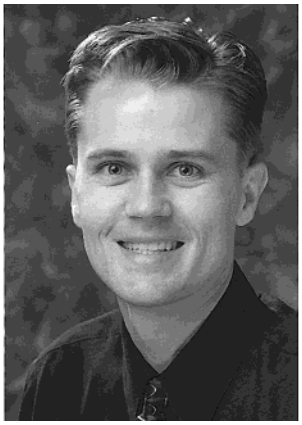
Designing Protein Energy Landscapes

Jeffery G. Saven

Deptartment of Chemistry, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Received March 13, 2001

Contents



Born in Manhattan, KS, Jeffery G. Saven obtained his B.A. degree in Chemistry from New College of the University of South Florida in 1988. As an NSF Graduate Fellow, he earned his Ph.D. degree in Chemical Physics from Columbia University in the city of New York. However, the bulk of his dissertation research was carried out working with James L. Skinner in the Department of Chemistry at the University of Wisconsin, Madison, where he studied applications of statistical mechanics and simulation to condensed phase spectroscopy. He was an NSF Postdoctoral Fellow at the University of Illinois, Urbana-Champaign, where he studied protein folding with Peter G.Wolynes. Saven has been an Assistant Professor in the Department of Chemistry at the University of Pennsylvania since 1997. His honors include a Research Innovation Award and an NSF CAREER Award, and he is also a Cottrell Scholar and an Arnold and Mabel Beckman Young Investigator. His research interests include the theory and simulation of molecular folding and combinatorial design.

I. Overview

The concepts of protein folding span biology, physics, and chemistry and have applications to biomedicine and biomaterials. Since proteins are the direct products of genes, folding is fundamental to the expression of genetic information in the cell. Folding is also of fundamental physical interest, since it involves spontaneous ordering at the molecular scale. With few exceptions, proteins fold reversibly to unique structures. The three-dimensional folded structure of a protein is encoded in its sequence of amino acids. Thus, it may eventually be possible to predict structure from sequence alone and to design desired folded structures through careful choice of sequence. Important goals in the field include determining structure from gene sequence, re-engineering existing proteins, and crafting new ones de novo. Using synthetic sequences, features important in protein stability and folding kinetics may be probed via selective mutations. Once particular structures can successfully designed, the opportunity then exists for the design of novel functional proteins. Potentially, these ideas can be expanded beyond the naturally occurring biopolymers. Folding polymers, both biological and synthetic, could yield new types of structures and properties and lead to novel pharmaceuticals, catalysts, and materials.

A predictive understanding of molecular folding is obscured by the complexity of proteins. The hallmark of folding is the ability of an amino acid sequence to reversibly acquire a well-defined, unique structure in a moderate amount of time even though an exponentially large number are possible. Another level of complexity concerns the uncertainty with which the forces that guide folding are quantitatively understood. The stabilizing interactions within a folded or partially ordered structure are some of the most controversial and difficult to treat theoretically and include hydrogren bonding, hydrophobic effects, and electrostatic interactions. How a protein finds its folded state in a reasonable amount of time is also another central question of protein folding. Clearly, not all conformations can be searched.[1] Attempts to design sequences that fold to particular structures, often called inverse folding, stand to inform these issues as well as lead to novel proteins. However, the design process opens up still another level of complexity. The number of possible protein sequences for a given target structure is exponentially dependent on the number of residues. There are numerous examples in nature of a small but diverse set of sequences that fold to similar structures. Discerning the properties and identities of those sequences that fold to a predetermined structure is one of the core goals of protein design.

10.1021/cr000058w CCC: $36.00
Published on Web 09/05/2001

The energy landscape view of protein folding is an attempt to unravel these various levels of complexity. This approach takes into account the energetics of a protein's conformational space in a reductionistic manner. Energy landscape ideas have led to advances in understanding protein stability, folding kinetics, energy function determination, and structure prediction.[2] In particular, energy landscape ideas have been honed via comparisons with simplified models of proteins where extensive characterization of a model protein's conformational space is possible.[3−8]

Here I discuss how energy landscape ideas and simple models can inform efforts in protein design. There have been many recent excellent reviews concerning the energy landscape theory,[2,9,10] simplified models of proteins,[11,12] and theoretical aspects of protein design.[13,14] Herein I will discuss some of the general issues involved in protein design and recent developments in the area over the past few years. It will be shown that the field continues to progress and in many cases makes direct contact with particular proteins and protein design experiments. I begin with a brief review of the energy landscape theory of protein folding in section II and then discuss the nature of some of the simplified models of proteins in section III. Two broad classes of methods are available to identify sequences that may be compatible with a particular folded state structure. One general approach involves searching for optimal sequences in a directed fashion so as to optimize sequence−structure compatibility, wherein sequences are explicitly generated and sampled. This and many of the important considerations in protein design are discussed in section IV. The second general approach involves indentifying the features of sequences sharing common properties in a statistical fashion. I review recent progress in this area in section V.

II. Energy Landscape Theory of Protein Folding

In the energy landscape picture, protein folding may be viewed as a collective, cooperative process.[9,15−18] The focus is on the global nature of a protein's free energy surface. Such a picture emphasizes those general characteristics that proteins in a structural class may share as well as which properties are specific to particular proteins. In such an approach, information about the energetics of the unfolded as well as folded states must be accounted for. This is an obvious consideration since the protein must recognize and acquire one conformation when a huge number are possible. Although much has been learned about the energetics and features of unfolded states from detailed, atom-based simulations,[19−21] these types of calculations are computationally time-consuming. Given the likely complexity of a protein's conformational free energy surface, the energy landscape picture focuses on developing concepts that simplify the description of such a complicated process as protein folding. Oftentimes studying simplified models of proteins, which can be extensively if not completely characterized, can be useful in this regard. A prime goal of energy landscape approaches is to identify a handful of thermodynamic and other simplifying quantities that characterize protein folding through their description of both the folded state and the partially folded ensemble of conformational states.

From a global perspective, the energy landscape theory provides a useful picture within which to discuss folding kinetics and thermodynamics. In this picture, the conformational energy surface of a protein is characterized in a statistical sense rather than painstakingly accounting for all intramolecular interactions. By "energies" here, what is meant is the free energy of a particular backbone conformation obtained after averaging over the solvent degrees of freedom. Such an "energy" could also be obtained from an effective potential that does not explicitly include solvent, such as one inferred from the protein structure database.[22] The discussion of energy and entropy then refer only to the polymer chain's degrees of freedom. One key perspective of the landscape approach is the recognition of the partially random nature of protein sequences.[15,23] Although there are exceptions,[24] by most tests protein sequences appear random. For arbitrary collapsed structures of the protein chain, incommensurate parts of the chain are likely to be in contact with one another, e.g. hydrophobic residues next to hydrophilic ones, leading to frustration. Frustration refers to the inability of all energetic interactions within the protein to be simultaneously satisfied. For most conformations, the covalent connectivity of the protein backbone prevents all the interactions between residues from being favorable. Because of this frustration, the energy surface of a random heteropolymer is rough. Small changes in conformation may lead to large changes in energy. In addition to a global minimum corresponding to the folded state, the surface has many additional local minima corresponding to partially misfolded states. When the thermal energy is much less than the typical contact energy, the protein can become trapped in these local minima. The roughness of the energy landscape leads to a glass transition at low temperature, wherein the protein can become trapped in these low-energy, non-native conformations.[2,25] As the temperature is lowered, the folding dynamics become increasingly sluggish and search through the minima of the landscape becomes difficult. To surmount this difficulty, quickly foldable proteins have an additional property that guides them through this multiple minima, conformational search problem: their sequences satisfy "the principle of minimal frustration"[15] or perhaps more precisely their sequences "sufficiently minimize frustration". This notion, which builds upon previous ideas in folding,[26] implies that the energy of the protein decreases more than would be expected for a random sequence as the conformations it assumes become progressively more similar to the native structure (the ground state). Energetic biases in the conformational energy landscape guide the protein toward the native state (see Figure 1). The biased, rough energy landscape picture has been applied to describe both experiments at a qualitative level and minimalist models of protein folding in a quantitative fashion.[2,12,27,28]

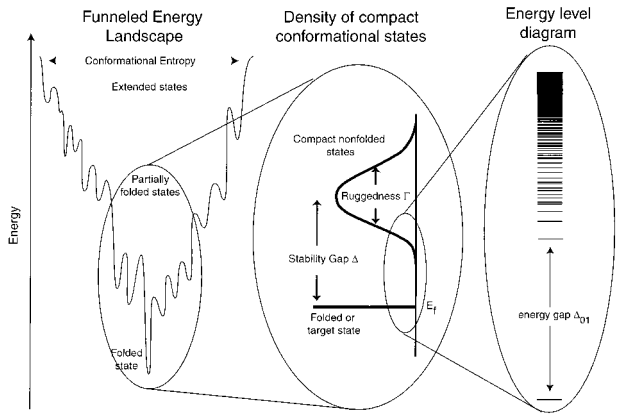


Figure 1. Schematic representations of the conformational energy landscape for a sequence that folds to a particular native state. (a) The "folding funnel" provides a quantitative rendering of the energy landscape. The surface is rugged with many local minima but has an overall bias toward the folded state. The width of the funnel indicates the conformational entropy.[8,179] (b) The distribution of conformational energy states for low-energy compact structures. The stability gap $\Delta = E_f - \langle E \rangle_u$, where E_f is the energy in the folded (target) structure and $\langle E \rangle_u$ is an average energy over compact conformations. The variation in energy, the ruggedness, is quantified by the variance of the energy over compact, unfolded structures $\Gamma^2 = \langle E^2 \rangle_u - \langle E \rangle_u^2$.[64] (c) The energy level diagram represents particular conformational states that are close in energy to the folded state. Note that the distribution in energy levels is essentially continuous above a certain threshold in the energy. Δ_{01} is the difference in energy between the folded state and the next highest distinct compact conformer.[61]

In the energy lanscape perspective, the assumption is that nature has selected for sufficiently nonfrustrated sequences over the course of evolution. Presumably, overly frustrated sequences do not fold and hence are not viable since they are likely to adversely affect the livelihood of an organism. Nature need not have found, however, optimal sequences. There are many examples where naturally occurring proteins can be made more stable in the laboratory through mutagenesis.[29,30] In the context of protein design, however, the degree of frustration is ostensibly at the control of the researcher since it is straightforward to synthesize any arbitrary sequence. Designing the energy landscape of a protein involves determining sequences whose energetic features are characterized for both the folded state and the ensemble of nonfolded structures. Including information about both stabilizing the folded (target) state and destabilizing unfolded (nontarget) states is often referred to as "negative design".[31] Indentifying practical quantities that characterize the conformational energy landscape for something as complex as a protein is nontrivial, but the energy landscape theory provides a framework for identifying such quantities, which will be discussed in section IV.D.

While energy landscape ideas have been used for structure prediction and for understanding folding kinetics,[2] the issue of protein design has some simplifying features relative to these other "folding problems". During folding (and structure prediction), amino acids are displaced so as to form favorable interactions in a collapsed structure. This must be done while maintaining the covalent connectivity of the backbone. There is a strong correlation between the locations of amino acid residues nearby in sequence. It is the frustration introduced by this covalent connectivity that motivates the use of partially random models developed for other frustrated model systems. On the other hand, in protein design, amino acids may be moved independently of one another so as to form a low-energy structure. The placement of amino acids is at the discretion of the researcher. This independence between amino acids simplifies the problem, but difficulties remain. In addition to addressing the energetics of a protein's conformational space, an exponentially large number of protein sequences is possible. For example, if all 20 amino acids are permitted at each position in an N-residue protein, 20^N sequences are possible. The difficulty comes in selecting suitable sequences from this large ensemble of possibilities. Computational algorithms are being developed to address this problem, while simplifying models can provide useful insight.

III. Atomistic and Minimal Models

A number of de novo designed proteins have been successfully created in the past few years, including proteins that mimic zinc fingers,[32] a novel right-handed coiled coil,[33] a helical dimer,[34] and a three-stranded β sheet.[35] Larger proteins (more than 50 residues) are more challenging targets. Nonetheless, several such proteins have been recently designed de novo, most commonly as helical bundles.[36−38] However, often such attempts yield proteins having substantial secondary structure but few well-defined tertiary interactions.[39] A variety of noncovalent interactions must be engineered. Such interactions include van der Waals forces, hydrogen bonds, complementary electrostatics, and hydrophobic interactions. The subtlety of these forces has frustrated many attempts to design particular structures. While structures with substantial symmetry such as coiled coils and helical bundles may now be successfully designed,[40] the complexity of proteins suggests that computational algorithms will be necessary to design structures as complex as those observed in nature, which need not possess any simplifying symmetry.

