THE PROTEIN FOLDING PROBLEM

Understanding and predicting the three-dimensional structures of proteins from their sequences of amino acids requires both basic knowledge of molecular forces and sophisticated computer programs that search for the correct configurations.

Hue Sun Chan and Ken A. Dill

Thousands of different types of proteins occur in biological organisms. They are responsible for catalyzing and regulating biochemical reactions, transporting molecules, the chemistry of vision and of the photosynthetic conversion of light to growth, and they form the basis of structures such as skin, hair and tendon. Protein molecules have remarkable structures. A protein is a linear chain of a particular sequence of monomer units. A major class of proteins, globular proteins, ball up into compact configurations that can have much internal symmetry. (See figure 1.) Each globular protein has a unique folded state, determined by its sequence of monomers.

The protein folding problem is to predict the compact three-dimensional structure from knowledge of the monomer sequence. It is one of the fundamental problems in biophysical science. Understanding the physics of protein conformations will be of great importance for biomedicine: in designing novel proteins, in decoding the genetic information obtained by the Human Genome Project, in designing new drugs and in trying to understand the structures and functions of the thousands of protein sequences that are being discovered every day in biotechnology labs.

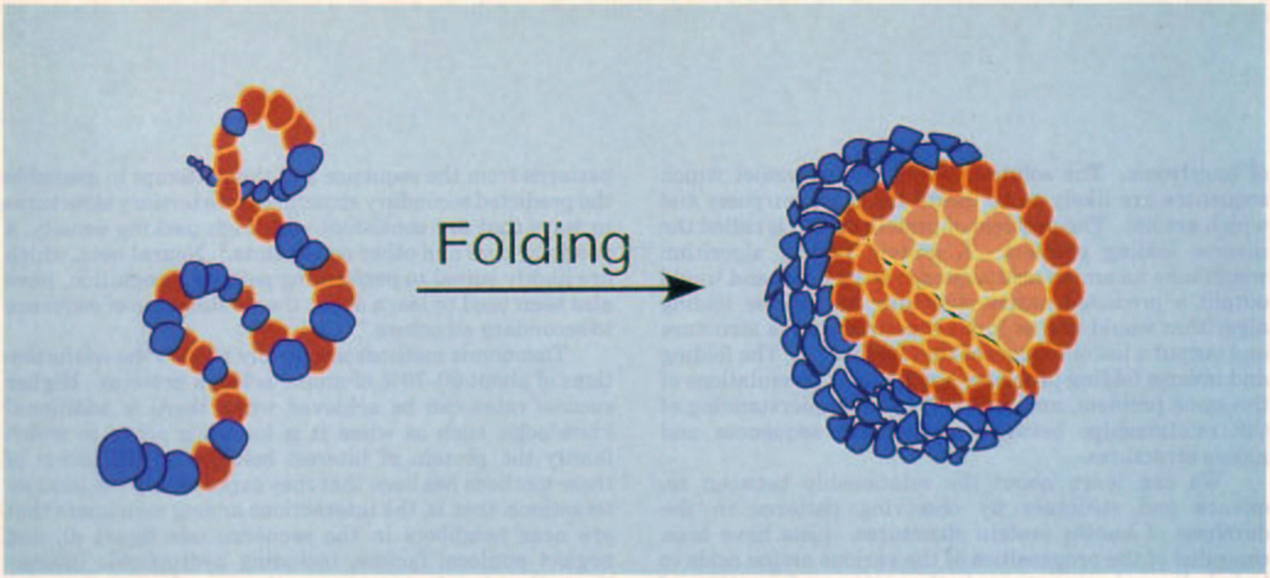
What are proteins?

A protein is a linear polymer molecule, a chain of tens to thousands of monomer units strung together like beads in a necklace. The monomers are the 20 naturally occurring amino acids. Different proteins have different sequences of the amino acid monomers, and the amino acid sequence is known as the primary structure of a protein. Proteins may be classified into three types: fibrous, membrane and globular. Fibrous proteins such as collagen, which contributes to tendon and bone, and α-keratin, which makes up hair, skin and feathers, serve mainly structural roles. Membrane proteins reside in cellular membranes, where they mediate the exchange of molecules and information across cellular boundaries.

The main focus of this article is globular proteins. Enzymes, which are the catalysts for virtually all biochemical reactions in living cells, are globular proteins. A typical cell contains 1000 to 4000 different enzymes. Although they are chain molecules, globular proteins have structures and properties quite different from those of other polymeric states of matter. Because the energy differences caused by internal bond rotations are small, polymers usually have many different conformations. (See figure 2.) Most synthetic polymeric materials are liquids, glasses, elastomers (such as rubber) or composite solutions, in which individual molecules have diverse conformations, most of which are open and interpenetrated by other molecules. In contrast, the most important state of a globular protein, known as its native or folded state, is extremely compact and is unique. That is, a given protein folds to only one native state (although the native states of different proteins can be quite similar). The so-called secondary structure of a globular protein includes hydrogen-bonded α-helices and β-sheets (the latter formed of two or more adjacent strands running parallel or antiparallel). (See figure 3.) The large-scale architecture of a protein—how the helices, sheets and other secondary structures fit together—is called its tertiary structure. Proteins are in their native states in aqueous solvents near neutral pH at 20–40 °C; this is the typical cellular environment. Under some nonphysiological conditions, such as high temperature, acidic or basic pH, or in some nonaqueous solvents, the unique folded structure of a protein unfolds or denatures, often reversibly, through a sharp transition to an ensemble of more expanded conformations.

The folding equilibrium is shown schematically in figure 1. Under physiological conditions the native state is marginally more stable (typically by about 40 kJ per mole of protein) than the ensemble of denatured conformations. Marginal stability may be necessary for biological function, since catalysis and binding properties of proteins must be responsive to the environment and to regulatory molecules. For example, hormones and biological signal-

Hue Sun Chan is an assistant research biophysicist and **Ken Dill** is a professor in the department of pharmaceutical chemistry and the department of biochemistry and biophysics at the University of California, San Francisco.



Folding of a globular protein from its denatured state (left) to its unique, compact native state (right) is encoded in its sequence of amino acid monomers. A complete understanding of this "second genetic code" continues to elude researchers. **Figure 1**

ing molecules cause conformational changes when they bind to their target proteins, and the ability to chemically degrade proteins is essential for the regulation of protein concentrations in the cell. Nevertheless marginal stability poses the problem for researchers of understanding the small net effect of large and diverse driving forces. Among the 20 amino acids, some have net charge, all can form hydrogen bonds, and about half of them are nonpolar to varying degrees. A large contribution to the balance of forces also comes from the decrease in conformational entropy upon folding.

