuclease A indicate that an intermediate possesses an increased volume akin to a molten globule, while NMR analysis provides evidence that an intermediate has features of a dry molten globule (96). Further investigation of the early intermediate (157) corroborates results from equilibrium folding studies. Because the authors discovered that the early intermediate is able to bind inhibitor, possesses hydrogen protection factors similar to the near native intermediate, and has a developed β-sheet, they believe that this intermediate also contains significant tertiary structure.

Using staphyloccocal nuclease, Vidugiris et al (158) found that pressure denaturation formed a transition state with a positive activation volume (basically an increase in volume of the protein/water system). The authors liken this swollen intermediate to a molten globule state. In another study looking at apomyoglobin unfolding, Barrick & Baldwin (159) describe an intermediate state with developed helices, no strong tertiary structure, and a Gibbs free energy closer to the unfolded state than the native. From these results, they conclude that side-chain packing is responsible for most of the stability of the native state. This apomyoglobin intermediate can be thought of as the initial burst state, seen in much of the kinetic work (121, 122), in which the protein is compact and yet contains secondary structure. As discussed above, Uversky & Ptitsyn liken the burst intermediate to a premolten globule state (120). Eliezer et al (160) provided a more general view of the solution structure of apomyoglobin's folding intermediate. Their small-angle X-ray scattering showed that the initial folding intermediates at 20 and 100 ms are as compact as the molten globule and almost as compact as the refolded native state. In a quite recent analysis of dihydrofolate reductase refolding, Hoeltzli & Frieden (manuscript submitted) monitored the resolved resonances of 6-19F-tryptophan and found strong

evidence that the search for the correct residue packing causes the slow rate-limiting step of refolding. In contrast, new techniques able to look at the formation of the burst phase intermediate suggest that it contains secondary structure and residues with native tertiary contacts (for a review see 161). Although these results are still preliminary, they provide support for nucleation events in folding.

Conclusions from Experiment

Analyses of protein crystal structures (6, 20, 25), as well as solution measurements (82), show that proteins in their native conformations possess tightly packed cores. Experimental results are not as clear as to when or how this well-packed core arises. It is clear that proteins follow more than one folding pathway; however, for all pathways, collapse occurs early. With the caveat that the data come from a limited set of proteins and experiments, we can construct the following general folding progression. From a denatured state, a protein collapses into an initial burst phase intermediate (or for the small, fast-folding proteins, folds directly to the native state). This proposed premolten globular state contains a certain amount of secondary structure and tertiary contacts, but the protein's overall topology is incomplete. Next, development of the general chain topology occurs. As yet, not all the side-chains have packed well. In the end, the protein attains its native conformation with a tightly packed core. Simulations of folding (for a review see 162) as well as examination of hinge motions (72) and mutational studies (86–89) support the idea that packing can drive the last step in folding. Kinetically trapped intermediates could occur at any point along this pathway. Although still speculative, this picture does point out that both nucleation and hydrophobic collapse play important roles in protein folding. It is uncertain exactly where and to what extent either affects each stage of the folding process. In any case, experiments show that a well-packed core is essential to achieving the native state during protein folding.

MODELING THE PACKING OF SIDE-CHAINS

Early Work—Defining the Problem

The difficulty of ab initio protein structure prediction originates from the enormous number of three-dimensional conformations that a chain of amino acids can adopt. A 100-residue protein has approximately 400 degrees of freedom: Each residue has two main-chain single-bond torsion angles, ψ and ϕ, and on average two side-chain single-bond torsion angles, $\chi 1$ and $\chi 2$ (small side-chains have one χ angle; large ones have four). Crudely assuming that a torsion angle accuracy of 10° is sufficient, each residue has $36 \times 36 = 1296$ independent (ϕ, ψ) main-chain conformations, giving a main-chain combinatorial complexity

of $1296^{100} = 10^{311}$. Making the same assumption for the two side-chain torsion angles also gives a complexity of 10^{311}. The two conformational spaces are the same size. However, the main-chain torsion angle errors propagate throughout the protein and are sequentially amplified. Side-chain angle errors only affect the local conformation and propagate less directly.

In 1987, Ponder & Richards (163) pointed out that using the criterion of "good packing" against the rigidly fixed native main-chain rules out the majority of side-chain rotamer conformations for residues in core regions. Side-chain rotamers, which are a tabulation of frequently observed conformations, have been proposed for many years (164), but Ponder & Richards (163) reduced these to a set of 67 different conformations that could account for most side-chains observed in real proteins (assuming an angle tolerance of $\pm 20^\circ$). While enumeration of these conformations is computationally feasible over a few neighboring residues, the task of enumerating all possibilities for each residue in a 200-residue protein is computationally intractable. (Specifically, there are on average 3.35 rotamers per amino acid (67/20), and this gives $3.35^{100} \cong 10^{53}$ combinations.)

One of the first attempts to actually predict the side-chain conformation given the correct conformation for the main-chain involved manual modeling (165). Working with the known X-ray conformation of the main-chain of flavodoxin, this test study yielded a final side-chain prediction error of 2.41 Å RMS (root mean square). Nevertheless, many large aromatic side-chains deep within the core of the protein were very badly predicted. This in turn led to an error propagation cascade throughout, causing satisfactory prediction for only 30–40% of the side-chain conformations.

Several investigators have performed local energy minimization of a very few residues in the field of otherwise fixed protein atoms (166–168). By restricting interest to situations where only a limited number of side-chains were replaced (e.g. by assuming that conserved residues remain in similar conformations when two sequences have very high sequence similarity), these methods effectively focused their efforts on neighboring residues. Their success suggested that, if the problem could be separated into small sets of residues that interact little with each other, the daunting combinatorics of the side-chain packing problem could be surmounted.

A Possible Solution?

In 1991, four groups working independently each discovered a method that naturally broke the combinatorial problem into manageable pieces (169–172). When a protein is stripped of all its side-chains, and the native main-chain is used as a rigid constraint to repack all the side-chain atoms, these varied methods could achieve an accuracy of 1.8 Å RMS error over all side-chain atoms.

These four methods all rely on the van der Waals energy to eliminate bad side-chain arrangements. They differ very much in how they generate possible side-chain conformations and how they choose between them. The method of Lee & Subbiah (169) utilizes no database information, making it the most physically based method of the four. Side-chains are allowed to explore torsion angles in 10° intervals, and simulated annealing is used to optimize the arrangement of neighboring side-chains by minimizing the van der Waals energy. Two of the methods use a set of rotamers taken from known protein conformations and optimize an energy function [which can include hydrogen-bonding and electrostatics (171)] using Monte Carlo (MC) minimization (170) or a genetic algorithm (171). The fourth method (172) relies more heavily on known protein structures and the surprising finding of Jones & Thirup (173) that almost all segments of main-chain conformation recur in proteins. In this method, van der Waals packing energy is used to select plausible segments of known protein structure, borrowing the side-chain conformation. Rather than optimizing the side-chain conformations, it introduces some variability in selecting chain segments, averages atomic coordinates to enhance the signal from the common conformations, and then regularizes the stereochemistry with energy refinement.