In addition to these experimental efforts, a number of important advances in algorithms and theories related to protein folding and design have been developed. These methods can be considered in the context of sequence−structure compatibility and can be loosely classified into two types: (a) those that consider the protein in nearly atomistic detail and (b) those that use reduced descriptions of the amino acids and backbone. Since the goal of these methods is to determine sequences that fold to a predetermined structure, the main chain of the protein is kept fixed or allowed to fluctuate only slightly. For a given target structure, nearly all sequence design algorithms vary sequence so as to optimize a given foldability criterion. Given the enormous number of possible sequences, methods of sequence design must search sequence space in a directed manner that does not consider every possible sequence or set of side-chain conformations.

Atomistic algorithms treat each protein sequence in molecular detail. These methods are discussed in several recent reviews.[40−43] The energy, as calculated using an atom-based molecular potential, is minimized by varying sequence and side-chain conformation.[44−47] In some algorithms, the free volume within the target structure is minimized.[48] In several cases, these methods have performed remarkably well in designing small proteins or those with limited sequence variability.[32,33,49,50] In dealing with the large number of possible sequences and side-chain conformations, sampling and pruning methods are often used. Sampling methods, which include genetic algorithms[47] or simulated annealing,[45,48] perform a partially random, directed search for sequences that minimize the energy or some other function that scores sequence−structure compatibility. Pruning methods such as "dead end elimination"[51−53] eliminate monomer types that cannot occur in the global optimum. For certain types of potential, pruning methods determine the global minimum. The computational demands of such techniques, which involve enormous numbers of degrees of freedom, limit the size and number of the sequences that may be considered in the search for those that optimize interresidue interactions.[54] Furthermore, the enormous number of degrees of freedom involved in these calculations impedes the consideration of alternate structures of the protein. Such atomistic methods cannot directly take into account the global features of a protein's energy surface. To explicitly include information about nontarget conformations is computationally prohibitive for such detailed descriptions of proteins. As a result, information about nontarget or unfolded states is included in an approximate manner, e.g., by including penalty terms for exposing hydrophobic surfaces.[55,56]

Alternatively, many researchers have examined the search for viable protein sequences using simplified models or simplified descriptions of real proteins. The goal with developing such models is to provide tractable representations of proteins that recover much of their important phenomenology. As such, these models have fewer degrees of freedom than more realistic representations of proteins. Much larger variation in sequence and in backbone conformation is possible. Side chains are replaced by effective atoms or are not represented at all. The complexity of the main chain peptide backbone is replaced with effective bonds between effective residues. In some cases, such models are simplified renderings of particular structures. A commonly used model type in this case are the so-called "off-lattice" models, where "off-lattice" simply implies that the coordinates that specify conformation are continuously valued.[6,57] These models are typically a chain of connected beads, where the beads represent the residue positions. A further simplification is to restrict the locations of the residues to the sites of a two- or three-dimensional lattice. This further reduction in conformational complexity permits much more extensive or in some cases complete sampling of conformations. Interestingly, realistic models of proteins are possible using high-dimensional or carefully chosen underlying lattices.[58−60] Smaller models such as the two-dimensional square lattice model and especially the three-dimensional 27-mer cubic lattice model are more commonly studied since it is possible to obtain extensive sampling of their energy landscapes.[11,13] Via Monte Carlo algorithms, it is also possible to examine the effective kinetics of folding in these systems.[61,62] Even these simplest explicit protein models have many of the stability properties of real proteins when only two amino acids are used (see Figure 2). Unique structures can be encoded in sequence that are stable at finite temperatures. The folding is often cooperative. Moreover, the kinetics of these lattice models resembles in many ways that of real proteins: there is a diversity of rates for different structures and sequences, sequences can be determined that fold in moderate amounts of time, and it is possible to identify features of the transition-state ensemble and other intermediate structures during folding.[12,13] In such models it is energetic interactions that are responsible for stabilizing a particular folded state structure. As a result a suitable potential or energy function must be chosen (see also section IV.B). For realistic representations of protein backbones without explicit side chains, the most commonly used simplified energy functions are the "information-based" potentials,[22,63−66] which are inferred from the database of known protein structures, but with model systems the effects of local stability can be examined by varying various energetic contributions. Last, the simplified protein models provide a well-understood "laboratory" within which algorithms for protein design can be tested.

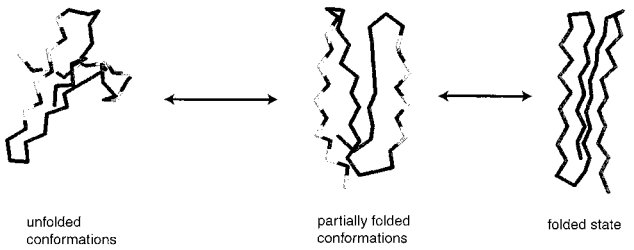


Figure 2. Minimalist models of proteins contain many of the features of real proteins: large numbers of unfolded conformations; partially ordered states, some of which can play a role in the transition state ensemble; and a stable, well-defined folded conformation. (The model shown is that of Thirumalai and co-workers.[6,119])

IV. Elements of Protein Design

Several considerations go into identifying heteropolymer sequences that are likely to fold to a predetermined structure. Obviously a target structure is necessary, and such a structure may be readily obtained from a structural database or as a result of molecular modeling, but some structures may be more difficult targets than others. Some indicator of "foldability" is required to identify sequences that rapidly fold to a unique, target structure. Since most foldability measures involve the energies of different conformations, accurate and tractable energy functions that reproduce the properties of proteins are also needed. Some means of identifying sequences having a desired foldability is necessary, but this

search is frustrated in part by the huge number of possible sequences. Reducing the number of possible amino acids can simpify the search in computational protein design, but deciding on a particular subset of the amino acids is a subtle issue. It is these issues of structure, energy function, foldability, sequence search, and amino acid alphabet that are the focus of this section.

A. Structure

It seems reasonable that backbone structures for protein design should look something like naturally occurring proteins. That is, they should be compact, have no steric conflicts or other "bad" interactions, and possess some degree of secondary (or more generally "local") order. The target structures of most experimental design efforts are analogues of naturally occurring proteins. Backbone target structures can be easily obtained from the protein database directly or from the molecular modeling of such structures. It may not be possible, however, to find sequences that fold to any arbitrary compact structure. Indeed, from the number of new types of structures that are being determined each year, it appears that the number of naturally occurring protein fold families may be finite.[67−69] That is, nature may have used a somewhat limited set of structures to construct a larger set of functional proteins. The TIM barrel is one example of a common protein motif that can fulfill multiple functionalities.[70] This potential conservation of structures, however, is not a limitation in the synthetic design of proteins. In addition to developing methods for designing naturally observed structures, one of the challenges for protein design will be to determine sequences for proteins whose backbone structures are radically different from those seen in nature.

For any potential target structure, it would be useful to have prior knowledge of the likely number of sequences, if any, that fold to that structure. Structures that support more sequences, so-called more "designable" structures,[71] should be easier targets for design. Finkelstein and co-workers addressed the issue of the number of proteins likely to fold to a particular structure using the simplest phenomenological model of a rugged energy landscape,[72] the random energy model (REM).[15,73,74] In such a model, explicit sequences are not considered and the overall energy distribution of conformational states is assumed to have a Gaussian form. Even within this simple model some structures support no sequences of a sufficiently low energy to fold to that structure. This group has also argued that commonly observed structures should be those that are easily stabilized by random mutations.[75] Though informative, such methods do not lead to specific predictions about the number of sequences likely to fold to particular three-dimensional structures.

With regard to general trends for specific kinds of structure, recently there has been much work on the connection between symmetry and the degree to which particular structures may support large numbers of sequences. Yue and Dill examined a simple HP model on 2D and 3D lattices,[76] wherein they searched for structures that only maximize the number of HH contacts. They found that the structural degeneracy, the number of structures with the same number of HH contacts, was lowest for highly symmetric structures, many of which the authors classified as the lattice equivalents of α-helical bundles and α/β-barrels. This suggests that symmetric structures should serve as promising design targets, since such structures have few structural neighbors energetically. For a specific choice of the contact energy function for the cubic 27-mer, a conformation recently highlighted by Li et al. is the lowest energy conformation for 3794 sequences.[71] In contrast, some conformations are the lowest energy state of only a few sequences or none at all for both 2D and 3D lattice models. Li et al. noted crude symmetries in their most designable structures and suggested that these symmetries may be the reason some sequences more "designable" than others.[71] From a more general perspective, it has been suggested that the symmetries observed in proteins arise from physical principles analogous to those that guide clusters of small molecules into symmetric shapes.[77,78] Nelson et al. studied an off-lattice (continuum) model of proteins, for which the folding kinetics and the susceptibility to mutations was examined.[79] Kinetic optimization, i.e., maximizing the folding rate, was found to be strongly correlated with ground-state symmetry, and high-symmetry ground states were the most robust with respect to mutation. Wang et al. found that highly designable structures also have a large degree of symmetry in a two-dimensional lattice model.[80] For a one-body (profile[81] or solvation[82]) energy function, the authors also point out that highly designable structures are also distant structurally from nearby structural neighbors. Kussell and Shakhnovich examined an analytically solvable model and applied it to a two-dimensional model of proteins.[83] Here the designability of particular structures is found to be sensitive to energy function. The results suggest that symmetrical structures are highly designable. All these studies suggest that target structures with a large degree of approximate symmetry should be the native states of a larger number of sequences and potentially easier to design than an arbitrary structure without such symmetry.

The issue of designability has also been approached from other perspectives. Goldstein and co-workers suggested that structures that have optimal values of a foldability criterion are also the native structures of a large number of sequences.[84] This notion is born out in their studies of lattice models of proteins. Buchler and Goldstein developed a structural measure based upon similiarity between structures.[85] The authors found that highly designable structures have a low density of nearby structurally similar conformations. Which particular structures are highly designable depends on the energy function used. Using HP-type lattice models, Ejtehadi et al. found that the designability of particular structures depends on the nature of the energy function.[86] They found a threshold with regard to the nonadditive contribution to the energy of the most stable contact. Below this threshold, designable structures are ro-

bust. Methods for estimating the number of sequences likely to fold to a given structure have been developed[87−89] and are discussed in section V.

B. Energy Function

Most foldability criteria are in some sense energy based. Energy functions are needed that quantitatively rank different conformations. These functions should account for many of the features of folded structures in an accurate but practical manner. Atomistic potentials have been developed for use with proteins and organic molecules.[43,90−93] Such potentials can accurately approximate both covalent and noncovalent interactions such as van der Waals forces, hydrogen bonding, and electrostatic interactions. These potentials are most useful when little is known about the structural tendencies of a molecule or when detailed modeling of molecular structures is desired. Such potentials have proved useful in simulations of folded proteins,[94] side-chain modeling,[95] structure refinement,[96] and protein design.[43] However, the use of such potentials in studying foldability and design is troubled for two reasons. Chain molecules have large numbers of possible conformations; in many cases, it is not computationally practical to use atomistic potentials to simulate unfolded conformations so as to obtain free energy differences. More importantly, the accuracy of such potentials for folding is uncertain. Atomistic potentials are most often derived from fits to small molecule data, and such potentials need not recover the folding of large molecules. Interestingly, however, in recent simulations of small proteins, the chains did acquire conformations that had substantial native structure.[97,98] While these atomistic potentials are useful for quantifying the packing of residues and side chains, their use in identifying folded state structures and folding sequences remains limited.