The native state of a typical globular protein has several remarkable properties:

▷ It is as tightly packed as a small-molecule crystal, but it is usually devoid of the simple spatial regularity of a crystal.

▷ Amino acids are of different types, often classified as hydrophobic monomers (denoted by H), which are oil-like and interact unfavorably with water, and polar or charged monomers (denoted by P), which interact favorably with water. An example of an H monomer is the amino acid leucine; an example of a P monomer is serine. In native conformations of globular proteins, the H monomers tend to be buried inside the core of the globule, implying that proteins are driven to compactness by the force that causes oil and water to separate (the hydrophobic interactions, which we will discuss further below). P monomers tend to reside on the surface of the globule, although exceptions are common.

▷ Some proteins have beautiful symmetries in their secondary and tertiary structures, but other globular proteins have little internal symmetry. Proteins come in families of structures,[1] such as bundles of helices, or barrels or sandwiches of β-sheets.

▷ Each amino acid sequence folds into a unique native structure. DNA molecules in the genes encode the amino acid sequences. Most natural sequences are not simple periodic repeats of monomers.

▷ Under folding conditions, the native state is often thermodynamically stable (apart from small-amplitude fluctuations in the atomic positions, which can show glassy dynamics). In contrast, many synthetic polymeric materials are glassy and metastable, and their structures are dependent on their preparation history.

This set of properties has not been found in nonbiological polymers.

The second genetic code

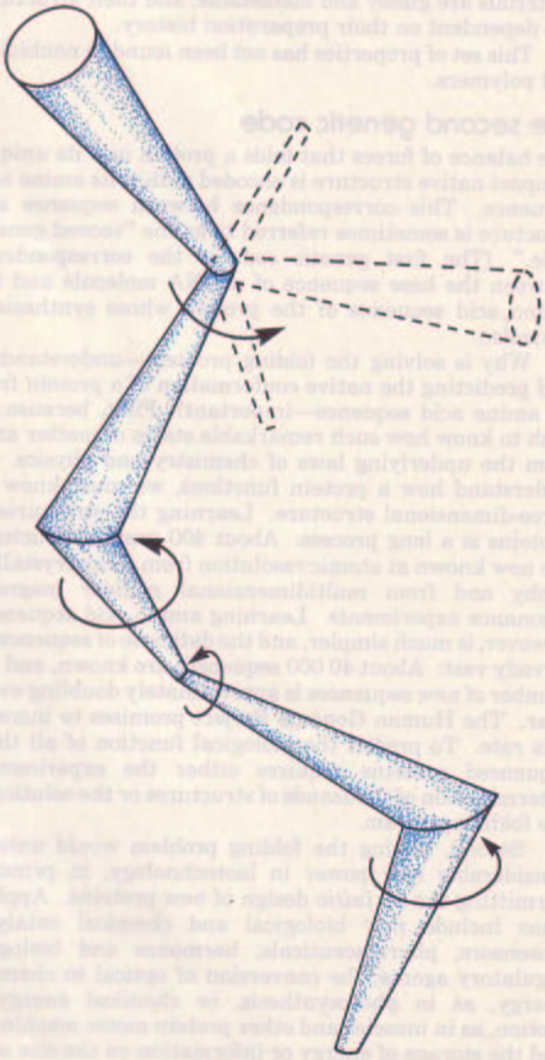
The balance of forces that folds a protein into its unique, compact native structure is encoded within its amino acid sequence. This correspondence between sequence and structure is sometimes referred to as the "second genetic code." (The first genetic code is the correspondence between the base sequence of a DNA molecule and the amino acid sequence of the protein whose synthesis it controls.)

Why is solving the folding problem—understanding and predicting the native conformation of a protein from its amino acid sequence—important? First, because we wish to know how such remarkable states of matter arise from the underlying laws of chemistry and physics. To understand how a protein functions, we must know its three-dimensional structure. Learning the structures of proteins is a long process: About 400 protein structures are now known at atomic resolution from x-ray crystallography and from multidimensional nuclear magnetic resonance experiments. Learning amino acid sequences, however, is much simpler, and the database of sequences is already vast: About 40 000 sequences are known, and the number of new sequences is approximately doubling every year. The Human Genome Project promises to increase this rate. To predict the biological function of all these sequenced proteins requires either the experimental determination of thousands of structures or the solution of the folding problem.

Second, solving the folding problem would unleash considerable new power in biotechnology, in principle permitting the *ab initio* design of new proteins. Applications include new biological and chemical catalysts; biosensors; pharmaceuticals; hormones and biological regulatory agents; the conversion of optical to chemical energy, as in photosynthesis, or chemical energy to motion, as in muscles and other protein motor machinery; and the storage of energy or information on the size scale of angstroms. The solution would let us predict which sequences are likely to be useful for these purposes and which are not. The problem of protein design is called the inverse folding problem: A protein folding algorithm would take an amino acid sequence as its input and would output a predicted native structure; an inverse folding algorithm would use as input a desired native structure and output a list of sequences that fold into it. The folding and inverse folding problems are different formulations of the same problem, and both call for an understanding of the relationships between amino acid sequences and native structures.

We can learn about the relationship between sequence and structure by observing patterns in the database of known protein structures. Lists have been compiled of the propensities of the various amino acids to be in helices, sheets or turns in the native conformations of proteins. Similar taxonomic lists now exist for pairs and triplets of amino acids, longer sequence fragments, short pieces of chains in loops and so on. Such lists form the basis of a strategy: Predict the likely secondary structure patterns from the sequence and then attempt to assemble the predicted secondary structures into tertiary structures in ways that are consistent with high packing density, a nonpolar core and other constraints.[2] Neural nets, which are highly suited to performing pattern recognition, have also been used to learn about the relationship of sequence to secondary structure.[3]

Taxonomic methods accurately predict the conformations of about 60–70% of amino acids in proteins. Higher success rates can be achieved when there is additional knowledge, such as when it is known *a priori* to which family the protein of interest belongs. A limitation of these methods has been that they capture only the local interactions, that is, the interactions among monomers that are near neighbors in the sequence (see figure 4), and neglect nonlocal factors, including hydrophobic interactions, that involve monomers that are far apart in the sequence. Peter G. Wolynes and his colleagues are developing a method that uses a neural-net-like procedure to "learn" the coefficients of both local and nonlocal energy terms in a semiempirical Hamiltonian function. Another approach to including the nonlocal interactions is that of Henrik Bohr and coworkers, in which neural nets "learn" distance matrices, that is, the spatial separations of pairs of amino acids in proteins.[3]