Since all these methods primarily rely on only extremely simple van der Waals packing in their energy functions, a better assay of accuracy is the predicted error in the well-buried side-chains. Considering only the half of all residues that are less solvent-exposed [<30% surface area accessible to the solvent (174)] significantly improves the prediction accuracy. The only ab initio method, using simulated annealing to minimize the van der Waals energies in a finely discretized torsional space (10° for the χ angles), was accurate to 1.25 Å RMS (169). The genetic algorithm approach (175, 176) that combinatorially mates rotamers selected from a 109-member rotamer database (171) was accurate to 1.54 Å RMS. The MC energy minimization over a similar rotamer database was accurate to 1.6 Å RMS (170). The segment-matching method was accurate to 1.37 Å RMS, in spite of its use of only the native Cα positions rather than the entire main-chain. This success by four different methods that all rely on packing to eliminate bad choices proved the foresight of Ponder & Richards (163) was indeed correct.

Recent Refinements—A Classification

Over the past four years, a flood of new methods, as well as improved versions of the early ones, have been reported. The best of these, like those of Lee (177) and Vásquez (178), consistently break the 1 Å RMS barrier over a large set of proteins, while a few others (179–181) hover between 1 and 1.1 Å RMS error. Of the remaining recent methods, all report average errors of less than 1.45 Å RMS over a test set of 10–60 proteins (182–186).

The four methods discovered independently between 1991 and 1992 employ surprisingly different approaches. Classifying these and the newer methods helps highlight what is necessary for successful prediction. Methods that predict side-chain conformation from a known backbone conformation involve two steps: (*a*) choosing a set of possible conformations for each side-chain, and (*b*) choosing the conformations of each side-chain to optimize packing for a given fixed main-chain.

POSSIBLE CONFORMATIONS The set of possible conformations is either knowledge based (taken from known three-dimensional structures of proteins) or defined by simple geometrical considerations. Most methods are knowledge based (170, 171, 178–180, 183, 187), following the use of rotamer libraries by Ponder & Richards (163). Variants in both the size and content of these libraries have been attempted (180, 187–189). The latter include some studies that use a rotamer set customized to match the local main-chain of the particular side-chain. (180, 189). Others (172, 181, 190) take one or a small number of fragments from known protein structures using a local fit to the main-chain to choose fragments. A few investigators (169, 177) disregard these database approaches and instead vary the side-chain single bond torsion angles in 10° increments.

OPTIMIZING PACKING Most methods use some type of search strategy to find the combination of side-chain conformations that optimizes packing. Good packing is generally assumed to correspond to a favorable value of the van der Waals energy, with its strong steric repulsion and weak long-range attraction, but more complicated energy terms are sometimes included (171). Of greater importance than the energy is the search method used to find the best combination of side-chain conformations. Simulated annealing is surprisingly effective at finding the optimal packing corresponding to side-chain arrangements found in native proteins (169, 186, 191), as is the related MC minimization method (170). Genetic algorithms have also been used (171, 192). More elaborate search methods have also been used, such as "dead end elimination" (184, 193) and the A* algorithm (194), and these have been combined with other heuristics (187, 193). More physically based methods search with MD simulations (179, 180, 189, 192, 195), self consistent mean-fields (177, 183), and Gibbs sampling utilizing heat baths (178). One method (172) simply pastes together segments found in known proteins, subject to their packing well into the growing structure.

AB INITIO METHODS Only a handful of methods that do not rely on protein-derived knowledge have worked well. One that relies on MD "annealing" of successively added atoms beyond the Cβ atoms enjoyed some success (196)

but has since been reported to be inferior to rotamer-based methods (192). A related method of annealing "sprouted" side-chain atoms, again using MD, has only been reported to work on small peptides (197). The most successful ab initio methods (169, 177), mentioned above, rely on simple van der Waals energy in conjunction with complete sampling of torsion angle space.

Assessing the Accuracy

RANDOM OR WORST RMS The success of these methods must be put into context by considering the RMS expected if all side-chain conformations were (*a*) randomly predicted or (*b*) predicted as badly as possible. The random RMS was estimated to be 3.1 Å and the worst RMS to be 4 Å for a 100-member rotamer library (169, 170). Later work gave similar random RMS between 3.3 Å and 3.5 Å, depending on the size of the rotamer library (187). Many studies have answered the opposite question of how well the best rotamer-based prediction can represent the native structure: RMS values range from about 0.5 Å for the large 624-member rotamer libraries (170, 178, 179, 187) to 1.0 Å for the original 67-member rotamer library (163).

EXPERIMENTAL ACCURACY The answer to the question "What error value corresponds to an excellent prediction?" can be found in a rotamer-independent manner. It has long been known that when X-ray structures of the same protein are determined by two different laboratories or in two different crystal forms, the main-chain atoms differ by about 0.5 Å RMS (198, 199). The side-chains can differ by as much as 1.5 Å RMS (199), but for the more buried side-chains not involved in crystal contacts, the difference can be up to 1 Å (198). Judged against a side-chain RMS of 3.1 Å being random and a RMS of 1 Å being the best possible, the fact that automatic methods routinely achieve values as low as 1.25 Å RMS suggests that the side-chain packing problem may be solved.

TORSION ANGLES Another measure of fit is the percentage of side-chains for which the torsion angles are correctly predicted. For the buried residues, the better side-chain packing algorithms usually predict correctly (within 40°) 90% of the $\chi 1$ angles and 80% of ($\chi 1$, $\chi 2$) angle pairs (169, 177, 178, 181, 187). When all residues are considered, these figures drop to 80% and 70%, respectively. The percentage correct obviously depends on the match criteria: With stricter criteria (within 20° or 30°), these values are reduced by about 10% (170, 172, 181, 200). These predicted values must be compared with the best that can be achieved by rotamer libraries. Allowing a deviation of less than 40° from the angle derived from X-ray information, even the smaller rotamer libraries can often correctly capture the native side-chain conformations for some 95% of the $\chi 1$ angles and 90% of the ($\chi 1$, $\chi 2$) pairs (178). With the stricter criterion of being within 20° of the angle from X-ray structures, these

values drop to 85% and 75%, respectively (200). It is encouraging that for the buried side-chains, the success rate of prediction is only 10% less than the best possible with rotamer libraries.

PREDICTIVE SUCCESS In terms of claimed accuracy, the ab initio method of Lee (177) and the rotamer-based method of Vásquez (178) are marginally superior to all others. Lee has published predictions prior to experimental X-ray determination that have proved to be accurate. He has reported RMS errors of 0.68–0.89 Å in side-chain prediction for T4 lysozyme mutants (201), 1.11 Å on λ-repressor mutants (202), and 0.97 Å RMS on polymeric HLA alleles (203). While some caution should be expressed since these predictions are only for a few buried residues, the results do suggest that the best side-chain packing methods can be useful.

Why Is This An Easy Problem?

Since it appears that the packing of side-chains can be well predicted, some investigators have suggested the problem is not really combinatorial in that the allowed side-chain conformations depend on the local main-chain environment (180, 187, 204). Methods that choose the side-chain conformation based only on the local main-chain are about 20% less accurate than methods that allow full combinatorial packing (178, 183). This remaining 20% in accuracy can only be obtained by considering combinatorial packing (178, 180, 183, 187).