As a result, other groups have taken an alternative approach to developing potentials for proteins. These potentials involve reduced descriptions of the amino acids, wherein in the side-chain degrees of freedom are subsumed by a united residue approximation. Solvent is not treated atomistically but can be accounted for implicitly via pairwise hydrophobic interactions or as a dielectric that modulates electrostatic interactions and may be conformationally sensitive.[99] Although water can be structurally well determined in crystallographic structures, such detail is beyond the scope of simpified models. A hierarchical approach, where atomic detail is included only for nearly folded structures,[100,101] would be necessary to account for such specific solvent interactions. From a database of protein structures, empirical potentials can be developed that recover folding. The simplest such potential is the so-called Gō model of proteins, wherein for a given protein with a known folded state structure, only native interactions between residues are stabilizing.[3,26] Such a potential has been used to investigate a wide range of folding kinetics[25,26,102,103] but has limited applicability to design since it is specific to one sequence and structure. Another class of energy functions are statistical potentials, where the energy of the interactions are related to the frequencies with which residues appear in protein environments or with which particular contacts occur among folded state structures.[22,104−106] While there is some concern about the reliability of such simplified potentials,[107−111] these potentials allow for the evaluation of large changes in sequence and structure. Other groups have developed optimized potentials under the assumption that naturally occurring proteins maximize a stability criterion.[64,112−114] Similarly, one group recently proposed a method for generating potentials for design when the minimum energy sequence is known.[115] Under the optimization assumption, it is easy to arrive at the optimal energy function parameters using a suitable training set of proteins and simple matrix inversion.[64] These potentials have been successfully used to predict structure from sequence,[64,116] and it has been shown that energy functions of this type do an excellent job of reproducing folding in model systems.[117] Recent attempts at improving optimized energy functions involve selectively optimizing with respect to low-energy unfolded structures.[116,118]

An advantage of examining simple models of proteins is the fact that the form of the potential is entirely at the control of the researcher. Potential parameters may be varied so as to examine the influence of particular interactions on folding rate and stability. While bond angle- and torsion angle-dependent potential terms can be readily included in most any model, by far the most studied form of potential consists primarily of pair interactions.

$$E = \sum_{ij} u_{ij}(a_i, a_j, r_{ij}) \tag{1}$$

Here the sum is over unique pair interactions, a_i and a_j refer to the amino acids at residues i and j, and r_{ij} is the distance between residues i and j. Such a potential is often used since interactions between residues distant in sequence play an important role in stabilizing native structures. For off-lattice models of folding, u_{ij} is typically a continuous function of r_{ij},[6,103,119−123] e.g., a Lennard−Jones potential.[119,124] For lattice models, a common form of the potential is a step function.

$$u_{ij}(a_i, a_j, r_{ij}) = \begin{cases} \epsilon(a_i, a_j) & \text{if } r_{ij} \le r_o \\ 0 & \text{if } r_{ij} > r_o \text{ or } |i-j| = 1 \end{cases} \tag{2}$$

In other words, two residues interact with energy $\epsilon(a_i, a_j)$ only if their distance from one another is less than r_o and they are not covalently connected. Often for a lattice model r_o is chosen to be the lattice spacing so that two residues only interact if they are nearest neighbors on the lattice. Pair interactions are usually considered to be symmetric, $\epsilon(a, a') = \epsilon(a', a)$. Note that for such a contact potential, the total energy of a conformation may also be written as sum over contacts between different types of residue pair interactions.

$$E = \sum_{ij} u(a_i, a_j, r_{ij}) = \sum_{aa'} \epsilon(a, a') n(a, a') \tag{3}$$

Here the second sum is over unique pair iteractions between two types of monomer. If only two types of amino acid are possible, such as an HP model, then there are three distinct types of interaction: $\epsilon(P, P)$, $\epsilon(H, P)$, and $\epsilon(H, H)$. For 20 different amino acids, there are 210 unique pair interactions.

Several different types of potentials merit mentioning. In the simple HP potential of Dill and coworkers, only hydrophobic contacts are stabilizing: $\epsilon(P, P) = \epsilon(H, P) = 0$ and $\epsilon(H, H) < 0$.[4,76,125] Oftentimes the minimum energy conformations of such models are less than compact.[126] Other choices of energy function make other types of contact stabilizing so that on average compact structures are preferred, e.g., $\epsilon(P, P) < 0$ or $\epsilon(H, P) < 0$ or both.[8,62,127−129] Li et al. made the following choice for the energy function parameters:[71] $\epsilon(P, P) = 0$ or $\epsilon(H, P) < 0$ and $\epsilon(H, H) = 2.3\epsilon(H, P)$. For this choice of the energy function, compact conformations have lower energies than extended ones. In addition, it satisfies the inequality $\epsilon_{ij}(H, H) < \epsilon_{ij}(H, P) < \epsilon_{ij}(P, P)$ and the energy decreases for those sequences that have larger number of contacts involving H monomers. The energy function favors H monomers in the interior of a structure, since interior positions have the largest coordination numbers. Thus, the energy function favors placing H monomers on the interior of a conformation, which conforms with the burial of hydrophobic residues seen in folded protein structures. This energy function also satisfies $\epsilon_{ij}(H, H) < 2\epsilon_{ij}(P, P)$, which implies that dissimilar monomers favor segregation within the collapsed globule. Li et al. mention that their results were not sensitive to the precise value of $\epsilon_{ij}(H, H)$, as long as the above inequalities are satisfied.

For larger numbers of amino acids, more complex potentials are necessary. Given that there are 210 possible interactions for 20 amino acids, Shakhnovich and co-workers examined potentials where the energy parameters u_{ij} are random quenched variables.[5,61,130] The Shakhnovich group also pointed out that there are some cases where design schemes that fail for a HP model but are valid when larger number of amino acids are available.[131] The most commonly used contact potentials derived from real proteins are the potential parameters derived statistically by Miyazawa and Jernigan.[22,106] Other groups have also developed contact type potentials inferred by optimizing foldability criteria across a training set of protein structures.[64,116]

It is also important to note that pair potentials such as that in eq 1 may not be sufficient for discriminating native structures in realistic representations of proteins.[132−135] Energy functions involving three-body and higher order terms may be invoked, but determining these potentials requires very large data sets. Nonetheless, even for simple model systems, pair potentials such as that in eq 1 can yield a rich array of protein-like thermodynamic and kinetic properties.

C. Search Methods

In the course of design, sequences must be selected that are compatible with a given target structure. The space of all possible sequences is exponentially large; it scales as m^N, where m is the number of monomers (20 if all amino acids are used) and N is the number of residues in the protein. The fraction of sequences likely to fold to a particular structure is an infinitesimal fraction of this number. Thus, searching for suitable sequences is nontrivial. For small, specifically defined models, it is sometimes possible to obtain analytic solutions for the optimal sequence and other sequence properties.[83] One advantage of very small models, such as two- and three-dimensional lattice models, is that in many cases it is possible to explicitly generate all possible sequences in the search for those that optimize a particular foldability function, particulary if the number of amino acids is small, e.g., $m = 2$. For large models having arbitrary numbers of amino acid types, however, it is necessary to use numerical sampling algorithms that search for viable sequences.

Given that the sequence energy landscape may have local minima as sequences are sampled, partially stochastic methods are often used that can avoid becoming locally trapped. The appropriate algorithm to use often depends on the particular model, the energy function, the way trial sequences are generated, and the type of information about the sequence space of a particular protein that is desired. Monte Carlo methods are some of the most mature and most commonly employed methods for sequence searching.[45,136] One useful aspect of Monte Carlo methods is the ability to associate an effective temperature with the search process; this defines the degree to which sequences having high energy (or other values of low foldability) are accessible during the course of the search.[137,138] The ensemble properties of suboptimal sequences can be investigated. Genetic algorithms have also seen extensive use in sequence design.[47,139] In addtion to yielding specific sequences, both Monte Carlo and genetic algorithm methods provide information about an ensemble of sampled sequences. With slow cooling of the effective temperature and an arbitrary amount of computer time, Monte Carlo methods can provide good estimates of global minima (simulated annealing).[45,48] Nonetheless, limited computer resources and the fact that the sequence energy surface may be rugged can motivate the use of other methods.

Self-consistent mean-field methods are useful for treating systems having many degrees of freedom. Theories of this type have been used extensively in condensed matter physics[140,141] and in biomolecular science for structural studies.[142−144] In particular, mean-field methods have been successfully applied for side-chain modeling problems with given backbone structures.[145−150] The effective thermally averaged energy (field) at each site is solved for self-consistently as the overall energy is lowered. The method is fast and usually scales as a polynomial in the number of residues N. A quantitative comparison of sequence search algorithms confirms the efficiency of mean-field theory, although for solvable test cases, the method sometimes fails to find the sequence of minimum energy.[54,151] In addition to identifying low-energy sequences (or sequence−rotamer combinations), mean-field methods may also be used to

estimate the probabilities of the amino acids that are energetically consistent with a particular backbone structure[87−89] (see section V).

A number of recent advances in search methods are likely to aid sequence searching. Recently, configurational biasing has been introduced into Monte Carlo sampling of sequences.[151] Such methods can permit faster cooling rates and speed toward low-energy sequences. Cootes et al. found that such methods speeded the Monte Carlo search, provided more accurate estimates of global optima for known test cases, but were not as fast as mean-field methods. Foreman et al. developed a new global optimization method which works well for rugged energy surfaces with an overall funnel topology.[152] Such methods may eventually prove useful for protein design.

Pruning methods such as "dead end elimination"[51−53] eliminate monomer types that cannot occur in the global optimum and can provide a more extensive search among sequences. For potentials comprising only site and pair interaction terms, pruning methods determine the global minimum. Pairs of interactions between residues are used to eliminate amino acids (or amino acid conformations) that cannot contribute to the global minimum. The process is continued until no further amino acids can be eliminated. If more than one sequence remains, these few are searched for the global optimum. The method has been applied with substantial success to the design of small proteins[32] and residue subsets within proteins.[153,154] The method is promising and is undergoing further refinement.[155,156] The method is not without limitations, however. It is restricted to scoring (energy) functions that involve only one- and two-body interactions. There is increasing evidence that the use of three-body and higher order terms may be necessary to recover folding to unique structures.[134,157] With the pruning methods it can be difficult to include global constraints on sequence such as the numbers of each amino acid, and calculating the energies of alternate compact, nontarget conformations is computationally prohibitive. Since amino acids at many sites are eliminated, it is difficult to use the method to examine the properties of ensembles of sequences, some of which may have less than optimal interactions.

While advances in search methods will surely be useful, many of these methods are sufficient to determine viable sequences, particularly for model systems of proteins. The choice of a particular search method is not likely to be a problem. Sequence design methods are more sensitive to the choice of potential, target structure, and foldability criterion.

Most protein design algorithms place no requirements on the properties of transiently generated sequences during the course of the search for viable sequences. It it is interesting to note, however, that such requirements would in some sense mimic evolution.[131] In addition it is useful to characterize the sequence landscape, i.e., how the folded state energy and energy landscape features change with sequence. In studying a small off-lattice model, Nelson and Onuchic found that different structures have different energetic "basins" in sequence space containing the sequences that fold to those structures.[158] They found that to traverse the landscape from one of structure to another, i.e., to move between basins, it was necessary to pass through a "barrier" region in sequence space comprising sequences without folded states. Tiana et al. examined the sequence energy landscape of a 36-mer lattice model with a 20-letter alphabet.[159] They perform Monte Carlo protein design simulations at different effective selection temperatures. For a target structure, sequence space is grouped into clusters of low-energy sequences having large similarity that are mutually accessible via essentially neutral pair mutations. At higher design temperatures, the authors interpret their findings in terms of superclusters; different superclusters have little sequence similarity and require high-energy (nonfoldable) mutations to convert from one to another.

D. Foldability Criterion

How might we choose to quantify the foldability of a particular sequence? The simplest measures of sequence−structure compatibility are qualitative and use information learned from known structures. Sequences have been designed using primarily hydrophobic patterning, steric complementarity, and secondary structure preferences, and these methods have often been successful.[39] For a quantitative approach, it is useful to have global criteria that consider the protein and its sequence as a whole, since folding is a cooperative process.[17,160−162] Also, proteins need not have all their intramolecular interactions satisfied since many examples exist in nature of proteins that can be stabilized through mutagenesis.[163] Thus, the stability of proteins need not be optimal, although there is some evidence that on average they may be close to being so.[164] Quantitative measures of sequence−structure compatibility are termed "foldability criteria". What is meant by a "good folder?" Generally, this refers to a sequence that folds reversibly to a unique structure in a reasonable amount of time.

Given the importantance of folding kinetics, the folding rate has been used to select for folding sequences.[121,129,131,165,166] The results of these studies suggest that nature may indeed select for sequences with fast folding rates, especially in light of comparisons with conserved residues in naturally occurring proteins.[166] Performing such simulations is extremely computationally intensive, however, since for each sequence visited the estimated folding rate of that sequence must be determined via repeated kinetic simulations. The results of such kinetics studies on multiple sequences suggest that there is a large correlation of folding rate with folded state stability. In particular, the folding rate is correlated with Δ/Γ (see Figure 1).[129,130,167] While there as a been a large body of work finding that folding rate is strongly correlated with stability,[10] the connection is subtle. Studying a cubic lattice model having simple "side chains", Li et al. identified interactions that do not appear in the native state but whose stabilization speeds folding.[166] Design of such a protein then could

include stabilizing nontarget structures (or substructures) in the unfolded state. A detailed understanding of the kinetic events in folding would provide an additional element of control in the design of such proteins.