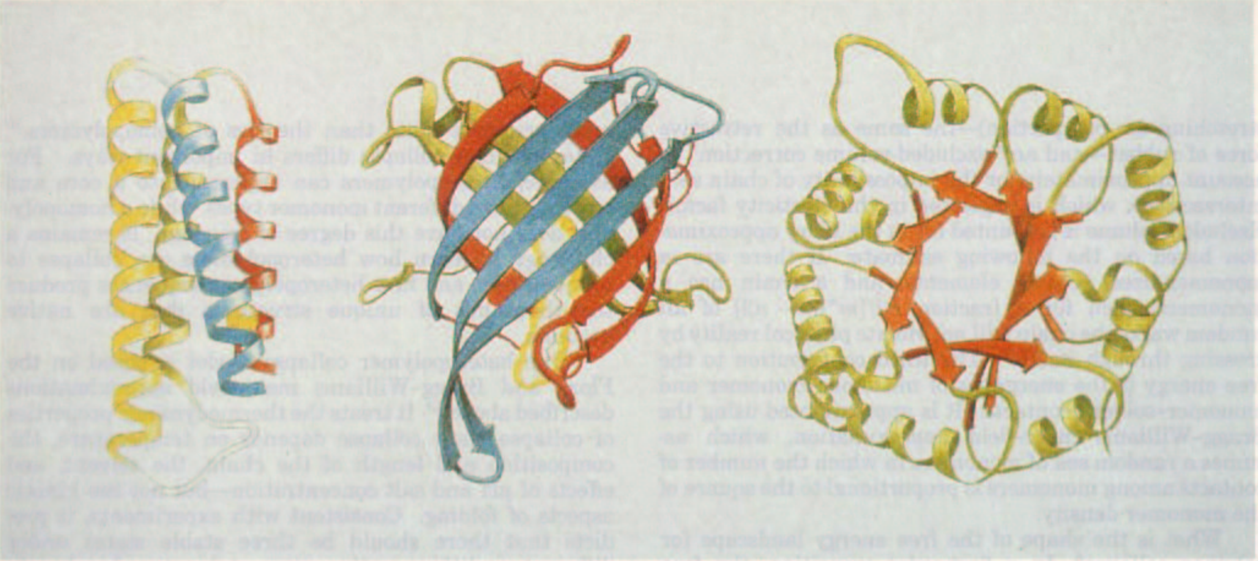


Physical driving forces

Ultimately we would like to go beyond such heuristic approaches and understand how to fold a protein based on the physical driving forces. To find the stable native state of a protein, ideally we should compute, for every possible conformation of the chain, the sum of the free energies of the atomic interactions within the protein and with the solvent and then find the conformation with the lowest free energy. But this is not feasible, because the number of conformations N of a chain molecule grows exponentially with the chain length: $N \sim \mu^n$, where n is the number of monomers and $\mu \simeq 2$–6 is the number of rotational isomers, determined by the types of monomers that make up the polymer. (See figure 2.) An exhaustive search is not a practical solution to the folding problem for a computer algorithm. Nor is it practical for a real protein: This is the Levinthal paradox, named after Cyrus Levinthal,[4] who first raised the question, How does a protein find the global optimum (its native state) without a global search? Proteins fold much faster—by tens of orders of magnitude—than the time a chain molecule would need to undertake a global search. What vast parts of conformational space does the protein avoid?

Proteins are not the only physical systems that find thermodynamically stable states in the face of large numbers of degrees of freedom. Physical systems settle into equilibrium states by processes that are seldom random or exhaustive; rather they are directed by energies. The free energy as a function of the degrees of freedom is the energy landscape, or conformational space.

Different interatomic bond conformations (generated by the indicated rotations) have only small energy differences, allowing many overall conformations of a polymer chain to arise. **Figure 2**



Structures of globular proteins include ordered assemblies of helices (left) and sheets (middle) and mixtures of helices and sheets (right). Other proteins have less regularity. (From C. Branden, J. Tooze, *Introduction to Protein Structure*, Garland, New York, 1991.) **Figure 3**

One way to explore such landscapes is by molecular dynamics techniques, in which a computer numerically solves Newton's laws of motion using interaction energies obtained from experiments on smaller molecules.

These force-field simulations[5] have contributed much to our understanding of proteins and polymers, but to find the global minimum for a protein is to search for a needle in a very large haystack. Even a small protein contains tens of thousands of atoms, and to treat all the major forces properly the simulation must include the surrounding water molecules, adding thousands of additional atoms. Moreover, because the harmonic motions of bonded atoms have characteristic times of around 10^{-14}–10^{-13} seconds, stable numerical integration requires femtosecond (10^{-15} second) time steps. Supercomputers can currently simulate up to nanoseconds of real-time protein dynamics with such short time steps, but this scale doesn't approach the 10^{-1}–10^{3} seconds typically required to fold real proteins. Even though parallel processing supercomputer power is increasing about a thousandfold every 10 years, it could be 10–30 years before brute-force molecular dynamics reliably folds proteins. (See figure 5.) Success in folding proteins by molecular dynamics will also require improvements in the accuracy of the simulated force fields.

The shape of the energy landscape is determined by the forces of folding. Alfred Mirsky and Linus Pauling proposed in 1936 that hydrogen bonding is the dominant force of folding.[6] Pauling, R. B. Corey and Herman R. Branson built models of chains of amino acids to determine the peptide bond geometry.[7] By finding conformations that make good hydrogen bonds, they discovered α-helices and β-sheets, and predicted they would be important components of proteins. As figure 3 illustrates, their prediction was correct.

But in the 1950s, Walter Kauzmann pointed out that hydrogen bonding would not strongly favor the folded state relative to unfolded states, because unfolded conformations can form hydrogen bonds with water that should be just as strong as the intrachain hydrogen bonds in the folded state.[8] He felt that hydrophobic interactions were a stronger force for folding proteins. Despite many theoretical and experimental studies since Kauzmann's work, however, the molecular details of hydrophobic interactions are not yet clearly understood. In thermodynamic terms, we know that the mixing of nonpolar, oil-like molecules with water has a large positive free energy, is disfavored by entropy near room temperature and leads to a large increase in heat capacity. The most common interpretations of these effects involve orientational ordering of water molecules upon the dissolving of a nonpolar substance.

In the 20 years following Kauzmann's observation, the view emerged that hydrophobic interactions nonspecifically favor compactness and that hydrogen bonds and local interactions determine the detailed internal architecture and sequence-dependent uniqueness of a native conformation. A different view has recently entered into protein folding research—that hydrophobicity and nonlocal interactions are a major factor in causing not only the compactness but also the uniqueness and internal architectures of globular proteins. (References 9 and 10 review this viewpoint.)