Main-Chain Movement

It is becoming increasingly clear that the assumption of a fixed main-chain during combinatorial repacking is not generally valid. Attempts have been made very recently to relax this assumption. A clever method, which allows main-chain and side-chain flexibility, has been applied to the special case of repeating coiled-coil structures; it is able to predict the buried side-chains almost as accurately as when the perfect main-chain is available (195). Koehl & Delarue (205) have applied the mean-field approach, so successful for side-chains (177, 183), to the main-chain with promising results. Wilson et al (179) have proposed the use of rounds of alternating side-chain packing onto a fixed main-chain and full MD minimization. The multiple copy and mean-field approaches also appear to be particularly well suited to allowing main-chain shifts (183, 206, 207).

GETTING TO THE ENDGAME

How Close Is Close Enough?

In order to get to the endgame, one needs a backbone that is very close to native. How close is close enough? The question of whether packing optimization schemes can model side-chains accurately upon fixed imperfect backbones has

been under intense scrutiny. Many studies have considered the repacking of correctly aligned target sequences onto fixed homologous template structures, employing the same algorithms used when the ideal backbone was provided (168, 179, 180, 183, 188). Recently, in a more systematic study that spanned the full range of possible sequence identities within certain protein families, Chung & Subbiah (208, 209) observed a monotonic decrease in buried side-chain prediction accuracy as the sequence identity diverged and backbone deviation increased. They estimate that when the template is more than 2 Å RMS error from the native backbone (corresponding to ~25% sequence identity), the side-chain prediction accuracy approaches the random expectation of 3.1–3.3 Å RMS (169, 187, 209).

In the absence of general methods that accommodate movable backbones in side-chain prediction, it appears that backbones within 2 Å RMS of the native structure are required for accurate modeling. Backbones as accurate as these are sometimes available if the structure of a close sequence homologue is known and the two sequences are correctly aligned. In the general case, how is it possible to obtain folded backbones that are sufficiently accurate?

Threading Methods

In one approach, known as threading or fold recognition, a new sequence is aligned upon a known three-dimensional structure, and each sequence-structure alignment is scored via an energy function. Threading has identified compatible folds that are undetectable by conventional sequence alignment methods (210, 211). However, success in recognizing a related fold does not imply success in building an accurate model using the related fold as a template. The alignment of the new sequence on the known backbone has to be almost perfectly correct to get the required 2-Å accuracy (adjacent residues are about 4 Å apart). Results from the threading predictions illustrated the various shortcomings of available alignment and/or scoring methods (212). Moreover, even given perfect alignments, backbones generated by threading methods may not be useful if the aligned sequences show less than 30% identity (208). Threading specializes in finding such folds, so it is unlikely to provide acceptable backbones for standard side-chain prediction methods, even if the alignment were optimal. In any case, at the present time many proteins of interest are new folds for which there is no threading target. Hence, we do not regard threading in its current form as a viable pathway to the endgame of folding.

Ab Initio Folding

In ab initio methods, a fold for the new sequence is generated without directly using the known fold of any other protein. This is accomplished either by (*a*) a broad and even sampling of conformational space by an energy-independent

method, followed by screening of the resulting candidate folds by an energy function or (*b*) minimizing the conformational energy of a polypeptide as it folds through an approximately continuous conformational space. In either case, the level of detail included in the structural representation must balance the computational tractability and geometric accuracy of the model. Lattice models can be computationally feasible, even to the point of enumerating folds exhaustively (213), but such folds sacrifice secondary structure features and are generally less accurate than 5 Å RMS. Off-lattice discrete models, such as those possessing six states per residue, can reproduce the native backbone to 2 Å RMS (214), but generating such folds exhaustively is beyond the power of today's computers (a chain of length 100 has $6^{100} = 10^{70}$ folds). For minimization methods, complex lattice and discrete representations hinder the search for the energy minimum because they make the energy landscape more rugged and increase the number of moves necessary to traverse the conformational space. In spite of these limitations, ab initio folding approaches have made progress, routinely achieving structures with accuracy up to ∼4 Å RMS error (215–227). Other methods, especially those that use experimental secondary structure constraints, fare even better. In the next section, we review approaches to ab initio folding that have produced folds within 2–4 Å RMS of the native structure.

Discrete-State Models and Energy Functions

A simple discrete-state model has been described by Park & Levitt (228). Their optimized four-state off-lattice representation is able to build backbones within 2 Å RMS from the native backbone. Even with this model, exhaustive enumeration is impossible, as $4^{100} = 10^{60}$ is intractable. If one enforces the native secondary structure as an external constraint, this model has no more than about 200,000 folds for each protein. This is a manageable number of folds that evenly and broadly sample phase space while providing candidates that take folding into the end-game.

Providing such candidates in itself is not useful unless an energy function can successfully distinguish the native-like folds from the entire set of folds generated. This issue presents two questions: (*a*) Can energy functions distinguish between the near-native folds and those that are grossly misfolded? (*b*) Can the energy functions distinguish between the native structure and the near-native folds? The first question depends as much on the quality of the representation as on the effectiveness of the energy function; i.e. energy functions are useful only in the context of a representation capable of generating suitably near-native structures. The second question asks whether or not the energy function can tell a true native fold from the best near-native decoys; it thus assesses the resolution of the function and indicates whether further minimization of the function can in principle drive the conformation towards a more native-like state.

Park & Levitt (229), using a basis set of six energy functions, report that the native fold can be recognized very effectively if one combines energy functions that stress complementary factors, such as nonspecific hydrophobicity (a general compacting force) and residue-specific pairings. Furthermore, the best of the native-like folds usually rank very high in the energy-sorted list. On average, the best combinations of energy functions place the native-like folds in the top 1% of the score-sorted list (229), although there are always many grossly misfolded decoys with energies more favorable than some of the near-native folds. Therefore, if one were to apply an effective energy function as a screen of the entire decoy set (for example, by taking the top half of the energy-sorted list), the concentration of the near-native folds in the high-scoring subset would increase, but the highest-scoring folds in the subset would not all be near-native. In other words, RMS deviation and energies are not highly correlated in the RMS range explored in this study (229). More encouraging is that the best energy functions typically score the native fold more favorably than all the decoys, including those within 2 Å RMS.

Energy Minimization and Search Strategies

Methods that use energy minimization to move through phase space have shown promise in folding to near-native conformations. Recent work by Mumenthaler & Braun (230) describes a self-correcting distance geometry method for predicting the tertiary arrangement of small globular helical proteins. This method, like the one by Park & Levitt (228, 229), assumes that the helical segments are known in advance; only the (ϕ, ψ) dihedral angles of loop residues are adjustable (though constrained to combinations that are commonly observed in the database for each residue type). First, the method predicts whether each residue is solvent-exposed ("outside") or buried ("inside"), using an algorithm that exploits multiple sequence alignment information (231). Upper limits for the distances between the three types of residue pairings (inside-inside, outside-outside, and inside-outside) are calculated as a function of the size of the protein. The minimization engine then applies these distance constraints in a clever algorithm that dynamically adjusts constraints over each iteration of the structure generation cycle. Thus, rather than having an energy function per se, the method relies on a "target function" that depends on the predicted constraints. The structures with the fewest constraint violations tend to cluster within 3 Å RMS of the experimentally determined structure, although only the helical residues were included in the RMS calculation. The final predicted structure, taken as the average structure in the low-violations cluster, can be accurate to 2.3 Å RMS of the native structure. Because the constraints are adjusted to the structures during the procedure, there is no path-independent energy function available for further minimization. Overall, six out of eight test proteins converged to

near-native predictions ($\leq$3 Å RMS error), but none were within 2 Å. Nevertheless, this method can be a useful tool for taking folding into the end-game, assuming that secondary structure prediction methods continue to improve.