Given the computational difficulty of directly assessing the kinetic viability of each trial sequence, much effort has been put into arriving at design algorithms from a thermodynamic viewpoint. For reversible folding, it is generally regarded that the folded state must be a pronounced free energy minimum.[168] At a desired temperature, the folded state stability (foldability) is best determined by the population of the folded state P_f or, equivalently, the free energy of folding ΔF.

$$P_f = \frac{\exp(-\beta E_f)}{Q} \tag{4}$$

Here E_f is the energy of the folded state and $\beta = 1/(k_B T)$, where k_B is Boltzmann's constant. The partition function Q involves a sum over all possible conformational states including the folded state. Here it has been taken that the folded state is essentially unique; there is just one folded state structure so that $Q_f = \exp(-\beta E_f)$, where Q_f is the partition sum over the "folded" part of conformational space. Note that P_f may also be written as

$$P_f = \frac{1}{1 + K_{eq}^{-1}} \tag{5}$$

where $K_{eq} = \exp(-\beta \Delta F) = Q_f/Q_u$ is the equilibrium constant for folding and Q_u is the partition sum for unfolded conformational states. Thus, mimizing ΔF (maximizing K_{eq}) with respect to sequence should yield proteins that are stable in a particular target structure.

$$\beta \Delta F = \beta E_f - \ln Q_u \tag{6}$$

As with many problems in statistical thermodynamics, it is determining (or approximating) Q or Q_u that is the main hurdle in developing practical, thermodynamically based foldability criteria. (Note that most design procedures are performed for a fixed temperature, β = constant.)

For small models of proteins, the free energy difference ΔF can be estimated and used to arrive at optimal sequences.[169] Seno et al. employed a dual Monte Carlo method to determine sequences having a high probability of residing in a target structure.[169] These authors sampled sequence space, and for each attempted sequence the free energy of unfolded state structures was estimated using a Monte Carlo growth procedure in conformational space, so that low-energy structures are preferentially sampled. Sequence design for two-dimensional small lattice models was performed. This method was extended to previously studied 48-mers, a set of design targets studied previously.[126] The unfolded partition sum was approximated as an average over conformations, $Q_u \propto \langle \exp(\beta E) \rangle^{-1}$.[170] With no constraints on the numbers of each type of monomer, sequences having a range of compositions, where the number of hydrophobic residues ranged from 10 to 24, were identified. Though powerful and accurate, these methods are too computationally intensive to apply to realistic models of proteins.

Several foldability criteria have been suggested that do not involve calculating free energy differences. From considerations of the features necessary in a protein's energy landscape, the notion arises that fast folding proteins should have large values of T_f/T_g, where T_g is a glass transition temperature below which the protein can become trapped in any one of many low-energy conformations.[171] T_f is the folding temperature; for $T < T_f$, the folded state is thermodynamically stable. Socci and Onuchic found that stable, rapid folding does correlate with T_f/T_g in 27-mer lattice models.[62,172] Also from studies of folding in model proteins, a strong correlation has been found between the folding rate and $(T_\theta - T_f)/T_f$, where T_θ and T_f are the collapse and folding temperatures.[12,128] It is not clear, however, how to determine sequences having a predetermined value of either of these ratios without doing extensive simulations. In particular, simulations are needed to determine T_θ, T_g, and T_f. For a simple model such as the random energy model, it is possible to obtain an analytic expression for T_f and T_g.[173] While these temperatures and their ratios can provide a fundamental characterization of the energy landscape of a protein, their use in protein design is limited.

In eq 6 it is shown that minimizing the energy of the folded state E_f should have an impact on stability, if the sequence dependence of the unfolded state ensemble is neglected, i.e., $\ln Q_u$ is not dependent on sequence. In such a context, it is possible to develop a statistical mechanical theory for protein design. A Monte Carlo-based search for sequences of low energy E_f corresponds to sampling sequences at an effective temperature, which can be understood as a design T_{des} or selection temperature T_{sel}.[137,138] Sequence design becomes much simpler if only the energetics within the folded state are of concern. This should be a valid assumption if all sequences considered have essentially the same behavior for $\ln Q_u$. This would be expected if the search process sequence is varied subject to "constant composition", i.e., the search considers only sequences having the same numbers of each amino acid.[136] This has been reviewed more extensively elsewhere.[13,14] For a sufficiently flexible protein where compact states dominate the contribution to $\ln Q_u$, this should be a good assumption since the unfolded ensemble involves an average over many structures having large numbers of fluctuating interresidue interactions. This has been verified by studies of lattice model proteins, where the free energy of nonfolded states is relatively invariant for fixed numbers of each monomer type.[170] Nonetheless, such an algorithm may fail for some models where the folded state is less than compact.[126] A related approach is molecular "imprinting", where separated monomers are annealed around a central substrate. The amino acids are then polymerized.[138] Configurations with stable interactions are selected, and in cubic lattice models, the sequences so selected often reversibly fold. The method is less robust,

however, because low-energy interactions between nearest neighbors along the chain in the disconnected state are not a factor in the polymer since covalent connectivity maintains their proximity to one another for all protein conformations. Guided by results from the random energy model of proteins, Shakhnovich and co-workers argued that minimizing the folded state energy at constant composition actually optimizes a more profound foldability criterion, the energy gap Δ_{01} (see Figure 1).[61] This is the difference in energy between the folded state and the next lowest energy compact conformation. Clearly, for a sequence to fold to a unique structure, the energy of the folded state must be removed for of competing structures. This method of minimizing energy at constant composition has been used extensively to design and study lattice models of proteins.[28,61,136,174] Studying a 36-mer lattice model, Tiana et al. found a hierarchy of allowed mutations, where the mutations are ranked according to the change in the energy of the folded state structure.[175] The impact on folded state stability was found to be directly related to the energy of the mutation, which affected primarily the energy of the native state rather than the energetics of the unfolded ensemble.

A recent lattice study has examined a design proceedure wherein energy is also minimized but in an attempt to examine the role of steric effects. Micheletti et al. developed a model wherein some "large" (L) monomers where assigned energetic penalties when at interior positions not having vacant nearest neighbor sites.[176] "Small" (S) monomers where less penalized when at sites of high coordination. Such a simple LS model exhibited many of the same features of proteins, including compact structures being the lowest energy, only a tiny fraction with a unique ground state, and encodable structures that are highly designable. The model was applied to two-dimensional lattice polymers. Though highly simplified, the model suggests that steric considerations can play a pronounced role in dictating sequence. For the 16-mer, nondegenerate ground-state sequences may be identified using only the target structure alone, whereas for an HP model, information about nonfolded states must be included to identify encodable structures and viable sequences. For such an LS model, it is not possible to mount alternate structures without violating steric constraints. This finding is in harmony with the success of some atomistic design algorithms that seek to appropriately pack protein cores with side chains.[46,47,52] Interestingly, Micheletti et al. found that when combined with an HP model to yield four types of monomer, the resulting diversity allows new, less-compact structures to be encoded.

The constraint of constant composition limits the sequences that can be used in the course of protein design. Explicit incorporation of more information about unfolded states in the design process permits relaxing constraints on composition. Alternate design algorithms take into account the energetics of structures other than the target. Truncated cumulant expansions provide a useful framework within which to discuss such criteria.[177,178]

The unfolded partition sum in eq 6 will be estimated, even in cases where the set of unfolded conformations is incomplete. In fact, only for simple model systems may all possible unfolded conformations be enumerated and ln Q_u determined exactly. More often a subset of nontarget conformations must be considered. Many have argued that a suitable subset is that which most closely resembles the set of compact alternative structures of the protein. Proteins consist of a large number of hydrophobic residues, so collapsed conformations are those most likely to compete with the folded state and to have the dominant contribution to ln Q_u.

The unfolded ensemble partition function can be written as an average over Ω_u unfolded conformations.

$$\sum_{i}^{\Omega_u} e^{-\beta E_i} = \Omega_u \left(\frac{1}{\Omega_u} \sum_{i}^{\Omega_u} e^{-\beta E_i} \right) = \Omega_u \langle e^{-\beta E_i} \rangle_u \qquad (7)$$

Cumulant expansions may be used to approximate the average.

$$\ln \langle e^{-\beta E_i} \rangle_u = -\beta \langle E \rangle_u + \frac{1}{2}\beta^2 (\langle E \rangle_u^2 - \langle E^2 \rangle_u) + \ldots \quad (8)$$

$$= -\beta \langle E \rangle_u + \frac{1}{2}\beta^2 \Gamma^2 + \ldots \qquad (9)$$

Here $\Gamma^2 = \langle E \rangle_u^2 - \langle E^2 \rangle_u$ is the variance in the energy among unfolded compact states (see Figure 1). Note that this is essentially an expansion in β and is likely to be most valid at high temperatures where β is small. ΔF can then be written as

$$\beta \Delta F = -\ln \Omega_u + \beta \Delta + \frac{1}{2}\beta^2 \Gamma^2 + \ldots \qquad (10)$$

where $\Delta = E_f - \langle E \rangle_u$. Usually in the process of protein design, sequences having the same length and similar physical chemical properties are compared. As a result, the term ln Ω_u is identical for all potential sequences and need not be considered in determining those compatible with a particular target structure. A foldability criterion can then be defined, Ψ.

$$\Psi = \beta \Delta + \frac{1}{2}\beta^2 \Gamma^2 + \ldots \qquad (11)$$

If just the first-order term in β in eq 11 is kept, then at constant temperature the foldability criterion is $\Psi \approx \beta \Delta$, where Δ is the so-called "stability gap" (see Figure 1).[179] This energy difference Δ between the target state and an ensemble of nonfolded states has been suggested as a potential foldability criterion.[177] An appealing feature of this and more advanced folding criteria is that the assumption of constant composition can be relaxed, and hence, a much wider range of sequences can be considered in the design process. Using a 3D 27-mer model, Zou and Saven showed that the number of sequences is dramatically increased if Δ is used to classify sequences and that in the absence of constrained composition there is little correlation between E_f and Δ. Rossi et al. go beyond a simple pair potential and

include three-body interactions in a design approach using three types of amino acids[180] to describe an off-lattice model of proteins. The stability gap $\Delta = E_f - \langle E \rangle_u$ is minimized in designing sequences and is also evaluated for several different effective selection temperatures. For the structures of several proteins (thioredoxin, CI2, and barnase), correlation between conserved residues among naturally occurring sequences and those residues whose identity is specified in the calculation even at high selection temperatures is observed.

A shortcoming of using Δ as a foldability criterion is that it characterizes the unfolded part of the energy landscape in a very crude way by only accounting for the mean $\langle E \rangle_u$. It would be useful to have more information about the distribution of unfolded state energies for each sequence. A simple measure of the width of the energetic distribution of unfolded states is the variance Γ^2. The cumulant expansion in eq 11 truncated at second order is a natural way to include this information. Morrissey and Shakhnovich used similar cumulant expansions in the search for sequences that are stable at a desired temperature.[178]

