PROTEINS: Structure, Function, and Genetics 31:10–20 (1998)

RESEARCH ARTICLES

Structure-Based Design of Model Proteins

Jayanth R. Banavar,[1] Marek Cieplak,[2]* Amos Maritan,[3] Gautham Nadig,[4] Flavio Seno,[5] and Saraswathi Vishveshwara[4]

[1]*Department of Physics and Center for Materials Physics, 104 Davey Laboratory, Pennsylvania State University, University Park, Pennsylvania*

[2]*Institute of Physics, Polish Academy of Sciences, Warsaw, Poland*

[3]*Istituto Nazionale di Fisica della Materia, International School for Advanced Studies, and INFN sezione di Trieste, Trieste, Italy*

[4]*Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India*

[5]*Istituto Nazionale di Fisica della Materia, Dipartimento di Fisica, Universita' di Padova, Padova, Italy*

ABSTRACT **A structure-based, sequence-design procedure is proposed in which one considers a set of decoy structures that compete significantly with the target structure in being low energy conformations. The decoy structures are chosen to have strong overlaps in contacts with the putative native state. The procedure allows the design of sequences with large and small stability gaps in a random-bond heteropolymer model in both two and three dimensions by an appropriate assignment of the contact energies to both the native and nonnative contacts. The design procedure is also successfully applied to the two-dimensional HP model. Proteins 31:10–20, 1998.**

Key words: protein design; lattice model; protein stability

INTRODUCTION

Functionally useful proteins are sequences of amino acids that fold rapidly under appropriate conditions into their native states, commonly assumed to be their ground state conformations. The functionality of a protein is controlled by its native state structure. An outstanding vital problem is that of protein design (Cordes et al., 1996; Kuroda et al., 1994; Quinn et al., 1994; Lombardi et al., 1997; Kamtekar et al., 1993; Yue et al., 1995; Shakhnovich and Gutin, 1993; Yue et al., 1992; Sun et al., 1995; Kurosky and Deutsch, 1995; Deutsch and Kurosky, 1996; Seno et al., 1996; Morrisey and Shakhnovich, 1996; Ponder and Richards, 1987; Bowie et al., 1991; Pabo, 1983)—the identification of the sequence (or sequences) of amino acids that has a desired target structure as its ground state. The principal theme of this article is the development of a new strategy for the design problem that entails not only the study of sequences but also the conformations that significantly compete with the target structure in being the ground state of candidate sequences. Our work underscores the importance of the role of the interaction energies of both native and nonnative contacts in the making of an ideal folder.

The simplest and the best-studied design strategy consists of an exploration in sequence space in order to identify the sequence that has the lowest energy in the target structure. This is usually done for a fixed composition of the amino acids because there exists no mechanism within this procedure for choosing between sequences having differing compositions. Furthermore, the strategy is only approximate because there is no guarantee that just because a given sequence has a low energy in the target structure that it would not have an even lower energy in an alternative structure.

A physical and rigorous approach consists of exploring both the family of sequences and the family of conformations to identify a sequence (or sequences) that has a lower energy in the target structure than in any other conformation (Deutsch and Kurosky, 1996; Seno et al., 1996; Morrisey and Shakhnovich, 1996). Stated mathematically, in order to perform an inverse design on a target structure, Γ, one needs to identify the sequence of amino acids, S, that maximizes the "occupation probability" according to Boltz-

Abbreviations: HP, hydrophobic-polar; DSKS, Dinner-Sali-Karplus-Shakhnovich.

Contract grant sponsor: KBN (Poland); Contract grant number: 2P03B-025-13; Contract grant sponsor: NASA; Contract grant sponsor: NATO; Contract grant sponsor: SERC at the Indian Institute of Science; Contract grant sponsor: Center for Academic Computing at Penn State.

*Correspondence to: Marek Cieplak, Institute of Physics, Polish Academy of Sciences, 02-668 Warsaw, Poland.

Received 27 June 1997; Accepted 28 October 1997

mann statistics,

$$P_\Gamma(S) = \frac{e^{-\beta E_S(\Gamma)}}{\sum_{\Gamma'} e^{-\beta E_S(\Gamma')}} = \frac{e^{-\beta E_S(\Gamma)}}{Z_s}, \qquad (1)$$

evaluated at any convenient but sufficiently low temperature. For sequences with a unique ground state, any temperature below the folding transition temperature, at which the probability of occupancy of the native state is ½, would suffice. Γ' denotes the family of conformations that the sequence can adopt, β is equal to $1/k_BT$, (κ_B is the Boltzmann constant and T is the temperature) and $E_S(\Gamma)$ is the energy of the sequence S in conformation Γ. A brute force application of this rigorous design procedure, involving the exhaustive search of the sequence with maximum $P_\Gamma(S)$, is often not feasible because of the computational difficulty of determining an accurate value of Z_S for each of the sequences considered.