A similar minimization approach was developed by Sun et al (232). Like the two procedures discussed above, this method also begins with the known secondary structure elements in order to reduce the conformational space to be searched. Their conformational search engine is two tiered and is powered by a genetic algorithm that operates on a string of paired (ϕ, ψ) dihedral angles describing the conformation of the protein. First, mutation and crossover operations are performed at randomly chosen rotatable residues (i.e. those not in secondary structure). Mutations are random selections from a set of dihedral angle pairs derived from the structure database. The second step refines the search by perturbing randomly chosen unconstrained torsion angles slightly in order to probe the local energy landscape for minima. The selection method is an energy function that models the hydrophobic interaction (1) and is an extension of the simple hydrophobic-polar models of Dill and coworkers (233). The results were encouraging. Out of ten test cases, four of the lowest-energy models were within 4 Å RMS error, but none of the minimized structures achieved 2-Å accuracy. Moreover, many of the native structures had energies much worse than the minimized structures, thus limiting the utility of their highly simple energy function in the endgame.

Let us summarize the strengths and shortcomings of the ab initio methods discussed above. The results of Park & Levitt (228, 229) suggest that an effective energy function (of which there are several) yoked with the proper search strategy can drive near-native folds towards the native fold. However, the same function cannot reliably recognize near-native folds, even the best ones, from the entire set of decoys. For near-native structure generation, the minimization method of either Mumenthaler & Braun (230) or Sun et al (232) might be a better alternative. However, these methods are not fail-safe, for they do not always converge near the native structure.

In the ab initio methods discussed above, folds were generated either exhaustively (229) or from random tertiary arrangements (230, 232). As close as these methods can get to the native fold, their accuracy is hampered by the reduced complexity of the model, the energy functions that drive the folding of the chain, or both. In the next section, we address these concerns. Energy functions are challenged to recognize native folds from all-atom representations very close in conformation to the native fold.

Discriminating Native from Near-Native Conformations

A key requirement of an energy function able to drive the search towards the endgame is that the native conformation have a lower energy than the near-native

conformations. Such sets of near-native conformations can be generated by deforming the experimentally determined structures using methods such as MC and MD simulations. Energy functions are then applied to these test sets in order to assay their discrimination power.

The method developed by Wang et al (234, 235) was the first attempt at recognizing the native fold from large decoy sets of near-native and compact structures. This method is based on the atomic solvation potential of Eisenberg & McLachlan (236), grouping atoms into 17 chemically related "molecular fragment types," each with its associated solvation parameter. These parameters were obtained by a training algorithm that maximizes the solvation energy difference between the native and a large set of compact nonnative structures generated by MC and MD simulations (235 and references therein). The solvation parameters were then used to evaluate native structures of a separate test set of decoy structures generated by MC and MD. The MC-generated structures were selected to be compact (the radius of gyration did not exceed that of the native structure plus 5%) and within predetermined RMS deviation from the native structure (up to 5 Å maximum). The MD simulations were carried out at room temperature (300 K) and high temperature (500 K); the average RMS errors for the 300-K and 500-K simulations were 4.1 Å RMS and 8.0 Å RMS, respectively. More than 8200 nonnative MC and MD decoys were furnished for each of 11 test proteins, of which only 7 on average were misrecognized as native (having a more favorable energy score than the experimentally determined structure). The solvation energy roughly correlated with the RMS deviation between the native and decoy structures. All of the misrecognized decoys, or false positives, were structures very close to the native ($<$1 Å RMS). Wang et al (234) also demonstrated that their method compared favorably against a battery of standard energy functions: MD force fields, statistically derived contact potentials, three-dimensional profile methods, knowledge-based potentials of mean force, and others (188, 210, 237–242).

In a related study, Huang et al (243) explored the ability of a very simple hydrophobic contact function (244) to recognize near-native decoys generated by MD simulation in solution at room (298 K) and high (498 K) temperatures. Five small proteins formed the test set. Overall, the average RMS deviations from the native structure were 1.5 Å (at 298 K) and 4.1 Å (at 498 K). As in the earlier studies (234, 235, 243), native structures were readily identified from the sets of decoy structures: There were only 330 false positives out of 10,000 (combined room and high temperature runs for the five proteins). Likewise, the energy function is strongly dependent on the extent to which the structures are deformed: Only one false positive exhibited an RMS deviation more than 2 Å from the native structure (243).

What is the impact of these two studies on how the endgame is played? Both appear to be successful at identifying native folds from compact, near-native

folds, a quality that other functions apparently lack (234). Huang et al (243) note that the decoy set used in their study is perhaps a more rigorous test, given the lower RMS deviations produced from MD simulations. Indeed, demonstrating that simple energy functions can discriminate native from near-native structures in this RMS range (0–2 Å) is important. Given that ab initio methods can provide folds that are quite close to the native (around 2 Å), it is important to use methods such as MC and MD simulations to probe the relationship between energy and molecular conformation within 2 Å RMS from the native. However, even more challenging near-native test sets are needed to assess the true discrimination power of existing potentials. High-temperature MD simulations (234, 235, 243) and the MC simulations of Wang et al (235) compromise the integrity of the secondary structure and loosen the packing of the tertiary structure. Even the 298-K MD simulations in solvent by Huang et al (243), which depart from the native by an average of only 1.5 Å RMS, undergo a 2–3% increase in the radius of gyration. A function that stresses hydrophobicity (i.e. nonspecific compacting force), such as the one by Huang et al (243), is sensitive to minute changes of this type. Corroborating evidence is seen in recent work by Levitt and coworkers, who have tested the performance of 18 energy functions on this set of MD structures (245). This study indicated that other energy functions emphasizing hydrophobicity also excelled at native fold discrimination.

Although RMS deviation imperfectly serves as a coordinate along the folding trajectory, it is encouraging nonetheless to confirm its strong correlation with energy functions (243). Although neither study attempted to minimize their respective energy functions using near-native structures as starting points, we challenge future studies to progress along these lines.

CONCLUSIONS

Proteins are close-packed both in the solid state and in solution. In fact, they are probably the most tightly packed form of organic matter. This close-packing is related to function in that it provides a rigid core on which to arrange catalytic side-chains in enzymes. Loose-packing is often associated with flexible hinges and conformational changes, whereas tight packing correlates with better stability. How such tight-packing arises in protein folding is still unclear, although there has been enormous progress in characterizing the packing of partially folded intermediates. However it arises, this close-packing limits the number of possible arrangements of the side-chains, which has led to methods capable of predicting side-chain packing on a known, rigid main-chain. These same methods are applicable to homology modeling provided the main-chain "borrowed" from the related structure is close enough (within 2 Å). If no homologous structure is known, other methods can sometimes generate main-chains that are almost close enough (<3.5 Å RMS). It is crucial to have

an energy function that can recognize the folds that are closer to the native structure.