Homopolymer collapse theories

What drives a polymer to become compact? A polymer chain composed of oil-like monomers will ball up in water to minimize the area of unfavorable monomer–water contacts. But since there are far fewer compact than expanded conformations of chain molecules, the greater conformational entropy in the expanded state will oppose collapse. The balance of these forces will determine the average chain compactness.

These ideas, rooted in the work of Paul J. Flory in 1949, led to the first theory of the collapse of homopolymers (polymers composed of a single species of monomer), developed by Oleg B. Ptitsyn and Yuili Eizner in 1965, and to subsequent mean-field models.[10,11] According to these models, changing the strength of the monomer–monomer attraction leads to a sharp collapse from open to compact conformations. In this approach, the chain is assumed to follow a three-dimensional random walk, and three terms contribute to the free energy as a function of the chain compactness. The first two come from the entropy, which is assumed to be factorable into two parts: "elasticity," which originates from the reduction of entropy on stretching (or compaction)—the same as the retractive force of rubber—and an "excluded-volume correction" to account approximately for the impossibility of chain self-intersections, which is neglected in the elasticity factor. Excluded volume is accounted for in the Flory approximation based on the following estimate: If there are m monomer-sized volume elements, and a chain has n monomers, then for a fraction $m!/[m^n(m-n)!]$ of all random walks the chain will not violate physical reality by crossing through itself.[10] The third contribution to the free energy is the energetics of monomer–monomer and monomer–solvent contacts. It is approximated using the Bragg–Williams mean-field approximation, which assumes a random sea of monomers in which the number of contacts among monomers is proportional to the square of the monomer density.

What is the shape of the free energy landscape for polymer collapse? In a first-order transition, the free energies of native and denatured states would be minima separated by a free-energy barrier. In a higher-order transition, there would be no barrier. Some mean-field models predict a first-order transition, but it is possible that the free energy barrier in those models is an artifact of their approximations.

There have been several improvements in collapse theories. Sam F. Edwards introduced a self-consistent field approach in 1965 to model self-avoiding chains more accurately. Here the excluded-volume repulsion between individual monomers is approximated as a field that is self-consistently determined as a function of the monomer density. In 1968 Ilya M. Lifshitz proposed a general self-consistent field formalism for the study of polymer collapse; it has been further developed by his coworkers Alexander Yu. Grosberg and Alexey R. Khokhlov. In a first approximation similar to the approach of Flory, they found a second-order transition for the collapse of infinitely long homopolymer chains. More recently, by allowing for a nonuniform spatial density distribution of monomers and by treating the entropic restrictions on chain turns at globule surfaces, Grosberg and Dmitry V. Kuznetsov found that coil–globule transitions for finite-length homopolymers are considerably sharper—more like first-order transitions. A principal difficulty in devising refined theories is the many-body nature of the chain self-collisions in compact conformations.[12]

Models of heteropolymer collapse

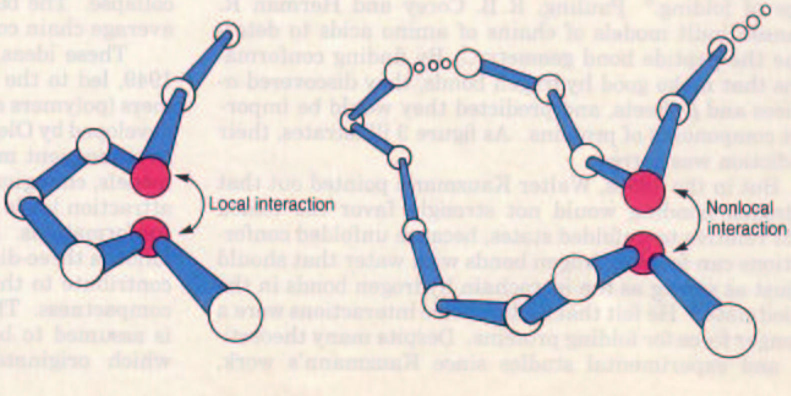
Theories of the collapse of heteropolymers (polymers made of more than one monomer type), such as proteins, are in a more primitive state than theories of homopolymers.[10] Heteropolymer collapse differs in important ways. For example, heteropolymers can organize into a core and surface of two different monomer types, while a homopolymer does not have this degree of freedom. It remains a challenge to learn how heteropolymers can collapse to unique states and how heteropolymer sequences produce the thousands of unique structures that are native proteins.

One heteropolymer collapse model is based on the Flory and Bragg–Williams mean-field approximations described above.[13] It treats the thermodynamic properties of collapse—how collapse depends on temperature, the composition and length of the chain, the solvent, and effects of pH and salt concentration—but not the kinetic aspects of folding. Consistent with experiments, it predicts that there should be three stable states under different conditions: native, compact denatured and highly unfolded.