Another foldability criterion that has received considerable attention is Δ/Γ.[64,181,182] For the REM model of the protein energy landscape, it may be shown that minimizing Δ/Γ is a direct result of maximizing T_f/T_g.[171] It is straightforward to show that minimizing $\beta\Delta + 1/2\beta^2\Gamma^2$ is equivalent to minimizing Δ/Γ in the context of a contact energy function (see eq 3). Since the energy can be written as a sum over distinct types of contact interaction $E = \sum_{aa'} u_{a,a'} n_{aa'}$, then

$$\Psi_2(0) = \sum_{aa'} (n^f_{aa'} - \langle n_{aa'} \rangle_u) u_{aa'} + \frac{1}{2} \sum_{aa',bb'} u_{aa'} \langle (n_{aa'} n_{bb'} \rangle_u - \langle n_{aa'} \rangle_u \langle n_{bb'} \rangle_u) u_{bb'} \quad (12)$$

$$= \mathbf{A} \cdot \beta \mathbf{u} + \frac{1}{2} \beta \mathbf{u} \cdot \hat{\mathbf{B}} \cdot \beta \mathbf{u} \quad (13)$$

where the vector $\mathbf{A}$ has elements given $A_{aa'} = n^f_{aa'} - \langle n_{aa'} \rangle_u$ and the matrix $\hat{\mathbf{B}}$ has elements $B_{aa',bb'} = \langle n_{aa'} n_{bb'} \rangle_u - \langle n_{aa'} \rangle_u \langle n_{bb'} \rangle_u$. Recall that typically there is a discrete set of amino acid types, but if the elements of the vector $\mathbf{u}$ are treated as being continuously valued, then it is easy to determine the minimum. The value of this set that minimizes $\Psi_2(0)$ is

$$\beta \mathbf{u}_{opt} = -\hat{\mathbf{B}}^{-1} \cdot \mathbf{A} \quad (14)$$

This is exactly the solution obtained for minimizing Δ/Γ for a particular structure. In fact, these ideas are essentially identical to one method for determining energy functions from a training set of proteins.[64] Energy function determination and protein design are in some sense parallel problems. In developing effective energy functions, typically a fixed set of sequences is used to solve for effective energy parameters, where as in protein design it is the energy function that is fixed so as to identify a suitable sequence. There is usually a finite set of possible amino acids, however, so that protein design is a discrete problem, whereas the parameters of an energy function may be continuously varied. From an analysis of the random energy model, Buchler and Goldstein showed a significant positive correlation of the energy gap Δ_{01} with Δ/Γ.[85] These results are also in good agreement with lattice studies. Abkevich et al. used Δ/Γ as the basis for a sequence design algorithm in a Monte Carlo search for sequences having low values of this ratio.[181] Optimal sequences have values of Δ/Γ that are large in magnitude and less than zero. Abkevich et al. go on to make the more specific statement that fast folding sequences have an additional feature that reduces their ruggedness; these sequences have a low energetic dispersion in their individual native contact energies. Three separate lattice model studies have found that Δ/Γ correlates well with the folding rate and stability;[129,130,167] however, the most rapid folders need not have the lowest value of Δ/Γ.[166]

Δ/Γ is often referred to as the "Z-score", borrowing notation from statistics where "z" is sometimes used to denote the separation from the mean of a particular sampled value as measured in units of the standard deviation. However, the term "Z-score" has also been used to denote a similar quantity, where the averaging is instead done over different sequences for a fixed structure[81] rather than different conformations for fixed sequence. In fact, Street et al. used this latter "sequence scrambling" definition of the Z-score to design mutants at the surface of a β-sheet protein with an all atom model.[154]

Δ/Γ results from a simple picture of the protein energy landscape. As mentioned, it may also be viewed as the result of a high-temperature expansion. More sophisticated criteria are likely to be needed to further characterize the relevant parts of the energy landscape, particularly the low-energy unfolded states that are populated at low temperatures. As a result, there have been a number of recent efforts to determine new algorithms for protein design.

In an attempt to further surmount the difficulty with estimating the free energy of unfolded states, Seno et al. developed a variational approach to protein design.[183] The unfolded free energy is expressed simply as a linear function of the numbers of each type of monomer, n_i: $F_u = \sum a_i n_i$. Using a set of predetermined folding structures, the coefficients a_i are determined by minimizing an intensive functional that is a function of $E_f - F_u$. For a four-letter model, this functional is minimized across a set of 500 sequences of two-dimensional square lattice 16-mers. The authors were able to identify foldable sequences after an appropriate energy function had been determined by variationally minimizing the same function. Simplification of representing the free energy of unfolded states greatly speeds the design process.

Micheletti et al. compared several different design methods using an HP model.[184] The algorithms they consider include the following: an algorithm similar to that of Sun et al.,[185] wherein assignment of H or P depends on the number of local neighboring resi-

dues; minimizing an approximation to the free energy difference between folded and unfolded states,[176] wherein an effective chemical potential for H-type residues was used to approximate F_u; and last, a third method where F_u was replaced by a cost term that maintains the approximately linear relationship between chain length and number of contiguous H-segments. Sequences and structures were obtained from the protein structure database (PDB) and coarse grained as either H or P. Success was measured as the percentage of times an H or P was identified as that observed in the "true sequence". For each of the methods the success rate is only in the range of 70−75%. The authors suggest that this is due to the HP coarse graning and confirm this with studies of reduction of a four-letter to two-letter alphabet in a square lattice model. For such an exactly solvable system, only 72% of sequences maintain unique ground states upon alphabet simplification and only 86% even if the energy parameters are re-optimized after simplification.

Rossi et al. developed an iterative method to determining folding sequences and tested it using a the 27-mer lattice model of proteins.[186] In this method, the free energy of unfolded states is approximated by summing only over a set of compact conformations. An optimal sequence is selected, and its ground state is determined. If this conformation is not the target, the structure is added to the set of nonfolded conformations and the process is repeated. The method imposes no composition constraint, and the designability of particular structures was found not to be dependent on the number of possible amino acids. Although this method shows striking success for lattice models, one drawback is that it requires conformational minimization of intermediate sequences, which is computationally intensive for realistic models of proteins. Irbäck et al. surmounted the usual nested Monte Carlo approach, wherein configurational Monte Carlo simulations are performed for each attempted sequence.[187,188] In what they term "multisequence" Monte Carlo, sequence and conformational space are placed on equal footing and sampled simultaneously. Sequences that are observed more frequently in such a method are less likely to have unique ground states and are pruned from consideration in the course of design. The authors compare this type of elimination of sequences with elimination based on finding sequences with ground-state energies other than the target structure. The method is applied to HP representations of lattices and to off-lattice models. The method is efficient but limited to models having small size and limited alphabets, since a set of sequences must be maintained. The sampling scheme can be improved by eliminating conserved sites, as obtained from trial runs. The authors verify that the method is an improvement upon design methods based upon minimizing energy at constant composition or high-temperature expansions.

Mirny et al. identified that selection for stable as well as fast folding sequences can occur by maintaining strong energetic interactions within a critical substructure of a particular folded state.[189] Many of the remaining residues of the 48-mer lattice model were less conserved, suggesting that in protein design it may be possible for seemingly complex structures to focus on optimizing the interactions among a handful of key residues.

While most efforts involving design have focused on specifying the global structure, some have considered just specifying the structure of a small subset of amino acids. The global fold of the sequence is unspecified, but the goal remains to find sequences that fold to well-defined conformations that maintain a desired structure of the subset. This mimics the selection of sequences based upon function, e.g., maintaining local structure at an enzymatic active site. Pande et al. presented similar ideas in their lattice-based imprinting studies.[190] Building upon results from a spin-glass model of proteins,[191] Yomo et al. selected sequences according to function, which was specified by selecting for sequences with a predetermined spatial arrangement of four residues.[165] The structure of the remaining residues in their 47-mer off-lattice model of the engrailed homeodomain were not predetermined. Mutations were accepted according to the usual importance sampling based upon the structure of the "active site" only, wherein the degree of convergence to the active site configuration was determined using a molecular dynamics simulation for each trial sequence. In addition to maintaining the desired "active site" structure, the sequences "evolved" to possess substantial helical content and compactness similar that of the homodomain. This remaining structure provided a scaffold upon which to present the "active site". This is an example of a case where design can be successful without complete specification of the folded state structure.

E. Alphabet or Monomer Set

Nature uses the 20 naturally amino acids to build protein structures. Synthetically, it is straightforward to use this same "alphabet" to create any desired sequence or even to use a larger number of artificial amino acids.[192,193] A subset of the 20 amino acids may also be considered in examining the effects of alphabet simplification on protein design. Reducing the number of amino acids reduces the number of possible sequences. From a computational standpoint, there is an additional subtlety concerning how amino acids are distinguished. This can include gross simplifications of the 20 amino acids by classifying each into one of several groups based upon molecular properties such as hydrophobicity or side-chain size. The effective number of different types of monomers may also be increased by associating conformational states with each amino acid and determining not only the amino acid identity but the side-chain orientation of each amino acid for a given backbone structure. This is a combinatorial problem encountered in explicit side-chain-based algorithms for protein design.[52] Nonetheless, there is much interest in simplifying the amino acid alphabet both from a practical viewpoint of reducing the size of the search space as well as from of a more fundamental standpoint of understanding the minimal set of

amino acids necessary to form particular structures or any structure.

Experimental studies with reduced numbers of amino acids have been recently reviewed,[194,195] and only some of them will be mentioned briefly here. Early efforts in protein design examined positioning hydrophobic and polar residues at positions in the heptad repeat of an α-helix to form helical bundles, primarily focusing on the positioning of leucine residues in the interior and glutamate and lysine residues at the exterior and interfacial positions.[39,196,197] Kamtekar et al. suggest a binary patterning of amino acids according to hydrophobicity, much akin to the minimal protein models having just two types of monomer.[11] Using such a patterning for an α-helical bundle, where hydrophobic residues reside on the interior and hydrophilic ones on the exterior, these authors identified a large fraction of sequences having many protein-like properties. This was not, however, a purely "two-letter" monomer alphabet, since six different hydrophobic amino acids and five different hydrophilic amino acids were possible. Riddle et al. used combinatorial methods to reduce the number of possible amino acids at 40 positions in the src SH3 domain.[198] Only five amino acids representing a range of physicochemical properties were allowed at these positions: I, K, E, A, and G. Functional proteins were identified in which 38 of these 40 positions had been mutated. From these studies the authors infer that at least five, but not three, types of amino acids are necessary to encode sequences that fold to the SH3 structure. Schafmeister et al. used just seven different amino acids to design a four-helix bundle.[36] In this study, a reduced amino acid set was used to create an entire protein.

Accompanying these experimental efforts are several studies that address the number of amino acids using theory and modeling. Studies of 27-mer lattice and off-lattice models of proteins suggest that the rapid folding capabilities of real proteins are better recovered with models having three rather than two amino acids.[18] Shakhnovich found it difficult to find low-energy structures in protein design using an HP model.[131] However, folding sequences can be found when 20 amino acids are used for such models with a Miyzawa−Jernigan-type contact potential.[22] Wolynes suggested that as the number of amino acids is increased, it becomes more facile to arrive at funneled energy landscapes.[199] Homopolymers have essentially a flat energy landscape, whereas proteins comprising only a few amino acids may have many competing minima. With increasing monomer complexity, it becomes possible to select (via evolution) or tailor (via design) sequences with good folding properties. By noting correlations among the elements of the BLOSUM50 similarity matrix used for sequence alignments and developing appropriately reduced matrixes, Murphy et al. found that approximately 10 different amino acid types are necessary to detect homologues in a clustered database.[200] In this study, more than 600 folding families of structures were considered. The authors suggest that the value of approximately 10 different amino acids applies to designing arbitrary structures but that a smaller alphabet may be sufficient for particular structures. Starting from a Miyazawa−Jernigan pair interaction matrix,[106] Wang and Wang reduced the 20-monomer alphabet by minimizing mismatches between pairs.[201] They were able to reduce this to a five-letter alphabet, consistent with the findings of Riddle et al.[198] Moreover, Wang and Wang use their reduction scheme in combination with folding studies of a 27-mer to examine how the average similarity (fraction of native contacts, Q) varies with number of amino acid types. This average similarity is small when only two amino acids are used ($\langle Q \rangle \approx 0.35$) but reaches $\langle Q \rangle \approx 0.9$ for five different amino acid types. There is only slight further improvement when 20 different amino acids are used. By studying an off-lattice model coupled with design via minimizing the energy at constant composition, Liang found that four rather than two amino acid types are necessary to design sequences that fold to collapsed structures of a 16-mer.[202] Studying a two-dimensional lattice model, Buchler and Goldstein found that the designability of a structure is highly dependent upon the number of amino acids that are used as well as on the criterion used to determine foldability.[203] Highly designable structures for HP models were not the most designable for larger alphabets. In determining designability, these authors did find a strong correlation between 20-letter Miyazawa−Jernigan-type alphabets and an a model that permitted an arbitrarily large number of amino acids.

Taken together, these experimental and theoretical studies suggest that some degree of simplification is available in determining sequences that fold to a desired structure. However, some diversity must be maintained in order to identify sequences with sufficiently smooth and biased energy landscapes. The results suggest that different structures may require different numbers of amino acids. As few as five amino acids may suffice to construct structures such as the SH3 domain,[198] but more may be required for larger or more structurally complex structures. The ability to quantitatively understand such alphabet simplification can only stand to accelerate progress in protein design. In addition, expanding the alphabet through the use of unnatural amino acids will expand the tool kit for protein designers and enlarge the number of "designable" structures (see also the article by Cheng, Gellman, and DeGrado in this issue[204]).