Throughout the review, we have argued that packing forms a strong constraint on protein structure, severely restricting the number of possible structures. However, in the earlier stages of the folding process, particularly those relating to the formation of the overall fold, it is believed that packing is much less important. This theory has been borne out in experimental studies demonstrating how tolerant a fold is to many random mutations (89, 246–248). It has also been substantiated in theoretical studies that show how surprisingly easy it is for a protein of random sequence (a "random hetropolymer") to close-pack in an approximate sense (249–251).

A number of challenges lie ahead. Perhaps the greatest is to understand how a protein close-packs its residues during the latter stages of folding. The early stages are generally considered to be dominated by nonspecific hydrophobic interactions. Another challenge is to understand how packing affects function: If loose-packing is essential for function, it should be possible to design proteins that are too stable to function as catalysts. In the area of computer simulations, we expect progress in the consideration of main-chain flexibility, derivation of strongly discriminating energy functions, and generation of diverse sets of decoy folds. For structure prediction, the problem of packing side-chains using a near-native backbone seems almost completely solved. The challenge now is to generate main-chains sufficiently close to the native backbone to allow packing algorithms to be successful. It also seems likely that designing small helical proteins will be easiest, and their detailed structure could be predicted over the next five years!

Visit the *Annual Reviews home page* at http://www.annurev.org.

Literature Cited

1. Kauzmann W. 1959. *Adv. Prot. Chem.* 14:1–63
2. Dill KA. 1990. *Biochemistry* 29:7133–55
3. Kuwajima K. 1996. *FASEB J.* 10:102–9
4. Levitt M. 1976. *J. Mol. Biol.* 104:59–107
5. Richards FM. 1974. *J. Mol. Biol.* 82:1–14
6. Richards FM. 1977. *Annu. Rev. Biophys. Bioeng.* 6:151–76
7. Daggett V, Levitt M. 1994. *Curr. Opin. Struct. Biol.* 4:291–95
8. Fersht AR. 1995. *Curr. Opin. Struct. Biol.* 5:79–84
9. Karplus M, Sali A. 1995. *Curr. Opin. Struct. Biol.* 5:58–73
10. Koehl P, Delarue M. 1996. *Curr. Opin. Struct. Biol.* 6:222–26
11. Vásquez M. 1996. *Curr. Opin. Struct. Biol.* 6:217–21
12. Richards FM. 1985. *Methods Enzymol.* 115:440–64
13. Richards FM, Lim WA. 1994. *Q. Rev. Biophys.* 26:423–98
14. Petitjean M. 1994. *J. Comp. Chem.* 15: 507–23
15. Gerstein M, Chothia C. 1996. *Proc. Natl. Acad. Sci. USA* 93:10167–72
16. Madan B, Lee B. 1994. *Biophys. Chem.* 51:279–89

17. Gerstein M, Tsai J, Levitt M. 1995. *J. Mol. Biol.* 249:955–66
18. Voronoi GF. 1908. *J. Reine Angew. Math.* 134:198–287
19. Bernal JD, Finney JL. 1967. *Discuss. Faraday Soc.* 43:62–69
20. Chothia C. 1975. *Nature* 254:304–8
21. Finney JL. 1975. *J. Mol. Biol.* 96:721–32
22. Richards FM. 1979. *Carlsberg Res. Commun.* 44:47–63
23. Finney JL, Gellatly BJ, Golton IC, Goodfellow J. 1980. *Biophys. J.* 32(1):17–33
24. Janin J, Chothia C. 1990. *J. Biol. Chem.* 265:16027–30
25. Harpaz Y, Gerstein M, Chothia C. 1994. *Structure* 2:641–49
26. Shih JP, Sheu SY, Mou CY. 1994. *J. Chem. Phys.* 100:2202–12
27. Tsai J, Gerstein M, Levitt M. 1996. *J. Chem. Phys.* 104:9417–30
28. Finney JL. 1978. *J. Mol. Biol.* 119:415–41
29. David CW. 1988. *Biopolymers* 27:339–44
30. Edelsbrunner H, Mucke E. 1994. *ACM Trans. Graph.* 13:43–72
31. Edelsbrunner H, Facello M, Ping F, Jie L. 1995. *Proc. 28th Hawaii Int. Conf. Syst. Sci.*, pp. 256–64
32. Connolly ML. 1983. *J. Appl. Cryst.* 16:548–58
33. Connolly ML. 1983. *Science* 221:709–13
34. Connolly ML. 1986. *J. Mol. Graph.* 4:3–6
35. Gregoret LM, Cohen FE. 1990. *J. Mol. Biol.* 211(4):959–74
36. Rashin AA, Iofin M, Honig B. 1986. *Biochemistry* 25:3619–25
37. Tilton RF Jr, Singh UC, Weiner SJ, Connolly ML, Kuntz ID Jr, et al. 1986. *J. Mol. Biol.* 2(2):443-56
38. Alard P, Wodak S. 1991. *J. Comp. Chem.* 12:918–22
39. Hubbard SJ, Argos P. 1994. *Protein Sci.* 3(12):2194–206
40. Hubbard SJ, Gross KH, Argos P. 1994. *Protein Eng.* 7(5):613–26
41. Kleywegt GJ, Jones TA. 1994. *Acta Cryst. D* 50:178–85
42. Williams MA, Goodfellow JM, Thornton JM. 1994. *Protein Sci.* 3(8):1224–35
43. Hubbard SJ, Argos P. 1995. *Protein Eng.* 8(10):1011–15
44. Sreenivasan U, Axelsen PH. 1992. *Biochemistry* 31:12785–91
45. Baker EN, Hubbard RE. 1984. *Prog. Biophys. Mol. Biol.* 44:97–179
46. Matthews BW, Morton AG, Dahlquist FW. 1995. *Science* 270:1847–49
47. Chan HS, Dill KA. 1990. *Proc. Natl. Acad. Sci. USA* 87:6388–92
48. Chan HS, Dill KA. 1991. *Annu. Rev. Biophys. Biophys. Chem.* 20:447–90
49. Gregoret LM, Cohen FE. 1991. *J. Mol. Biol.* 219(1):109–22
50. Hunt NG, Gregoret LM, Cohen FE. 1994. *J. Mol. Biol.* 241:214–25
51. Franks F. 1983. *Water.* London: R. Soc. Chem.
52. Chandler D, Weeks JD, Andersen HC. 1983. *Science* 220:787–94
53. Zichi DA, Rossky PJ. 1986. *J. Chem. Phys.* 84:2814–22
54. Gerstein M, Lynden-Bell RM. 1993. *J. Phys. Chem.* 97:2991–99
55. Gerstein M, Lynden-Bell RM. 1993. *J. Mol. Biol.* 230:641–50
56. Chothia C, Finkelstein AV. 1990. *Annu. Rev. Biochem.* 59:1007–39
57. Jones S, Thornton J. 1996. *Proc. Natl. Acad. Sci. USA* 93:13–20
58. Shoichet BK, Kuntz ID. 1991. *J. Mol. Biol.* 221:327–46
59. Cherfils J, Duquerroy S, Janin J. 1991. *Proteins: Struct. Funct. Genet.* 11:271–80
60. Walls PH, Sternberg MJ. 1992. *J. Mol. Biol.* 228:277–97
61. Cherfils J, Janin J. 1993. *Curr. Opin. Struct. Biol.* 3:265–69
62. Ponder JW, Richards FM. 1987. In *Evolution of Catalytic Function*, 52:421–28. Cold Spring Harbor, NY: Cold Spring Harbor Lab. Press
63. Lesk AM, Chothia C. 1984. *J. Mol. Biol.* 174:175–91
64. Gerstein M, Lesk AM, Chothia C. 1994. *Biochemistry* 33:6739–49
65. Lawson CL, Zhang R, Schevitz RW, Otwinowski Z, Joachimiak A, Sigler PB. 1988. *Proteins* 3:18–31
66. McPhalen CA, Vincent MG, Picot D, Jansonius JN, Lesk AM, Chothia C. 1992. *J. Mol. Biol.* 227:197–13
67. Frauenfelder H, Sligar SG, Wolynes PG. 1991. *Science* 254:1598–1603
68. Hubbard SJ, Argos P. 1996. *J. Mol. Biol.* 261:289–300
69. Lesk AM, Chothia C. 1988. *Nature* 335:188–90
70. Segawa S, Richards FM. 1988. *Biopolymers* 27:23–40
71. Gerstein M, Chothia CH. 1991. *J. Mol. Biol.* 220:133–49
72. Gerstein M, Anderson BF, Norris GE, Baker EN, Lesk AM, Chothia C. 1993. *J. Mol. Biol.* 234:357–72
73. Gerstein M, Schulz G, Chothia C. 1993. *J. Mol. Biol.* 229:494–501
74. Jolicoeur C, Riedl B, Desrochers D, Lemelin LL, Zamojska R, Enea O. 1986. *J. Solut. Chem.* 15:109–28
75. Kharakoz DP. 1989. *Biophys. Chem.* 34:5634–42