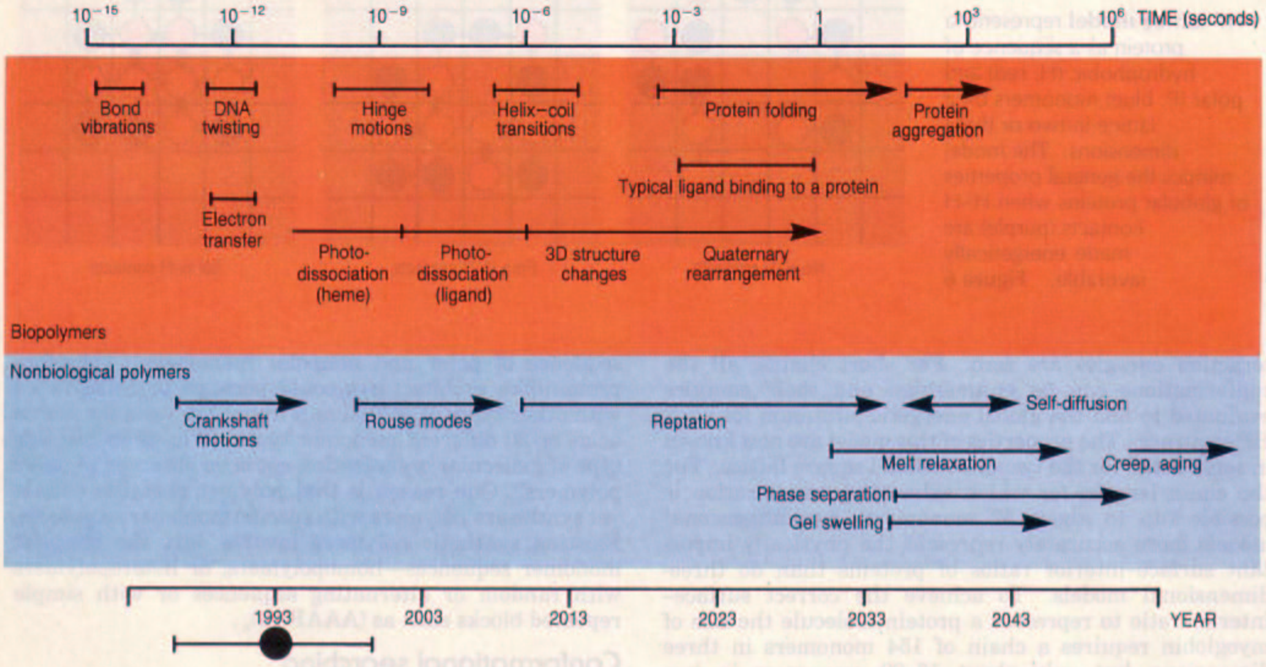
Another approach is based on spin glass models.[14] The concept of spin glass was first proposed by Edwards and Philip W. Anderson in 1975 to account for the magnetic properties of dilute alloys of manganese in copper. (See the Reference Frame columns by Anderson in PHYSICS TODAY, January, March, June and September 1988, July and September 1989, and March 1990.) Applications of spin glass methods to proteins do not try to model the folding of a specific amino acid sequence to a specific structure. Rather they consider statistical ensembles of amino acid sequences, modeled by assigning random interaction energies between monomers on a chain. By averaging over the ensemble these methods seek to learn about the folding process itself.

Joe D. Bryngelson and Wolynes were the first to apply spin glass concepts to the coil-to-globule folding of proteins, in 1987. In their model, interactions between monomers are assumed to be randomly distributed, as in Bernard Derrida's 1981 random-energy model. They predict different folding and "freezing" transitions of a heteropolymer: A chain may *fold* into a given native structure specified in advance or *freeze* into a collection of "misfolded" (non-native) structures that have extremely slow dynamics of interconversion. Other model studies have also found that the kinetic accessibility of the native structure is strongly sequence dependent.[15]

Other spin glass models[14] include one introduced in 1988 by Thomas Garel and Henri Orland. In their heteropolymer model of freely jointed chains, the pair interaction B_{ij} between monomers i and j is a random



Interactions in polymers may be divided into local (those among near neighbors in the sequence) and nonlocal (those among monomers that are far apart in the sequence). The importance of both types of interaction contributes to the difficulty of modeling folding. **Figure 4**



Time scales for various motions within biopolymers (red) and nonbiological polymers (blue). The year scale at the bottom shows estimates of when each such process might be accessible to force-field simulation on supercomputers, assuming that parallel processing capability on supercomputers increases at about the rate of 10^3 every 10 years and neglecting new approaches or breakthroughs. At current capabilities, a given allotment of computer time can be used for one run performed over a few hundred picoseconds for a small protein in a few thousand water molecules, or for one thousand runs to explore a thousand processes that have relaxation times of hundreds of femtoseconds; this range is indicated by the error bar below the year scale. **Figure 5**

variable. Then the spin glass procedure of averaging over different "replicas" is carried out. (See Anderson's June 1988 column.) B_{ij} is a parameter that represents the heterogeneity of interactions. Eugene I. Shakhnovich and Alexander M. Gutin developed a heteropolymer model in which the distribution of monomer-pair interaction strengths B_{ij} was assumed to be Gaussian and which included a three-monomer hard-core repulsion term. In their theory the width B of the heterogeneity distribution B_{ij} plays the crucial role of determining the number of lowest-energy states of the model. If the sequences are sufficiently heterogeneous (B large), Shakhnovich and Gutin find that only a few states dominate in the low-temperature phase. Thus they conclude that unique protein folds can arise simply from sequence heterogeneity.

Collapse theories show that heteropolymers can undergo sharp transitions that resemble protein folding, from open ensembles to compact conformations with solvent-averse (H) monomers sequestered into a core. Moreover, from the spin glass mode!s described above and exact models described in the next section, it is clear that heteropolymers can collapse to only a very small number of compact conformations. This contrasts sharply with the situation for homopolymers, which collapse to large ensembles of compact conformations, and it suggests that the uniqueness of protein native states may be largely encoded in the nonlocal interactions (mainly the pattern of hydrophobic monomers in the sequence) rather than in the local interactions. The limitation of existing heteropolymer collapse theories is that they consider only the composition (the number of monomers of each type) of a sequence and, in some of the spin glass models, the distribution of interaction energies, but otherwise they assume the sequences are random.

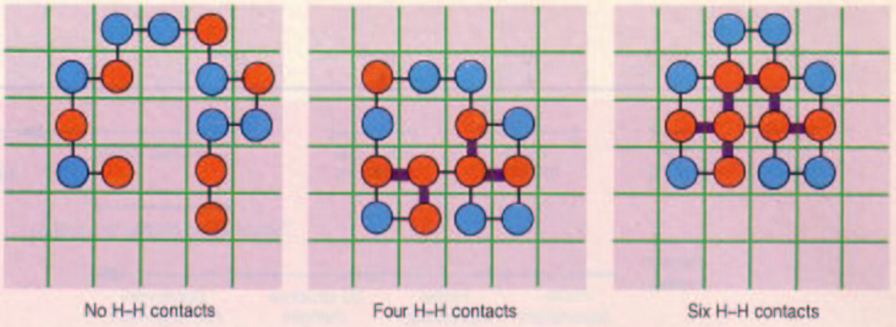
Simplified exact models

Exploring the relationships of amino acid sequences to native structures requires models different from existing collapse theories, which average out the effects of the sequence, and atomic-resolution molecular dynamics simulations, which are limited by computational restrictions. To explore sequence–structure relationships, a class of model has emerged in which proteins are represented as self-avoiding walks on lattices. Specific sequences of monomers are studied in chains short enough that the full conformational space can be enumerated exhaustively.

The first exact enumeration of short chains on lattices was the work of W. J. C. Orr in 1947. Orr enumerated all the self-avoiding conformations on lattices for chains less than 10 monomers long. With improvements in computer technology, Cyril Domb, M. F. Sykes and their coworkers explored longer chains,[16] providing the underpinnings for many of the modern developments in polymer theory, including scaling laws and renormalization-group methods.