V. Statistical Approaches to Design

Despite some of the successes of the discussed computational design methods, a variety of issues hinder their use to probe the full range of allowed sequences for particular structures. Methods based on Monte Carlo sampling or genetic algorithms can be applied to arbitrarily large proteins, but there is no guarantee of convergence to an appropriate minimum with respect to sequence. Indeed, in some cases, designed sequences fold to structures other than the target.[126] Pruning methods such as "dead end elimination"[51,53] can provide a more extensive search among sequences, but these methods are restricted

to scoring (energy) functions that involve only one- and two-body interactions. There is increasing evidence that the use of three-body and higher order terms may be necessary to recover folding to unique structures.[134,157] With the pruning methods, it can be difficult to include global constraints on sequence such as the numbers of each amino acid, and calculating the energies of alternate compact, nontarget conformations is computationally prohibitive. Pruning methods are designed to determine the global optimum, but many naturally occurring proteins are marginally stable, so it would be useful to develop methods to identify less than optimal sequences. Because they rely on the explicit generation of sequences, both stochastic search and pruning methods can only sample very small portions of sequence space for realistic representations of proteins (ca. 100 residues and 20 amino acids). Both the atomistic and the simplified approaches are sensitive to the energy or scoring function used. All energy functions in use for protein design are approximate, but the results of any search algorithm depend sensitively on the this energy function. This is not an issue in the context of model systems, but for applications to real proteins, uncertainties in the energy function may not merit such detailed search algorithms. In arriving at sequences that function, e.g., those that bind another molecule, modifications of the sequence from the optimum are likely to be necessary. Thus, it is important to develop methods that can provide information about suboptimal sequences for a given structure and that can include arbitrary constraints on sequences, such as those related to function or the features of the energy landscape. Such computational methods will also have application to a new class of protein design studies, combinatorial experiments.

Protein combinatorial experiments, wherein libraries of sequences are created and screened for evidence of folding to a predetermined structure, provide a means for broad-scale investigation of sequence variability. Recent developments in the use of combinatorial methods for de novo protein design are discussed in the article by Moffet and Hecht.[205] Such combinatorial libraries are usually created using recombinant methods, and molecules are selected for "protein-like" properties, oftentimes using a binding assay. Experiments of this type can explore a large number of sequences, and their results can shed light on the properties that foldable sequences in a library share. Peptides with protein-like properties have been isolated from random sequence libraries.[206,207] Combinatorial surveys have also been used to identify a variety of sequences that are consistent with a particular folded state structure.[208−213] Hecht and coworkers showed that a simple patterning of polar and nonpolar residues consistent with a four-helix bundle can yield sequences that are protein-like in being compact and having significant secondary structure.[209,214] These studies also reveal that dictating the specific details of contacts between residues may not be necessary for designing novel proteins. Axe et al. found combinatorial mutants of barnase wherein almost the entire hydrophobic core of the enzyme was modified.[213] As mentioned in the previous section, using combinatorial experiments, folding sequences have been found using a reduced set of amino acids.[198] The Baker group also studied other issues in protein folding, including the evolutionary selection of protein stability vis a vis folding kinetics[211] and the role of hydrophobic residues on the surface of a protein.[215] Ruan et al. found that for the pro region of Subtilisin, the number of sequences with maximal stability is small.[210] Thus, combinatorial experiments provide new routes to probe the determinants and features of folding. The large numbers of possible sequences, however, complicate these types of experiments, and limitations must be placed on the sequences so that the results are interpretable. Such limitations are often guided by qualitative chemical considerations, but a more quantitative computational theory would helpful in designing and interpreting these types of experiment.

Surveying the complete sequence landscape of proteins and other chain molecules seems at first glance intractable to both experiment and computation. Even a moderately sized protein of $N = 100$ residues has more than 10^{130} possible sequences. Recent studies using minimalist models, however, have found that the number of sequences that fold to a given structure is directly related to the degree of difficulty in arriving at these sequences.[158] Hence, estimating the number of foldable sequences is an important component of understanding the determinants of folding. Much has been learned from nature's set of protein sequences, and surveying whole libraries of folding sequences can reveal trends as to what interactions stabilize particular structures. In addition, many examples exist in nature of dissimilar sequences folding to essentially the same structure. Given these important issues, effort has been made in developing a statistical theory of sequences compatible with a given structure.[87−89] Such a theory will be extremely useful for designing combinatorial libraries of proteins, where huge numbers of sequences are possible. The theory of combinatorial libraries for folding molecules addresses the large numbers of possible sequences. The theory also incorporates a molecular understanding of the interactions involved in folding. This statistical approach is complementary to search methods and provides a different vantage on the sequence design problem. The theory addresses the whole space of available compositions, not just the small fractions that are accessible to experiment and to computational enumeration and sampling. One of the main goals is to aid protein chemists in discovering sequences that fold to a desired protein architecture through the use of appropriately designed combinatorial experiments. The theory also provides an aerial view of the "sequence landscape" and gives clues as to the relative importance of the different intramolecular interactions that stabilize particular three-dimensional structures.

The statistical approach to characterizing protein libraries addresses the number and composition of sequences compatible with a particular folded protein structure. Because of the exponential dependence on the number of residues N, S is focused on instead,

where S is the logarithm of the number of sequences for a target structure. If the energy of the sequences is fixed, S is equivalent to a microcanonical entropy, the *sequence entropy*. As in statistical thermodynamics, a maximum entropy approach is used where S is maximized with respect to any unconstrained internal parameters. Here the internal parameters are the probabilities $w_i(\alpha)$ that each residue position i in a sequence is occupied by amino acid type α.

$$S = -\sum_{i=1}^{N} \sum_{\alpha_i=1}^{m} w_i(\alpha_i) \ln w_i(\alpha_i) \qquad (15)$$

N is the chain length, and m is the number of possible monomer types. S is maximized subject to constraints on the sequences using the method of Lagrange multipliers.[124] The constraints need only be functions of the monomer probabilities. The constraints may specify values of global quantities that appear in the energy landscape theory, such as the folded state energy E_f or the stability gap Δ. "Patterning" constraints can also be included, where certain amino acids are precluded from occupying particular sites,[209] as well as composition constraints, where a specified number of each type of monomer is used in making the sequences in the library. For a given structure and energy function, a set of coupled, self-consistent equations is solved numerically to yield $w_i(\alpha)$. In this way, the number and composition of sequences having particular values of E_f, Δ, or any other physical or synthetic preconditions may be determined. Other constraints may be easily included. Upon introducing constraints, the number of different molecules in a particular chemical ensemble decreases. This reduction in library size is due to a fundamental concept in statistical thermodynamics: the imposition of any internal constraints in a system decreases the overall entropy. Thus, the theory may be used to design and focus combinatorial experiments. For a given target structure, the ramifications of such constraints on the number and identities of allowed sequences can be quickly investigated. For example, correlations between monomers can be examined by constraining the identity of one position and examining how this affects the identities of nearby residues. Being a form of heterogeneneous mean-field theory, the computational time necessary for the method goes as N^a, where $a = 1-2$. In contrast, the time required for explicit tabulation is exponentially dependent on N. Thus, the theory provides a tractable method of characterizing and designing sequence ensembles of proteins, where typically $N = 10^1-10^3$.

The theory has been applied to an exactly solvable system, a 27-mer cubic lattice polymer having only two types of amino acid. The exact enumeration of all 2^{27} sequences is computationally facile.[71,87] For a "protein-like" energy function,[71] the theory is in excellent agreement with the exact results for both $S(E_f)$ and the sequence identity probabilities $w_i(\alpha)$. The theory may also be used to focus libraries on regions having lower values of the target state energy E_f by fixing the hydrophobicity of buried residues. The theory may also be used to directly determine the "designability" of a structure, since it provides an estimate of the number of sequences as a function of the energy.[87] These methods have been extended so that the distribution of the stability gap Δ may also be estimated. To specify Δ, the set of all 103 346 compact, cubic conformations was used as an ensemble of nonfolded states.[61] Using the theory, the number and composition of sequences in a library as functions of both E_f and Δ may be examined. The theory is in excellent agreement with the results of the exact enumeration. The range and shape of the sequence entropy are recovered quantitatively by the theory. There is only a weak correlation between E_f and Δ, in agreement with the notion that energy minimization alone is likely to be insufficient for sequence design, but these two quantities are strongly correlated when the number of each amino acid is constrained, in agreement with previous design algorithms.[136] The theoretical estimates for $w_i(\alpha)$ are also in excellent agreement with the exact results for different values of E_f and Δ.[88]

In a recent study, this method has been extended and applied to realistic representations of proteins, which include the effects of side-chain packing in an atom-based manner.[89] The method has been applied to calculate the sequence probabilities of the immunoglobulin light chain-binding domain of protein L. This protein is an excellent target for the theory. Twenty-one different backbone models consistent with the NMR data are known. The different models permit the backbone sensitivity of the results to be examined. Combinatorial experiments on the protein using phage display selection have provided many variant folding sequences that bind to IgG.[211,215] The probabilities of amino acids have been calculated for three different secondary structural subunits of protein L and compared with experimentally observed amino acid frequencies. The theory alleviates the dependence of the amino state probabilities on backbone structure observed in many atom-based design algorithms.[216] The method is sufficiently rapid that many backbone structures may be considered and those features that are robust with respect to minor structure modifications may be identified. The folded state energy E_c° (or effective temperature T_c°) at which the generality of the results across different similar backbone structures breaks down is identified by a peak in an effective heat capacity C_v, which is directly related to fluctuations in the energies of the sequence−rotamer states. In the application to protein L, the probabilities for each amino acid in each of three secondary structure elements are consistent with the experimental studies, particularly with regard to the placement of hydrophobic residues. Although there is striking agreement between the theoretical and observed amino acid probabilities in some cases, precise quantitative agreement is not obtained throughout, which may be due in part to the sparse sampling of experimental sequences. The theory provides an important first step in the development of computational methods that address arbitrary numbers of sequences. Furthermore, such a theory should provide a useful framework to motivate and design combinatorial experiments that provide a larger sampling of possible sequences.

VI. Concluding Remarks

As this special issue attests, the field of protein design has substantially advanced in recent years and has had some spectacular successes.[32−35,40] While many of these efforts were guided by atom-based modeling about a fixed target backbone, energy landscape ideas have proven powerful in the "laboratory" of simple protein models and are beginning to have an impact in experimental protein design. These ideas are particulary important in discussing the stability of the target state in light of competing nontarget structures. For example, Hill and DeGrado indentified residues that are important for destabilizing nonfolded structures in a helical dimer.[217] Explicit consideration of both folded and unfolded states has led to the successful design of totally nonbiological folding polymers.[204,218] The simplifications provided by the energy landscape picture and verified using simple models will also likely prove useful in the design of larger proteins having 100 residues or more.

Understanding how sequence variation affects foldability will become increasingly important as more elements are included in the design process. The experimental design of not only stable but fast folding sequences is likely to see much progress, especially in light of recent advances in understanding how folding kinetics is reflected in a particular native state structure.[219−226] Predetermining which residues are variable and which are conserved will also be vital in the design of particular protein functions such as metal binding,[42] molecular recognition, and electron transfer,[227] so that residues can be identified whose variation will improve these properties without affecting the stability of a particular structural scaffold. Given that sequence variation can affect stability, kinetics, and function, it will be important to develop unified pictures for real proteins that treat each of these issues. Energy landscape-based methods for identifying sequences and characterizing the sequence space of a target structure provide such a vehicle for doing so.

VII. Acknowledgments

The author gratefully acknowledges support from the University of Pennsylvania and its Research Foundation, from the NSF in the form of a CAREER Award (NSF CHE-9984752), from the Research Corporation in the form of a Research Innovation Award and a Cottrell Scholars Award, and from the Arnold and Mabel Beckman Foundation via its Young Investigators Program.