76. Kunzt ID, Kauzmann W. 1974. *Adv. Protein Chem.* 28:239–45
77. Lee JC, Gekko K, Timasheff SN. 1979. *Methods Enzymol.* 61:26–49
78. Gavish B, Gratton E, Hardy CJ. 1983. *Proc. Natl. Acad. Sci. USA* 80:750–54
79. Gekko K, Hasegawa Y. 1986. *Biochemistry* 25:6563–71
80. Kharakoz DP, Mkhitaryan AG. 1986. *Mol. Biol.* 20:312–21
81. Iqball M, Verrall RE. 1987. *J. Phys. Chem.* 91:1935–41
82. Chalikian TV, Totrov M, Abagyan R, Breslauer KJ. 1996. *J. Mol. Biol.* 260: 588–603

82a. Harrison SC, Durbin R. 1985. *Proc. Natl. Acad. Sci. USA* 82:4028–30

82b. Matthews BW. 1995. *Adv. Prot. Chem.* 46:249–78

83. Baldwin EP, Matthews BW. 1994. *Curr. Opin. Biotechnol.* 5:396–402
84. Hubbard SJ, Argos P. 1995. *Curr. Opin. Biotechnol.* 6:375–81
85. Russell RJM, Taylor GL. 1995. *Curr. Opin. Biotechnol.* 6:370–74
86. Ishikawa K, Nakamura H, Morikawa K, Kanaya S. 1993. *Biochemistry* 32:6171–78
87. Anderson DE, Hurley JH, Nicholson H, Baase WA, Matthews BW. 1993. *Protein Sci.* 2:1285–90
88. Lim WA, Hodel A, Sauer RT, Richard FM. 1994. *Proc. Natl. Acad. Sci. USA* 91:421–27
89. Munson M, Balasubramanian S, Fleming KG, Nagi AD, O'Brien R, et al. 1996. *Protein Sci.* 5:1584–93
90. Ramachandran S, Udgaonkar JB. 1996. *Biochemistry* 35:8776–85
91. Tsong TY, Baldwin RL. 1972. *J. Mol. Biol.* 63:453–75
92. Wetlaufer DB. 1973. *Proc. Natl. Acad. Sci. USA* 70:697–701
93. Ptitsyn OB. 1973.*Dokl. Akad. Nauk SSSR* 210:1213–15
94. Kim PS, Baldwin RL. 1990. *Annu. Rev. Biochem.* 59:631–60
95. Baldwin RL. 1996. *Fold. Des.* 1:R1–R8
96. Kiefhaber T. 1995. *Proc. Natl. Acad. Sci. USA* 92:9029–33
97. Deleted in proof
98. Kim PS, Baldwin RL. 1982. *Annu. Rev. Biochem.* 51:459–89
99. Ptitsyn OB. 1987. *J. Protein Chem.* 6: 273–93
100. Creighton TE, Darby NJ, Kemmink J. 1996. *FASEB J.* 10:110–18
101. Baldwin RL. 1995. *J. Biomol. NMR* 5:103–9
102. Wolynes PG, Luthey-Schulten Z, Onuchic JN. 1996. *Chem. Biol.* 3:425–32
103. Jackson SE, Fersht AR. 1991. *Biochemistry* 30:10428–35
104. Khorasanizadeh S, Peters ID, Butt TR, Roder H. 1993. *Biochemistry* 32:7054–63
105. Milla ME, Sauer RT. 1994. *Biochemistry* 33:1125–33
106. Kuszewski J, Clore GM, Gronenborn AM. 1994. *Protein Sci.* 3:1945–52
107. Sosnick TR, Mayne L, Hiller R, Englander SW. 1994. *Nat. Struct. Biol.* 1:149–56
108. Kragelund BB, Robinson CV, Knudsen J, Dobson CM, Poulsen FM. 1995. *Biochemistry* 34:7217–24
109. Schindler T, Herrler M, Marahiel MA, Schmid FX. 1995. *Nat. Struct. Biol.* 2:663–73
110. Huang GS, Oas TG. 1995. *Proc. Natl. Acad. Sci. USA* 92:6878–82
111. Deleted in proof
112. Ohgushi M, Wada A. 1983. *FEBS Lett.* 164:21–24
113. Ptitsyn O. 1996. *Nat. Struct. Biol*. 3:488–90
114. Finkelstein AV, Shakhnovich EI. 1989. *Biopolymers* 28:1681–94
115. Kiefhaber T, Labhardt AM, Baldwin RL. 1995. *Nature* 375:513–15
116. Fink AL. 1995. *Annu. Rev. Biophys. Biomol. Struct.* 24:495–522
117. Wilson G, Ford SJ, Cooper A, Hecht L, Wen ZQ, Barron LD. 1995. *J. Mol. Biol.* 254:747–60
118. Dolgikh DA, Abaturov LV, Brazhnikov EV, Lebedev IO, Chirgadze IN, Ptitsyn OB. 1983. *Dokl. Akad. Nauk SSSR* 272:1481–84
119. Dolgikh DA, Kolomiets AP, Bolotina IA, Ptitsyn OB. 1984. *FEBS Lett.* 165:88–92
120. Uversky VN, Ptitsyn OB. 1996. *J. Mol. Biol.* 255:215–28
121. Gilmanshin RI, Ptitsyn OB. 1987. *FEBS Lett.* 223:327–29
122. Kuwajima K, Yamaya H, Miwa S, Sugai S. 1987. *FEBS Lett.* 227:115–18
123. Kuwajima K. 1989. *Proteins: Struct. Funct. Genet.* 6:87–103
124. Alexandrescu AT, Evans PA, Pitkeathly M, Baum J, Dobson CM. 1993. *Biochemistry* 32:1707–18
125. Peng Z-Y, Kim PS. 1994. *Biochemistry* 33:2136–41
126. Wu LC, Peng Z-Y, Kim PS. 1995. *Nat. Struct. Biol*. 2:281–86
127. Dolgikh DA, Abaturov LV, Bolotina IA, Brazhnikov EV, Bychkova VE, et al. 1985. *Eur. Biophys.* 13:109–21
128. Goto Y, Nishikiori S. 1991. *J. Mol. Biol.* 222:679–86
129. Foygel K, Spector S, Chatterjee S, Kahn PC. 1995. *Protein Sci.* 4:1426–29