Similar in spirit is the HP lattice model for proteins,[17] shown in figure 6. Chains are configured as self-avoiding walks on two-dimensional square lattices or three-dimensional simple cubic lattices. Based on the assumption that the hydrophobic interaction is the dominant force in protein folding,[9] a protein is modeled as a specific sequence of hydrophobic (H) and hydrophilic (P) monomers (for example, PHHPHP . . .). Each interaction between two H monomers that are adjacent in space but not covalently linked is favored by a contact energy $\epsilon < 0$, and all other in-

HP lattice model represents a protein as a sequence of hydrophobic (H, red) and polar (P, blue) monomers on a lattice in two or three dimensions. The model mimics the general properties of globular proteins when H–H contacts (purple) are made energetically favorable. **Figure 6**



teraction energies are zero. For short chains, all the conformations can be enumerated and their energies evaluated to find the global energetic minimum for each HP sequence. The properties of this model are now known in some detail for the two-dimensional square lattice. For the chain lengths for which exhaustive enumeration is possible (up to about 30 monomers), two-dimensional models more accurately represent the physically important surface–interior ratios of proteins than do three-dimensional models. To achieve the correct surface–interior ratio to represent a protein molecule the size of myoglobin requires a chain of 154 monomers in three dimensions, but only about 16–20 monomers in two dimensions.

The two-dimensional HP lattice model mimics the general properties of globular proteins. Under conditions that favor denaturation (H–H attraction small), the chains populate a relatively large ensemble of conformations, corresponding to the denatured states of proteins. With increasing H–H attraction, the chains undergo a relatively sharp transition to a small ensemble of conformations (for many sequences, only one or a few) that are compact and have nonpolar cores. Helices and sheets arise in these lattice models as a consequence of the compactness of the chain.[10] (See figure 7.) That is, with increasing compactness, helices and sheets (particularly short ones) become increasingly probable because the severe steric constraints make alternative configurations nonviable. Hence the driving forces for collapse contribute substantially to the development of secondary structure. The length distributions of the helices and sheets that arise from the compactnesses of chains on lattices are similar to those for the known proteins, but these "packing forces" that stabilize secondary structures are not very specific: To acquire the precise bond geometries of secondary structures in real proteins also requires hydrogen bonding and local interactions.

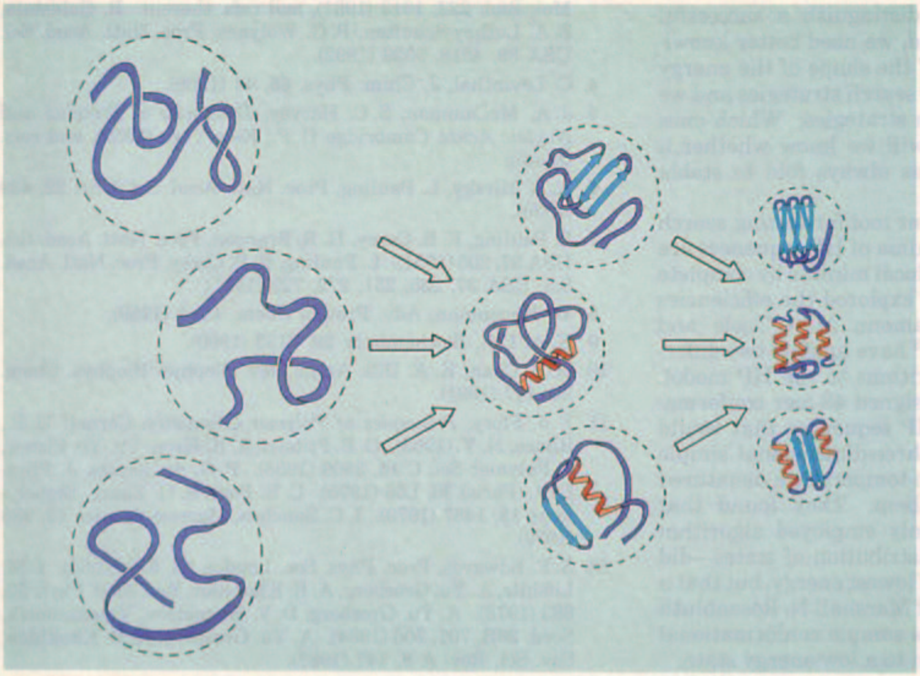
Evolutionary aspects of the HP lattice model, such as the effect of mutations, have also been studied. The model shows "mutational plasticity" corresponding to that of real proteins. (Plasticity is the ability of a protein to retain its native state in the face of small changes in sequence.) Sequence convergence is common. That is, a given native structure can be encoded by a large number of different sequences, consistent with protein mutation experiments.[18]

The view that emerges from the simple models is that the nonlocal interactions encoded in the HP sequence are sufficient to lead to several of the main features of proteins listed at the beginning of this article. A typical sequence can collapse through a relatively sharp transition to a small number of compact conformations, each with a core of H monomers and made up of helices and sheets. This suggests that the unique architecture of each individual globular protein may be encoded mainly by the specific sequence of polar and nonpolar monomers. Therefore protein-like architectures could perhaps be constructed with other types of polymers, without the need for amino acids or 20 different monomer types. Why then has this type of molecular organization not been observed in other polymers? One reason is that polymer chemists cannot yet synthesize polymers with specific monomer sequences. Existing synthetic polymers involve only the simplest monomer sequences—homopolymers, or heteropolymers with random or alternating sequences or with simple repeated blocks such as $(AAABBB)_n$.

Conformational searching

A major obstacle to folding proteins by computer is the challenge of searching the large and complex energy landscape to find the most stable states. Various statistical sampling methods are being explored, including Monte Carlo techniques, simulated annealing and so-called genetic algorithms.[19] Efficiency can be increased in these studies by using lower-resolution representations of proteins, that is, by averaging over certain degrees of freedom. For example, rather than representing each atom explicitly, one can take whole amino acids or clusters of a few atoms as the individual sites of interaction. Or, instead of using a continuum representation, one can treat chains as self-avoiding walk conformations on lattices. Jeffrey Skolnick, Andrzej Kolinski and their coworkers have developed high-resolution lattices, such as the lattice designated (2,1,0), in which each bond involves 2 steps along one of the three axes of a cubic lattice, 1 step along another axis, and 0 steps along the third. They have developed potential functions for this and other lattices that include local and nonlocal interactions. The end states of their Monte Carlo simulations resemble several of the major structural motifs of globular proteins.