It has been recognized that a simple binary pattern of hydrophobic and hydrophilic residues along the polypeptide chain encode structure at the coarse-grained level (Cordes et al., 1996; Kamtekar et al., 1993). Thus, the simplest model of proteins consists of sequences made up of just two kinds of amino acids (H and P, representing hydrophobic and polar residues) configured as self-avoiding chains on a lattice and described by a contact Hamiltonian (Lau and Dill, 1989, 1990; Chan and Dill, 1991). Alternatively, one may consider multiple kinds of amino acids with a specified interaction matrix analogous to the one deduced by Miyazawa and Jernigan (1985) for the amino acids of real proteins using a quasi-chemical method. Such models are known to adequately describe proteins at the coarse-grained level with the advantage that the native states can be determined exactly (Lau and Dill, 1989, 1990; Chan and Dill, 1991; Dill et al., 1991; Onuchic et al., 1995; Camacho and Thirumalai, 1993; Sali et al., 1994; Bryngelson et al., 1995). Furthermore, they provide a controlled laboratory for theoretical investigations and rigorous testing of concepts and ideas for future use in studies on real proteins.

The design of "good" real proteins that are thermodynamically stable and rapidly folding may be facilitated somewhat by the fact that the diversity of the constituent amino acids allow for their harmonious packing into the native structure, so that frustrating influences are minimized (Cordes et al., 1996; Ponder and Richards, 1987; Pabo, 1983; Go and Abe, 1981; Bryngelson and Wolynes, 1987; Shrivastava et al., 1995; Cieplak et al., 1996; Cieplak and Banavar, 1997). It has been suggested (Unger et al., 1996; Abkevich et al., 1995) that a dominant factor controlling the foldability of a sequence is the strength of possible interactions between residues close in sequence. It has also been recognized that a key ingredient in a successful design strategy is the necessity to take into account, for any given sequence, the conformations that compete significantly with the target structure to be its ground state, an idea called negative design (Yue and Dill, 1992; Richardson et al., 1992).

RANDOM-BOND MODEL

We begin with a heteropolymer model in which one considers self-avoiding chains N monomers long, comprised of one each of N amino acids, configured as a self-avoiding chain on a lattice. The amino acids are labeled 1, 2, 3, —, N in sequence. For a given self-avoiding conformation, the contact Hamiltonian is given by a symmetric interaction matrix, B_{ij}, denoting the interaction energy between the i-th and j-th monomers that are in contact (next to each other on the lattice but yet not next to each other along the chain). The physically correct way to generate a new sequence from the original one (assuming that the composition of the amino acids is fixed) is by interchanging one of the amino acids along the sequence with another. We will begin a much simpler bond-random analog of the site-random problem which instead allows an interchange of one of the B_{ij} values with another. In this approach a "sequence" is specified when a B_{ij} is assigned to each pair of nodes i and j. We will return to the physically correct approach later in the article to show that our principal results are independent of this simplifying assumption.

For a given choice of the B_{ij} interaction matrix, a sequence corresponding to minimal frustration can be designed using a generalization of the Go-Abe (Go and Abe, 1981) approach by assigning the most attractive of the B_{ij} values to the native contacts of the target structure. This ensures that the ground state of the designed sequence is indeed in the target structure and the ground state is a deep minimum that is thermodynamically stable (Shrivastava et al., 1995). Studies of the kinetics of such a designed sequence have shown that it folds rapidly and consistently to the target structure and that one may observe the formation of a folding funnel into the native state (Shrivastava et al., 1995; Cieplak et al., 1996; Cieplak and Banavar 1997). Strikingly, a random assignment of repulsive interactions to the nonnative contacts leads to a further stabilization of the native state and even more rapid folding kinetics (Shrivastava et al., 1995).

In this article, we study the $N = 27$ case on a cubic lattice, for which all maximally compact conformations can be enumerated exactly and the $N = 16$ case on a square lattice for which all conformations (maximally compact and all others) can be accounted for. The calculations of the thermodynamics of the three-dimensional model are carried out approximately using only the maximally compact conformations and the validity of the results are assessed within the two-dimensional model for which exact calculations are feasible. The use of only the maxi-

mally compact conformations in the three-dimensional calculations is a fairly standard approximation and is justified when the contact energies have a large overall attractive shift that promotes the compact structures.

For a maximally compact target structure, there are K native contacts and M nonnative contacts. In the three-dimensional $N = 27$ model, $K = 28$ and $M = 128$, whereas in the two-dimensional $N = 16$ model, $K = 9$ and $M = 40$. We begin by considering sequences obtained by full rank ordering of the contact energies, B_{ij}'s. These sequences are designed by assigning the K strongest attractive contacts to the native contacts of a target maximally compact conformation. The remaining M nonnative contacts are assigned to the other couplings randomly.

In our studies, we assume that $B_{ij} = \tilde{B}_{ij} + B_0$ where the values of $\tilde{B}_{ij}$ are from a Gaussian distribution with zero mean and unit variance and the shift B_0 is -2 and -1 in the three- and two-dimensional models, respectively.