130. Chalikian TV, Gindikin VS, Breslauer KJ. 1995. *J. Mol. Biol.* 250:291–306
131. Jeng MF, Englander SW, Elöve GA, Wand AJ, Roder H. 1990. *Biochemistry* 29:10433–37
132. Roder H, Elöve GA, Englander SW. 1988. *Nature* 335:700–4
133. Ochi H, Hata Y, Tanaka N, Kakudo M, Sakurai T, et al. 1983. *J. Mol. Biol.* 166:407–18
134. Marmorino JL, Pielak CJ. 1995. *Biochemistry* 34:3140–43
135. Colón W, Elöve GA, Wakem LP, Sherman F, Roder H. 1996. *Biochemistry* 35:5538–49
136. Doniach S, Bascle J, Garel T, Orland H. 1995. *J. Mol. Biol.* 254:960–67
137. Goto Y, Fink AL. 1990. *J. Mol. Biol.* 214:803–5
138. Nishii I, Kataoka M, Tokunaga F, Goto Y. 1994. *Biochemistry* 33:4903–9
139. Hughson FM, Wright PE, Baldwin RL. 1990. *Science* 249:1544–48
140. Jennings PA, Wright PE. 1993. *Science* 262:892–96
141. Richmond TJ, Richards FM. 1978. *J. Mol. Biol.* 119:537–55
142. Weaver DL. 1992. *Biopolymers* 32:477–90
143. Hughson FM, Barrick D, Baldwin RL. 1991. *Biochemistry* 30:4113–18
144. Barrick D, Baldwin RL. 1993. *Protein Sci.* 2:869–76
145. Waltho JP, Feher VA, Mertuka G, Dyson HJ, Wright PE. 1993. *Biochemistry* 32: 6337–47
146. Goto Y, Takahashi N, Fink AL. 1990. *Biochemistry* 29:3480–88
147. Kiefhaber T, Baldwin RL. 1996. *J. Mol. Biol.* 252:122–32
148. Kay MS, Baldwin RL. 1996. *Nat. Struct. Biol.* 3:439–45
149. Kataoka M, Nishii I, Fujisawa T, Ueki T, Tokunaga F, Goto Y. 1995. *J. Mol. Biol.* 249:215–28
150. Yamaguchi T, Yamada H, Akasaka K. 1995. *J. Mol. Biol.* 250:689–94
151. Cook KH, Schmid FX, Baldwin RL. 1978. *Proc. Natl. Acad. Sci. USA* 76: 6157–61
152. Schmid FX, Blaschek H. 1981. *Eur. J. Biochem.* 114:111–17
153. Schmid FX. 1983. *Biochemistry* 22: 4690–96
154. Udgaonkar JB, Baldwin RL. 1990. *Proc. Natl. Acad. Sci. USA* 87:8197–8201
155. Ybe JA, Kahn PC. 1994. *Protein Sci.* 3:638–69
156. Tamura Y, Gekko K. 1995. *Biochemistry* 34:1878–84
157. Udgaonkar JB, Baldwin RL. 1995. *Biochemistry* 34:4088–96
158. Vidugiris GJA, Markley JL, Royer CA. 1995. *Biochemistry* 34:4909–12
159. Barrick D, Baldwin RL. 1993. *Biochemistry* 32:3790–96
160. Eliezer D, Jennings PA, Wright PE, Doniach S, Hodgson KO, Tsuruta H. 1995. *Science* 270:487–88
161. Eaton WA, Thompson PA, Chen CK, Hagen SS, Hofrichter J. 1996. *Structure* 4:1133–39
162. Hinds DA, Levitt M. 1995. *Trends Biotechnol.* 13:23–27
163. Ponder JW, Richards FM. 1987. *J. Mol. Biol.* 193:775–91
164. Janin J, Wodak S, Levitt M, Maigret B. 1978. *J. Mol. Biol.* 125:357–86
165. Reid LS, Thornton JM. 1989. *Proteins: Struct. Funct. Genet.* 5:170–82
166. Snow ME, Amzel LM. 1986. *Proteins: Struct. Funct. Genet.* 1:267–79
167. Summers NL, Karplus M. 1989. *J. Mol. Biol.* 210:785–812
168. Schiffer CA, Caldwell JW, Kollman PA, Stroud RM. 1990. *Proteins* 8:30–43
169. Lee C, Subbiah S. 1991. *J. Mol. Biol.* 217:373–88
170. Holm L, Sander C. 1991. *J. Mol. Biol.* 218:183–94
171. Tufféry P, Etchebest C, Hazout S, Lavery R. 1991. *J. Biomol. Struct. Dyn.* 8:1267–89
172. Levitt M. 1992. *J. Mol. Biol.* 226:507–33
173. Jones TA, Thirup S. 1986. *EMBO J.* 5:819–22
174. Lee B, Richards FM. 1971. *J. Mol. Biol.* 55:379–400
175. Holland JH. 1975. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence.* Ann Arbor, MI: Univ. Mich. Press
176. Goldberg A. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Reading, MA: Addison-Wesley
177. Lee C. 1994. *J. Mol. Biol.* 236:918–39
178. Vásquez M. 1995. *Biopolymers* 36:53–70
179. Wilson C, Gregoret LM, Agard DA. 1993. *J. Mol. Biol.* 229:996–1006
180. Eisenmenger F, Argos P, Abagyan R. 1993. *J. Mol. Biol.* 231:849–60
181. Laughton CA. 1994. *J. Mol. Biol.* 235: 1088–97
182. Holm L, Sander C. 1992. *Proteins* 14:213–23
183. Koehl P, Delarue M. 1994. *J. Mol. Biol.* 239:249–75
184. Desmet J, Maeyer MD, Hazes B, Lasters I. 1992. *Nature* 356:539–42