One general strategy for solving the protein folding problem is based on the premise that low-resolution methods can survey the broad landscape to reach near-native states and thereby reduce any subsequent conformational search by higher-resolution methods. Low-resolution models require energy parameters. Recently, a popular approach has been to use "statistical potentials," in which the frequencies of pairs and sometimes triplets of amino acids within close spatial proximity in known native proteins are tabulated and used as if they were interaction free energies.[20,21] These potentials have been tested by exhaustive searching of restricted regions of conformational space within which the native structure is known to lie. The results have been encouraging: The true native structures have low energies. Usually, however, there are also structures with lower energies that have some incorrect folds. Statistical potentials have also been very useful for inverse folding, that is, for testing a sequence against a known structure to see if it will fold to it.[20]

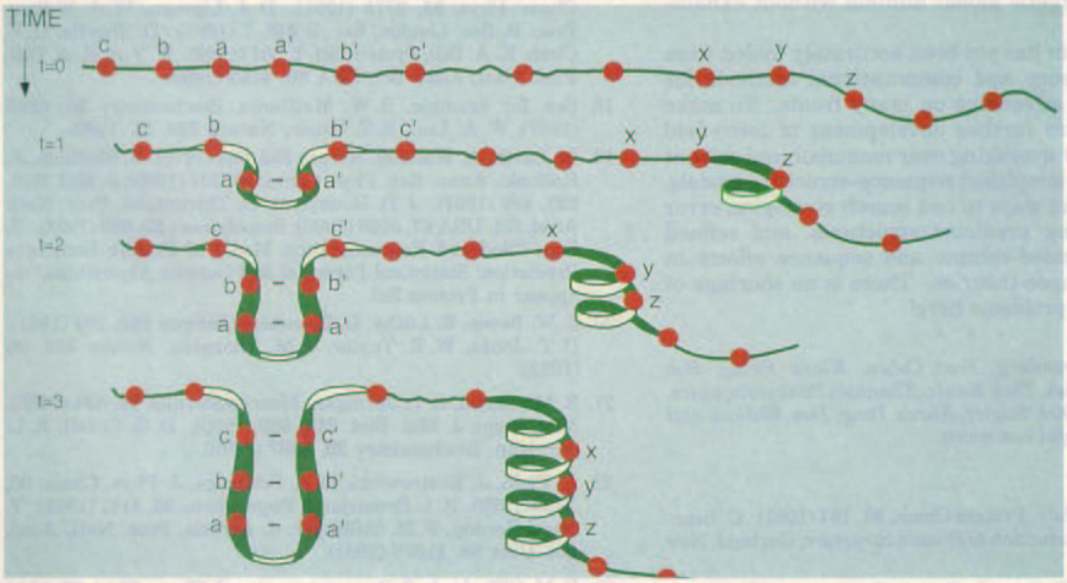


Confinement of a polymer chain to a compact region is an important factor causing helical and sheet substructures. In proteins, compactness appears to be driven mainly by hydrophobic interactions. **Figure 7**

Another approach has been to transform the energy landscape to remove local minima and make the global minimum more accessible.[22] For example, Lucjan Piela and coworkers suppose that the hills and valleys of the energy landscape are like the peaks and troughs of a hypothetical temperature distribution, which is then transformed according to the time-dependent heat conduction equation to become a smoother surface. Teresa Head-Gordon and coworkers change the landscape by using physical-chemical insights about amino acid dimers to allow only the correct chiralities, favored isomers and so on. These approaches are still in the early stages.

The ultimate aim of these efforts in conformational searching is to use only knowledge of the amino acid sequence to predict protein structure. However, even for low-resolution methods, conformational space is still a very large and tortuous place! To make the problem more manageable, some current strategies require help from prior knowledge of the native structure, such as assumed secondary structures, surface shape or experimental distance constraints, or they require extra nonphysical forces that incorporate some information about the desired native structure.

In our view, three hurdles remain to folding proteins by computer. First, we need better ways to assess errors and to determine the structural family of any arbitrary



Hydrophobic zipper hypothesis for how proteins might find their native states without exhaustive exploration of conformational space: Hydrophobic monomers (red circles) first pair up locally. This brings other hydrophobic monomer pairs into spatial proximity so they too can then pair up, and so on. This process can lead to compact chains that have cores of hydrophobic monomers and contain helices and sheets. **Figure 8**

conformation, so that we can distinguish a successful prediction from a failure. Second, we need better knowledge of the driving forces and of the shape of the energy landscape. Third, we need faster search strategies and we need to know more about search strategies: Which ones can find global minima? How will we know whether a minimum is global? Do proteins always fold to stable states?

The HP model is a convenient tool for testing search strategies because the global minima of HP sequences are known and distinguishable from local minima by complete enumeration. Two studies have explored the efficiencies of search strategies. First, Eamonn M. O'Toole and Athanassios Z. Panagiotopoulos[23] have applied two different Monte Carlo sampling algorithms to the HP model. O'Toole and Panagiotopoulos designed 48-mer conformations, to which they assigned HP sequences that would have a hydrophobic core on a three-dimensional simple cubic lattice. They sampled high-temperature denatured conformations, then recooled them. They found that Metropolis sampling—a commonly employed algorithm for estimating the Boltzmann distribution of states—did not return the system to a state of lowest energy, but that a variant of the sampling method of Marshall N. Rosenbluth and Arianna W. Rosenbluth does sample conformational space efficiently enough to return to a low-energy state.

The second search strategy explored using the HP model is based on the hypothesis that proteins may avoid exhaustive searching by folding along "hydrophobic zipper" pathways. That is, H monomers close to each other in the sequence come together first to form an H–H contact, with only small conformational searching, and this in turn brings other H monomers into proximity to form a next H–H contact—and so on until many H monomers have paired together to form a hydrophobic core. (See figure 8.) It is found that hydrophobic zipper processes, while not exhaustive, nevertheless lead to single global minima for about 70% of all possible HP sequences in the short-chain two-dimensional HP lattice model.[24]

These are just two among many possible search strategies by which proteins, and perhaps computers, might find the energetic global minima without exhaustive searching.

While no protein has yet been accurately folded from first principles, theory and computational methods for protein folding are advancing on many fronts. To make progress will require further development of force-field models, methods for averaging over monomer and solvent degrees of freedom, simplified sequence–structure models, search strategies and ways to test search strategies, error measures for testing predicted structures, and refined treatments of excluded-volume and sequence effects in heteropolymer collapse theories. There is no shortage of interesting physics problems here!

* * *

We thank Sarina Bromberg, Fred Cohen, Klaus Fiebig, Bob Jernigan, Peter Kollman, Tack Kuntz, Thanasis Panagiotopoulos, Jan W. H. Schreurs, Dirk Stigter, Karen Tang, Don Wallace and Peter Wolynes for helpful comments.