VIII. References

(1) Levinthal, C. How to fold graciously. In *Mossbauer spectroscopy in biological systems*; DeBrunner, P., Tsibris, J., Munck, E., Eds.; Universtiy of Illinois Press: Urbana, IL, 1969.
(2) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. *Annu. Rev. Phys. Chem.* **1997**, *48*, 539.
(3) Gō, N. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 559.
(4) Lau, K. F.; Dill, K. A. *Macromolecules* **1989**, *22*, 3986.
(5) Shakhnovich, E.; Gutin, A. *J. Chem. Phys.* **1990**, *93*, 5967.
(6) Honeycutt, J. D.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 3526.
(7) Skolnick, J.; Kolinski, A. *J. Mol. Biol.* **1990**, *212*, 819.
(8) Leopold, P. E.; Montal, M.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 8721.
(9) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins* **1995**, *21*, 167.
(10) Onuchic, J. N.; Nymeyer, H.; Garcia, A. E.; Chahine, J.; Socci, N. D. *Adv. Protein Chem.* **2000**, *53*, 87.
(11) Dill, K. A.; Bromberg, S.; Yue, K.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D.; Chan, H. S. *Protein Sci* **1995**, *4*, 561.
(12) Veitshans, T.; Klimov, D.; Thirumalai, D. *Folding Des.* **1996**, *2*, 1.
(13) Shakhnovich, E. I. *Folding Des.* **1998**, *3*, R45.
(14) Pande, V. S.; Grosberg, A. Y.; Tanaka, T. *rmp 72*, 259.
(15) Bryngelson, J. D.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524.
(16) Shakhnovich, E. I.; Finkelstein, A. V. *Biopolymers* **1989**, *28*, 1667.
(17) Dill, K. A.; Fiebig, K. M.; Chan, H. S. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 1942.
(18) Wolynes, P. G.; Onuchic, J. N.; Thirumalai, D. *Science* **1995**, *267*, 1619.
(19) Daggett, V.; Levitt, M. *J. Mol. Biol.* **1993**, *232*, 600.
(20) Creamer, T. P.; Srinivasan, R.; Rose, G. D. *Biochemistry* **1995**, *34*, 16245.
(21) Creamer, T. P.; Srinivasan, R.; Rose, G. D. *Biochemistry* **1997**, *36*, 2832.
(22) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *218*, 534.
(23) Shakhnovich, E. I.; Gutin, A. M. *Biophys. Chem.* **1989**, *34*, 187.
(24) Wootton, J. C. *Curr. Opin. Struct. Biol.* **1994**, *4*, 413.
(25) Nymeyer, H.; Garc×c1a, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5921.
(26) Gō, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183.
(27) Saven, J. G.; Wolynes, P. G. *J. Mol. Biol.* **1996**, *257*, 199.
(28) Shakhnovich, E. I. *Folding Des.* **1996**, *1*, R50.
(29) Pjura, P.; Matthews, B. W. *Protein Sci.* **1993**, *2*, 2226.
(30) Richards, F. M.; Lim, W. A. *Q. Rev. Biophys.* **1993**, *26*, 423.
(31) Hecht, M. H.; Richardson, J. S.; Richardson, D. C.; Ogden, R. C. *Science* **1990**, *249*, 884.
(32) Dahiyat, B. I.; Mayo, S. L. *Science* **1997**, *278*, 82.
(33) Harbury, P. B.; Plecs, J. J.; Tidor, B.; Alber, T.; Kim, P. S. *Science* **1998**, *282*, 1462.
(34) Kuroda, Y.; Hamada, D.; Tanaka, T.; Goto, Y. *Folding Des.* **1996**, *1*, 255.
(35) Kortemme, T.; Ramirez-Alvarado, M.; Serrano, L. *Science* **1998**, *281*, 253.
(36) Schafmeister, C. E.; LaPorte, S. L.; Miercke, L. J. W.; Stroud, R. M. *Nat. Struct. Biol.* **1997**, *4*, 1039.
(37) Hill, R. B.; DeGrado, W. F. *J. Am. Chem. Soc.* **1998**, *120*, 1138.
(38) Walsh, S. T. R.; Cheng, H.; Bryson, J. W.; Roder, H.; DeGrado, W. F. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 5486.
(39) Bryson, J. W.; Betz, S. F.; Lu, H. S.; Suich, D. J.; Zhou, H. X.; O'Neil, K. T.; DeGrado, W. F. *Science* **1995**, *270*, 935.
(40) DeGrado, W. F.; Summa, C. M.; Pavone, V.; Nastri, F.; Lombardi, A. *Annu. Rev. Biochem.* **1999**, *68*, 779.
(41) Street, A. G.; Mayo, S. L. *Struct. Folding Des.* **1999**, *7*, R105.
(42) Summa, C. M.; Lombardi, A.; Lewis, M.; DeGrado, W. F. *Curr. Opin. Struct. Biol.* **1999**, *9*, 500.
(43) Gordon, D. B.; Marshall, S. A.; Mayo, S. L. *Curr. Opin. Struct. Biol.* **1999**, *9*, 509.
(44) Ponder, J. W.; Richards, F. M. *J. Mol. Biol.* **1987**, *193*, 775.
(45) Hellinga, H. W.; Richards, F. M. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 5803.
(46) Kono, H.; Doi, J. *Proteins* **1994**, *19*, 244.
(47) Desjarlais, J. R.; Handel, T. M. *Protein Sci.* **1995**, *4*, 2006.
(48) Jiang, X.; Bishop, E. J.; Farid, R. S. *J. Am. Chem. Soc.* **1997**, *119*, 838.
(49) Bryson, J. W.; Desjarlais, J. R.; Handel, T. M.; DeGrado, W. F. *Protein Sci.* **1998**, *7*, 1404.
(50) H. Kono, M. Nishiyama, M. T.; Doi, J. *Protein Eng.* **1998**, *11*, 47.
(51) Desmet, J.; Demaeyer, M.; Hazes, B.; Lasters, I. *Nature* **1992**, *356*, 539.
(52) Dahiyat, B.; Mayo, S. L. *Protein Sci.* **1996**, *5*, 895.
(53) Dahiyat, B. I.; Sarisky, C. A.; Mayo, S. L. *J. Mol. Biol.* **1997**, *273*, 789.
(54) Voigt, C. A.; Gordon, D. B.; Mayo, S. L. *J. Mol. Biol.* **2000**, *299*, 789.
(55) Street, A. G.; Mayo, S. L. *Folding Des.* **1998**, *3*, 253.
(56) Raha, K.; Wollacott, A. M.; Italia, M. J.; Desjarlais, J. R. *Protein Sci.* **2000**, *9*, 1106.
(57) Levitt, M.; Warshel, A. *Nature* **1975**, *253*, 694.
(58) Kolinski, A.; Milik, M.; Skolnick, J. *J. Chem. Phys.* **1990**, *94*, 3978.
(59) Kolinski, A.; Skolnick, J. *Proteins* **1994**, *18*, 338.
(60) Hinds, D. A.; Levitt, M. *J. Mol. Biol.* **1996**, *258*, 201.
(61) Šali, A.; Shakhnovich, E.; Karplus, M. *J. Mol. Biol.* **1994**, *235*, 1614.
(62) Socci, N. D.; Onuchic, J. N. *J. Chem. Phys.* **1994**, *101*, 1519.
(63) Tanaka, S.; Scheraga, H. *Macromolecules* **1976**, *9*, 945.
(64) Goldstein, R.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 9029.

(65) Jones, D. T.; Thornton, J. M. *Curr. Opin. Struct. Biol.* **1996**, *6*, 210.
(66) Skolnick, J.; Kolinski, A.; Ortiz, A. R. *J. Biomol. Struct. Dyn.* **1998**, *16*, 381.
(67) Chothia, C. *Nature* **1992**, *357*, 543.
(68) Orengo, C. A.; Jones, D. T.; Thornton, J. M. *Nature* **1994**, *372*, 631.
(69) Wolf, Y. I.; Grishin, N. V.; Koonin, E. V. *J. Mol. Biol.* **2000**, *299*, 897.
(70) Farber, G. K.; Petsko, G. A. *Trends Biochem. Sci.* **1990**, *15*, 228.
(71) Li, H.; Helling, R.; Tang, C.; Wingreen, N. *Science* **1996**, *273*, 666.
(72) Finkelstein, A. V.; Gutin, A. M.; Badretdinov, A. Y. *FEBS Lett.* **1993**, *325*, 23.
(73) Derrida, B. *Phys. Rev. B.* **1981**, *24*, 2613.
(74) Shakhnovich, E. I.; Gutin, A. M. *Nature* **1990**, *346*, 773.
(75) Finkelstein, A. V.; Badretdinov, A. Y.; Gutin, A. M. *Proteins* **1995**, *23*, 142.
(76) Yue, K.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 146.
(77) Lingård, P.-A.; Bohr, H. *Phys. Rev. Lett.* **1996**, *77*, 779.
(78) Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 14249.
(79) Nelson, E. D.; Teneyck, L. F.; Onuchic, J. N. *Phys. Rev. Lett.* **1997**, *79*, 3534.
(80) Wang, T.; Miller, J.; Wingreen, N. S.; Tang, C.; Dill, K. A. *J. Chem. Phys.* **2000**, 8329.
(81) Bowie, J. U.; Lüthy, R.; Eisenberg, D. *Science* **1991**, *253*, 164.
(82) Li, H.; Tang, C.; Wingreen, N. S. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 4987.
(83) Kussell, E. L.; Shakhnovich, E. I. *Phys. Rev. Lett.* **1999**, *83*, 4437.
(84) Govindarajan, S.; Goldstein, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 3341.
(85) Buchler, N. E. G.; Goldstein, R. A. *J. Chem. Phys.* **1999**, *111*, 6599.
(86) Ejtehadi, M. R.; Hamedani, N.; Seyed-Allaei, H.; Shahrezaei, V.; Yahyanejad, M. *J. Phys. A: Math. Gen.* **1998**, *31*, 6141.
(87) Saven, J. G.; Wolynes, P. G. *J. Phys. Chem. B* **1997**, *101*, 8375.
(88) Zou, J.; Saven, J. G. *J. Mol. Biol.* **2000**, *296*, 281.
(89) Kono, H.; Saven, J. G. *J. Mol. Biol.* **2001**, *306*, 607.
(90) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
(91) Weiner, J. S.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G. Profeta, S. J.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765.
(92) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657.
(93) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. J.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
(94) Brooks, C. L.; Case, D. A. *Chem. Rev.* **1993**, *93*, 2487.
(95) Dunbrack, R. L.; Cohen, F. E. *Protein Sci.* **1997**, *6*, 1661.
(96) Karplus, M.; Brünger, A. T.; Elber, R.; Kuriyan, J. Molecular dynamics: applications to proteins. In *Cold Spring Harbor Symposia on Quantitative Biology*; Cold Spring Harbor Laboratory: Cold Spring Harbor, New York, 1987; Vol. 52.
(97) Boczko, E. M.; Brooks, C. L. *Science* **1995**, *269*, 393.
(98) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740.
(99) Takada, S.; Luthey-Schulten, Z.; Wolynes, P. G. *J. Chem. Phys.* **1999**, *110*, 11616.
(100) Collet, O.; Premilat, S.; Maigret, B.; Scheraga, H. A. *Biopolymers* **1997**, *47*, 363.
(101) Pillardy, A.; Czaplewski, C.; Liwo, A.; Lee, J.; Ripoll, D. R.; Kazmierkiewicz, R.; Oldziej, S.; Wedemeyer, W. J.; Gibson, K. D.; Arnautova, Y. A.; Saunders: J.; Ye, Y. J.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 2329.
(102) Pande, V. S.; Rokhsar, D. S. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 1273.
(103) Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. G. *Proteins* **1999**, *34*, 281.
(104) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859.
(105) Godzik, A.; Kolinski, A.; Skolnick, J. *Protein Sci.* **1995**, *4*, 2107.
(106) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623.
(107) Bryngelson, J. D. *J. Chem. Phys.* **1994**, *100*, 6038.
(108) Pande, V. S.; Grosberg, A. Y.; Tanaka, T. *J. Chem. Phys.* **1995**, *103*, 9482.
(109) Rooman, M. J.; Wodak, S. J. *Protein Eng.* **1996**, *8*, 849.
(110) Thomas, P. D.; Dill, K. A. *J. Mol. Biol.* **1996**, *257*, 457.
(111) Pereira de Araújo, A. F.; Pochapsky, T. C. *Folding Des.* **1996**, *1*, 299.
(112) Maiorov, V. N.; Crippen, G. D. *J. Mol. Biol.* **1992**, *227*, 867.
(113) Mirny, L. A.; Shakhnovich, E. I. *J. Mol. Biol.* **1996**, *264*, 1164.
(114) Hao, M.-H.; Scheraga, H. A. *Curr. Opin. Struct. Biol.* **1999**, *9*, 184.
(115) Chiu, T.; Goldstein, R. A. *Protein Eng.* **1998**, *11*, 749.
(116) Koretke, K. K.; Luthey-Schulten, Z.; Wolynes, P. G. *Protein Sci.* **1996**, *5*, 1043.
(117) Hao, M.-H.; Scheraga, H. A. *J. Phys. Chem.* **1996**, *100*, 14540.
(118) Hao, M.-H.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 4984.
(119) Guo, Z.; Thirumalai, D.; Honeycutt, J. D. *J. Chem. Phys.* **1992**, *97*, 525.
(120) Sippl, M. J. *Curr. Opin. Struct. Biol.* **1995**, *5*, 229.
(121) Sasai, M. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 8438.
(122) Guo, Z.; Thirumalai, D. *J. Mol. Biol.* **1996**, *263*, 323.
(123) Koretke, K. K.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 2932.
(124) McQuarrie, D. A. *Statistical Mechanics*; Harper & Row: New York, 1976.
(125) Lau, K. F.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 638.
(126) Yue, K.; Fiebig, K. M.; Thomas, P. D.; Chan, H. S.; Shakhnovich, E. I.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 325.
(127) Socci, N. D.; Onuchic, J. N. *J. Chem. Phys.* **1995**, *103*, 4732.
(128) Klimov, D. K.; Thirumalai, D. *Phys. Rev. Lett.* **1996**, *76*, 4070.
(129) Klimov, D. K.; Thirumalai, D. *J. Chem. Phys.* **1998**, *109*, 4119.
(130) Dinner, A. R.; Abkevich, V.; Shakhnovich, E.; Karplus, M. *Proteins* **1999**, *35*, 34.
(131) Shakhnovich, E. I. *Phys. Rev. Lett.* **1994**, *72*, 3907.
(132) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Oldziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849.
(133) Vendruscolo, M.; Domany, E. *J. Chem. Phys.* **1998**, *109*, 11101.
(134) Betancourt, M. R.; Thirumalai, D. *Protein Sci.* **1999**, *8*, 361.
(135) Tobi, D.; Elber, R. *Proteins* **2000**, *41*, 40.
(136) Shakhnovich, E. I.; Gutin, A. M. *Protein Eng.* **1993**, *6*, 793.
(137) Ramanathan, S.; Shakhnovich, E. *Phys. Rev. E* **1994**, *50*, 1303.
(138) Pande, V. S.; Grosberg, A. Y.; Tanaka, T. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 12976.
(139) Jones, D. T. *Protein Sci.* **1994**, *3*, 567.
(140) Toda, M.; Kubo, R.; Saito, N. *Statistical Physics I. Equilibrium Statistical Mechanics*, 2nd ed.; Springer-Verlag: Berlin, 1991.
(141) Pathria, R. K. *Statistical Mechanics*, 2nd ed.; Butterworth Heinemann: Oxford, 1996.
(142) Elber, R.; Karplus, M. *J. Am. Chem. Soc.* **1990**, *112*, 9161.
(143) Finkelstein, A. V.; Reva, B. A. *Nature* **1991**, *351*, 497.
(144) Reva, B. A.; Finkelstein, A. V. *Protein Eng.* **1992**, *5*, 625.
(145) Lee, C. *J. Mol. Biol.* **1994**, *236*, 918.
(146) Koehl, P.; Delarue, M. *J. Mol. Biol.* **1994**, *239*, 249.
(147) Koehl, P.; Delarue, M. *Curr. Opin. Struct. Biol.* **1996**, *6*, 222.
(148) Vásquez, M. *Biopolymers* **1995**, *36*, 53.
(149) Kono, H.; Doi, J. *J. Comput. Chem.* **1996**, *17*, 1667.
(150) Mendes, J.; Soares, C. M.; Carrondo, M. A. *Biopolymers* **1999**, *50*, 111.
(151) Cootes, A. P.; Curmi, P. M. G.; Torda, A. E. *J. Chem. Phys.* **2000**, *113*, 2489.
(152) Foreman, K. W.; Phillips, A. T.; Rosen, J. B.; Dill, K. A. *J. Comput. Chem.* **1999**, *20*, 1527.
(153) Strop, P.; Mayo, S. L. *J. Am. Chem. Soc.* **1999**, *121*, 2341.
(154) Street, A. G.; Datta, D.; Gordon, D. B.; Mayo, S. L. *Phys. Rev. Lett.* **2000**, *84*, 5010.
(155) Gordon, D. B.; Mayo, S. L. *J. Comput. Chem.* **1998**, *1998*, 1505.
(156) Pierce, N.; Spriet, J. A.; Desmet, J.; Mayo, S. L. *J. Comput. Chem.* **2000**, *21*, 999.
(157) Liwo, A.; Kazmierkiewicz, R.; Czaplewski, C.; Groth, M.; Oldziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A. *J. Comput. Chem.* **1998**, *19*, 259.
(158) Nelson, E. D.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 10682.
(159) Tiana, G.; Broglia, R. A.; Shakhnovich, E. I. *Proteins* **2000**, *39*, 244.
(160) Finkelstein, A. V.; Shakhnovich, E. I. *Biopolymers* **1989**, *28*, 1681.
(161) Kuwajima, K. *Proteins* **1989**, *6*, 87.
(162) Miranker, A. D.; Dobson, C. M. *Curr. Opin. Struct. Biol.* **1996**, *6*, 31.
(163) Matthews, B. W. *Annu. Rev. Biochem.* **1993**, *62*, 139.
(164) Kuhlman, B. K.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10383.
(165) Yomo, Y.; Saito, S.; Sasai, M. *Nat. Struct. Biol.* **1999**, *6*, 743.
(166) Li, L.; Mirny, L. A.; Shakhnovich, E. I. *Nat. Struct. Biol.* **2000**, *7*, 336.
(167) Mélin, R.; Li, H.; Wingreen, N. S.; Tang, C. *J. Chem. Phys.* **1999**, *110*, 1252.
(168) Anfinsen, C. B. *Science* **1973**, *181*, 223.
(169) Seno, F.; Vendruscolo, M.; Maritan, A.; Banavar, J. R. *Phys. Rev. Lett.* **1996**, *77*, 1901.
(170) Micheletti, C.; Seno, F.; Maritan, A.; Banavar, J. R. *Phys. Rev. Lett.* **1998**, *80*, 2237.
(171) Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4918.
(172) Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. *Proteins* **1998**, *32*, 136.
(173) Goldstein, R. A.; Luthey-Schulten, Z.; Wolynes, P. G. Protein Tertiary Structure Recognition Using Optimized Associative Memory Hamiltonians. In *26th Hawaii International Conference on System Sciences*; T. N. Mudge, V. M., Hunter, L., Eds.; IEEE Computer Society Press: Los Alamitos, CA, 1993; p 699−707.
(174) Shakhnovich, E.; Abkevich, V.; Ptitsyn, O. *Nature* **1996**, *379*, 96.