185. Dunbrack RL, Karplus M. 1993. *J. Mol. Biol.* 230:543–74
186. Hwang JK, Liao WF. 1995. *Protein Eng.* 8:363–70
187. Tanimura R, Kidera A, Nakamura H. 1994. *Protein Sci.* 1994:2358–65
188. Holm L, Sander C. 1992. *J. Mol. Biol.* 225:93–105
189. Dunbrack RL, Karplus M. 1994. *Nat. Struct. Biol*. 1:334–40
190. Wendoloski JJ, Salemme FR. 1992. *J. Mol. Graph.* 10:124–26
191. Hellinga HW, Richards FM. 1994. *Proc. Natl. Acad. Sci. USA* 91:5803–7
192. Cregut D, Liautard JP, Chiche L. 1994. *Protein Eng.* 7:1333–44
193. Goldstein RF. 1994. *Biophys. J.* 66:1335–40
194. Leach AR. 1994. *J. Mol. Biol.* 235:345–56
195. Harbury PB, Tidor B, Kim PS. 1995. *Proc. Natl. Acad. Sci. USA* 92:8408–12
196. Correa PE. 1990. *Proteins: Struct. Funct. Genet.* 7:366–77
197. David CW. 1993. *J. Comp. Chem.* 14:715–17
198. Flores TP, Orengo CA, Moss DS, Thornton JM. 1993. *Protein Sci.* 2:1811–26
199. Wlodawer A, Deisenhofer J, Huber R. 1987. *J. Mol. Biol.* 193:145–56
200. Schrauber H, Eisenhaber F, Argos P. 1993. *J. Mol. Biol.* 230:592–612
201. Lee C. 1996. *Folding & Design* 1:1–12
202. Lee C, Levitt M. 1997. *Pac. Symp. Biocomput.*, Hawaii, pp. 245–55. River Edge, NJ: World Sci.
203. Lee C, McConnell HM. 1995. *Proc. Natl. Acad. Sci. USA* 92:8269–73
204. Lasters I, De Maeyer M, Desmet J. 1995. *Protein Eng.* 8:815–22
205. Koehl P, Delarue M. 1995. *Nat. Struct. Biol.* 2:163–70
206. Roitberg A, Elber R. 1991. *J. Chem. Phys.* 95:9277–87
207. Zheng Q, Kyle DJ. 1994. *Protein* 19:324–29
208. Chung SY, Subbiah S. 1995. *Protein Sci.* 4:2300–9
209. Chung SY, Subbiah S. 1996. *Pac. Symp. Biocomput.*, Hawaii, pp. 126–41. River Edge, NJ: World Sci.
210. Bowie JU, Läthy R, Eisenberg D. 1991. *Science* 253:164–70
211. Jones DT, Taylor WR, Thornton JM. 1992. *Nature* 358:86–89
212. Lemer CM-R, Rooman MJ, Wodak SJ. 1995. *Proteins: Struct. Funct. Genet.* 23:337–55
213. Hinds DA, Levitt M. 1994. *J. Mol. Biol.* 243:668–82
214. Rooman MJ, Kocher JPA, Wodak SJ. 1991. *J. Mol. Biol.* 221:961–80
215. Wilson C, Doniach S. 1989. *Proteins: Struct. Funct. Genet.* 6:193–209
216. Covell DG. 1992. *Proteins: Struct. Funct. Genet.* 14:409–20
217. Bowie JU, Eisenberg D. 1994. *Proc. Natl. Acad. Sci. USA* 91:4436–40
218. Covell DG. 1994. *J. Mol. Biol.* 235:1032–43
219. Dandekar T, Argos P. 1994. *J. Mol. Biol.* 236:844–61
220. Kolinski A, Skolnick J. 1994. *Proteins: Struct. Funct. Genet.* 18:338–52
221. Vieth M, Kolinski A, Brooks CLI, Skolnick J. 1994. *J. Mol. Biol.* 237:361–67
222. Wallqvist A, Ullner M. 1994. *Proteins: Struct. Funct. Genet.* 18:267–80
223. Monge A, Lathrop EJP, Gunn JR, Shenkin PS, Friesner RA. 1995. *J. Mol. Biol.* 247:995–1012
224. Srinivasan R, Rose GD. 1995. *Proteins: Struct. Funct. Genet.* 22:81–99
225. Vieth M, Kolinski A, Brooks CLI, Skolnick J. 1995. *J. Mol. Biol.* 251:448–67
226. Rose GD, Srinivasan R. 1996. *Biophys. J.* 70:A378
227. Yue K, Dill KA. 1996. *Protein Sci.* 5:254–61
228. Park BH, Levitt M. 1995. *J. Mol. Biol.* 249:493–507
229. Park B, Levitt M. 1996. *J. Mol. Biol.* 258:367–92
230. Mumenthaler C, Braun W. 1995. *Protein Sci.* 4:863–71
231. Hänggi G, Braun W. 1994. *FEBS Lett.* 344:147–53
232. Sun S-J, Thomas PD, Dill KA. 1995. *Protein Eng.* 8:769–78
233. Dill KA, Bromberg S, Yue K-Z, Fiebig KM, Yee DP, et al. 1995. *Protein Sci.* 4:561–602
234. Wang Y-H, Zhang H, Li W, Scott RA. 1995. *Proc. Natl. Acad. Sci. USA* 92:709–13
235. Wang Y-H, Zhang H, Scott RA. 1995. *Protein Sci.* 4:1402–11
236. Eisenberg D, McLachlan AD. 1986. *Nature* 319:199–203
237. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, et al. 1984. *J. Am. Chem. Soc.* 106:765–84
238. Miyazawa S, Jernigan RL. 1985. *Macromolecules* 18:534–52
239. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, et al. 1990. *J. Mol. Biol.* 216:167–80
240. Godzik A, Skolnick J. 1992. *Proc. Natl. Acad. Sci. USA* 89:98–102
241. Maiorov VN, Crippen GM. 1992. *J. Mol. Biol.* 227:876–88

242. Ouzounis C, Sander C, Scharf M, Schneider R. 1993. *J. Mol. Biol.* 232:805–25
243. Huang ES, Subbiah S, Tsai J, Levitt M. 1996. *J. Mol. Biol.* 257:716–25
244. Huang ES, Subbiah S, Levitt M. 1995. *J. Mol. Biol.* 252:709–20
245. Park BH, Huang ES, Levitt M. 1997. *J. Mol. Biol.* 266:831–36
246. Lim WA, Sauer RT. 1989. *Nature* 339:31–36
247. West MW, Hecht MH. 1995. *Protein Sci.* 4:2032–39
248. Sosnick TR, Jackson S, Wilk RR, Englander SW, De Grado WF. 1996. *Proteins: Struct. Funct. Genet.* 24:427–32
249. Finkelstein AV, Ptitsyn OB. 1987. *Prog. Biophys. Mol. Biol.* 50:171–90
250. Gerstein M, Sonnhammer D, Chothia C. 1994. *J. Mol. Biol.* 236:1067–78
251. Kapp OH, Moens L, Vanfleteren J, Trotman CNA, Suzuki T, Vinogradov SN. 1995. *Protein Sci.* 4:2179–90