References

1. J. S. Richardson, Adv. Protein Chem. **34**, 167 (1981). C. Branden, J. Tooze, *Introduction to Protein Structure*, Garland, New York (1991).
2. W. R. Taylor, ed., *Patterns in Protein Sequence and Structure*, Springer-Verlag, New York (1992).
3. D. G. Kneller, F. E. Cohen, R. Langridge, J. Mol. Biol. **214**, 171 (1990). H. Bohr, J. Bohr, S. Brunak, R. M. J. Cotterill, H. Fredholm, B. Lautrup, S. B. Petersen, FEBS Lett. **261**, 43 (1990). M. S. Friedrichs, R. A. Goldstein, P. G. Wolynes, J. Mol. Biol. **222**, 1013 (1991), and refs. therein. R. Goldstein, Z. A. Luthey-Schulten, P. G. Wolynes, Proc. Natl. Acad. Sci. USA **89**, 4918, 9029 (1992).
4. C. Levinthal, J. Chim. Phys. **65**, 44 (1968).
5. J. A. McCammon, S. C. Harvey, *Dynamics of Proteins and Nucleic Acids*, Cambridge U. P., New York (1989), and refs. therein.
6. A. E. Mirsky, L. Pauling, Proc. Natl. Acad. Sci. USA **22**, 439 (1936).
7. L. Pauling, R. B. Corey, H. R. Branson, Proc. Natl. Acad. Sci. USA **37**, 205 (1951). L. Pauling, R. B. Corey, Proc. Natl. Acad. Sci. USA **37**, 235, 251, 272, 729 (1951).
8. W. Kauzmann, Adv. Protein Chem. **14**, 1 (1959).
9. K. A. Dill, Biochemistry **29**, 7133 (1990).
10. H. S. Chan, K. A. Dill, Annu. Rev. Biophys. Biophys. Chem. **20**, 447 (1991).
11. P. J. Flory, *Principles of Polymer Chemistry*, Cornell U. P., Ithaca, N. Y. (1953). O. B. Ptitsyn, A. K. Kron, Yu. Ye. Eizner, J. Polymer Sci. C **16**, 3509 (1968). P.-G. de Gennes, J. Phys. Lett. (Paris) **36**, L55 (1975). C. B. Post, B. H. Zimm, Biopolymers **18**, 1487 (1979). I. C. Sanchez, Macromolecules **12**, 980 (1979).
12. S. F. Edwards, Proc. Phys. Soc. London **85**, 613 (1965). I. M. Lifshitz, A. Yu. Grosberg, A. R. Khokhlov, Rev. Mod. Phys. **50**, 683 (1978). A. Yu. Grosberg, D. V. Kuznetsov, Vysokomolek. Soed. **26B**, 701, 706 (1984). A. Yu. Grosberg, A. R. Khokhlov, Sov. Sci. Rev. A **8**, 147 (1987).
13. K. A. Dill, Biochemistry **24**, 1501 (1985). D. O. V. Alonso, K. A. Dill, D. Stigter, Biopolymers **31**, 1631 (1991).
14. B. Derrida, Phys. Rev. B **24**, 2613 (1981). J. D. Bryngelson, P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987); J. Phys. Chem. **93**, 6902 (1989); Biopolymers **30**, 177 (1990). T. Garel, H. Orland, Europhys. Lett. **6**, 307, 597 (1988). E. I. Shakhnovich, A. M. Gutin, Biophys. Chem. **34**, 187 (1989).
15. P. G. Wolynes, in *Biologically Inspired Physics*, L. Peliti, ed., Plenum, New York (1991), p. 15. E. Shakhnovich, G. Farztdinov, A. M. Gutin, M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991). P. E. Leopold, M. Montal, J. N. Onuchic, Proc. Natl. Acad. Sci. USA **89**, 8721 (1992).
16. W. J. C. Orr, Trans. Faraday Soc. **43**, 12 (1947). C. Domb, Adv. Chem. Phys. **15**, 229 (1969), and refs. therein.
17. K. F. Lau, K. A. Dill, Macromolecules **22**, 3986 (1989); Proc. Natl. Acad. Sci. USA **87**, 638 (1990). H. S. Chan, K. A. Dill, J. Chem. Phys. **95**, 3775 (1991). D. J. Lipman, W. J. Wilbur, Proc. R. Soc. London, Ser. B **245**, 7 (1991). D. Shortle, H. S. Chan, K. A. Dill, Protein Sci. **1**, 201 (1992). K. Yue, K. A. Dill, Proc. Natl. Acad. Sci. USA **89**, 4163 (1992).
18. See, for example, B. W. Matthews, Biochemistry **26**, 6885 (1987); W. A. Lim, R. T. Sauer, Nature **339**, 31 (1989).
19. M. Levitt, A. Warshel, Nature **253**, 694 (1975). J. Skolnick, A. Kolinski, Annu. Rev. Phys. Chem. **40**, 207 (1989); J. Mol. Biol. **221**, 499 (1991). J. D. Honeycutt, D. Thirumalai, Proc. Natl. Acad. Sci. USA **87**, 3526 (1990); Biopolymers **32**, 695 (1992). S. Sun, "Reduced Representation Model of Protein Structure Prediction: Statistical Potential and Genetic Algorithms," to appear in Protein Sci.
20. J. W. Bowie, R. Lüthy, D. Eisenberg, Science **253**, 164 (1991). D. T. Jones, W. R. Taylor, J. M. Thornton, Nature **358**, 86 (1992).
21. S. Miyazawa, R. L. Jernigan, Macromolecules **18**, 534 (1985). M. J. Sippl, J. Mol. Biol. **213**, 859 (1990). D. G. Covell, R. L. Jernigan, Biochemistry **29**, 3287 (1990).
22. L. Piela, J. Kostrowicki, H. A. Scheraga, J. Phys. Chem. **93**, 3339 (1989). R. L. Somorjai, J. Phys. Chem. **95**, 4141 (1991). T. Head-Gordon, F. H. Stillinger, J. Arrecis, Proc. Natl. Acad. Sci. USA **88**, 11076 (1991).
23. E. M. O'Toole, A. Z. Panagiotopoulos, J. Chem. Phys. **97**, 8644 (1992).
24. K. A. Dill, K. M. Fiebig, H. S. Chan, "Cooperativity in Protein Folding Kinetics," to appear in Proc. Natl. Acad. Sci. USA **90** (1993). K. M. Fiebig, K. A. Dill, "Protein Core Assembly Processes," to appear in J. Chem. Phys. **98** (1993). ■