(175) Tiana, G.; Broglia, R. A.; Roman, H. E.; Vegezzi, E.; Shakhnovich, E. *J. Chem. Phys.* **1998**, *108*, 757.
(176) Micheletti, C.; Banavar, J. R.; Maritan, A.; Seno, F. *Phys. Rev. Lett.* **1998**, *80*, 5683.
(177) Deutsch, J. M.; Kurosky, T. *Phys. Rev. Lett.* **1996**, *76*, 323.
(178) Morrissey, M. P.; Shakhnovich, E. I. *Folding Des.* **1996**, *1*, 391.
(179) Onuchic, J. N.; Wolynes, P. G.; Luthey-Schulten, Z.; Socci, N. D. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 3626.
(180) Rossi, A.; Micheletti, C.; Seno, F.; Maritan, A. *Biophys. J.* **2000**, *80*, 480.
(181) Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. *Folding Des.* **1996**, *1*, 221.
(182) Zhang, L.; Skolnick, J. *Protein Sci.* **1998**, *7*, 1201.
(183) Seno, F.; Micheletti, C.; Maritan, A.; Banavar, J. R. *Phys. Rev. Lett.* **1998**, *81*, 2172.
(184) Micheletti, C.; Seno, F.; Maritan, A.; Banavar, J. R. *Proteins* **1998**, *32*, 80.
(185) Sun, S.; Brem, R.; Chan, H. S.; Dill, K. A. *Protein Eng.* **1995**, *8*, 1205.
(186) Rossi, A.; Maritan, A.; Micheletti, C. *J. Chem. Phys.* **2000**, *112*, 2050.
(187) Irbäck, A.; Peterson, C.; Potthast, F.; Sandelin, E. *Phys. Rev. E* **1998**, *58*, R5249.
(188) Irbäck, A.; Peterson, C.; Potthast, F.; Sandelin, E. *Structure* **1999**, *7*, 347.
(189) Mirny, L. A.; Abkevich, V. I.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 4976.
(190) Pande, V. S.; Grosberg, A. Y.; Tanaka, T. *J. Chem. Phys.* **1994**, *101*, 8246.
(191) Saito, S.; Sasai, M.; Yomo, T. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 11324.
(192) Mendel, D.; Cornish, V. W.; Schultz, P. G. *Annu. Rev. Biophys. Biomol. Struct.* **1995**, *24*, 435.
(193) Kowal, A. K.; Khrer, C.; RajBhandary, U. L. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 2268.
(194) Plaxco, K. W.; Riddle, D. S.; Grantcharova, V.; Baker, D. *Curr. Opin. Struct. Biol.* **1998**, *8*, 80.
(195) Chan, H. S. *Nat. Struct. Biol.* **1999**, *6*, 994.
(196) Eisenberg, D.; Wilcox, W.; Eshita, S. M.; Pryciak, P. M.; Ho, S. P.; DeGrado, W. F. *Proteins* **1986**, *1*, 16.
(197) Regan, L.; DeGrado, W. F. *Science* **1988**, *241*, 976.
(198) Riddle, D. S.; Santiago, J. V.; Bray-Hall, S. T.; Doshi, N.; Grantcharova, V. P.; Yi, Q.; Baker, D. *Nat. Struct. Biol.* **1997**, *4*, 805.
(199) Wolynes, P. G. *Nat. Struct. Biol.* **1997**, *4*, 871.
(200) Murphy, L. R.; Wallqvist, A.; Levy, R. M. *Protein Eng.* **2000**, *13*, 149.
(201) Wang, J.; Wang, W. *Nat. Struct. Biol.* **1999**, *6*, 1033.
(202) Liang, H. *J. Chem. Phys.* **2000**, *113*, 4827.
(203) Buchler, N. E. G.; Goldstein, R. A. *Proteins* **1999**, *34*, 113.
(204) Cheng, R. P.; Gellman, S. H.; DeGrado, W. F. *Chem. Rev.* **2001**, *101*, 3219−3232.
(205) Moffet, D. A.; Hecht, M. H. *Chem. Rev.* **2001**, *101*, 3191−3204.
(206) Davidson, A. R.; Sauer, R. T. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 2146.
(207) LaBean, T. H.; Kauffman, S. A.; Butt, T. R. *Mol. Diversity* **1995**, *1*, 29.
(208) Reidhaar Olson, J. F.; Sauer, R. T. *Science* **1988**, *241*, 53.
(209) Kamtekar, S.; Schiffer, J. M.; Xiong, H.; Babik, J. M.; Hecht, M. H. *Science* **1993**, *262*, 1680.
(210) Ruan, B.; Hoskins, J.; Wang, L.; Bryan, P. N. *Protein Sci.* **1998**, *7*, 2345.
(211) Kim, D. E.; Gu, H. D.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 4982.
(212) Sauer, R. T. *Folding Des.* **1996**, *1*, R27.
(213) Axe, D. D.; Foster, N. W.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5590.
(214) Xiong, H.; Buckwalter, B. L.; Shieh, H.-M.; Hecht, M. H. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 6349.
(215) Gu, H.; Doshi, N.; Kim, K. T.; Simons, K. T.; Santiago, J. V.; Nauli, S.; Baker, D. *Protein Sci.* **1999**, *8*, 2734.
(216) Huang, E. S.; Koehl, P.; Levitt, M.; Pappu, R. V.; Ponder, J. W. *Proteins* **1998**, *33*, 204.
(217) Hill, R. B.; DeGrado, W. F. *Structure* **2000**, *8*, 471.
(218) Nelson, J. C.; Saven, J. G.; Moore, J. S.; Wolynes, P. G. *Science* **1997**, *277*, 1793.
(219) Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 6170.
(220) Plaxco, K. W.; Simons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 984.
(221) Muñoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11311.
(222) Galzitskaya, O. V.; Finkelstein, A. V. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11299.
(223) Shoemaker, B. A.; Wolynes, P. G. *J. Mol. Biol.* **1999**, *287*, 657.
(224) Shoemaker, B. A.; Wolynes, P. G. *J. Mol. Biol.* **1999**, *287*, 675.
(225) Clementi, C.; Nymeyer, H.; Onuchic, J. N. *J. Mol. Biol.* **2000**, *298*, 937.
(226) Dinner, A. R.; Karplus, M. *Nat. Struct. Biol.* **2001**, *8*, 21.
(227) Gibney, B. R.; Dutton, P. L. *Adv. Inorg. Chem.* **2001**, *51*, 409.

CR000058W