There are $K!\ M!$ ways of constructing such sequences for a given set of values of the contact energies. Each sequence has its own folding temperature T_f and stability gap, Δ. T_f is defined as a temperature at which the probability to find a sequence in its native state is ½ and Δ is the difference in energy between the native state and the first excited state corresponding to another maximally compact conformation. It is expected that, statistically, large Δ's correlate with the larger values of T_f. Furthermore, a high measure of thermodynamic stability (as measured by a large Δ or equivalently a large T_f) is associated with an increased stability against mutations (Vendruscolo et al., 1997).

It is impossible to study the stability of all rank-ordered sequences. In order to assess the approximate spread in the values of Δ of these sequences, we devised a simple Monte Carlo procedure that involves starting from a randomly chosen rank-ordered sequence and making mutations corresponding to the interchange of the contact energies. The random swaps are carried out separately for the native and the nonnative sets of contacts with no mixing to ensure that one stays within the family of fully rank-ordered sequences. Figure 1 shows the result of our study in the form of a probability distribution of Δ after 10,000 swaps of the contact energies (2,000 of these swaps were among the 28 native contacts, whereas the remaining 8,000 corresponded to swaps of the 128 nonnative contacts). The inset of Figure 1 shows a plot of T_f versus Δ for randomly chosen sequences with a range of values of Δ and demonstrates a nice correlation between the two quantities. T_f here is calculated approximately, using only the maximally compact conformations in the denominator of Eq. 1.

We note that the simple design procedure of rank ordering is one that involves an optimal search in sequence space. We now turn to a design procedure

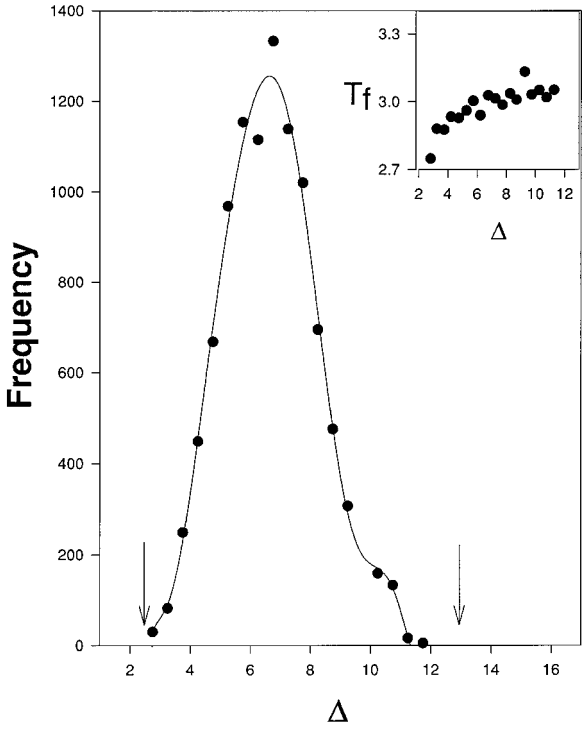

Fig. 1. Histogram of the stability gap (the energy difference between the maximally compact native state and the first excited maximally compact conformation) as derived from a randomly chosen, fully rank-ordered starting sequence followed by 2,000 random swaps of native contact energies and 8,000 random swaps of nonnative contact energies for the three-dimensional random-bond model. Arrows indicate the stability gap of the structure-based designed sequences S^- and S^+. The inset shows a correlation between the stability gap and the folding transition temperature (determined from the use of just the maximally compact conformations).

with information from the space of conformations that is complementary to the rank-ordering selection procedure. The basic idea behind this *structure-based design* is the identification of conformations (in a sequence-independent manner) that are significant competitors of the native state in being low energy states for candidate sequences. High thermodynamic and mutational stability would be expected to result from ensuring that the energies of these "excited" states be as high as possible. For a given set of contact energies, our goal is to find the sequences which are the "end members" and which correspond to maximum and minimum stability, as measured by Δ. These sequences will be denoted by S^+ and S^-, respectively.

Our strategy of search for the end members begins with the selection of a set of decoy structures that have maximal overlap with the target structure. The overlaps are defined in terms of the common contacts with the native state and their occurrence is based on the geometry of the conformations and not on any sequence-dependent properties. The more complete and faithful the set of the decoy structures is, the more accurate is the determination of the distribu-

tion of the overlaps and, thus, the better the approximation to the true S^+ and S^-. In the two- (three) dimensional model, the set of decoys may consist of all conformations (maximally compact conformations) which have more than a certain cutoff number of common contacts with the native state. In three dimensions, all maximally compact conformations with more than 13 common contacts with the native state, and in two dimensions all structures with two or more common contacts with the native state, were usually considered as decoy structures.

The rank-ordered procedure is highly degenerate because there is no selection principle for the assignment of the K strongest interaction energies among the native contacts and the M other interaction energies among the nonnative contacts. The key goal of the structure-based design procedure is the breaking of this degeneracy and a unique assignment of each of the contact energies. We developed a simple hierarchical scheme for doing this. The basic idea is to divide the decoy conformations into groups depending on the number of common contacts they have with the native state. The first step in the design procedure is to ensure that the energies of this group of conformations are increased as much as possible to promote thermodynamic stability. This usually leads to a breaking of some but not all of the degeneracies. Respecting the choices made at this stage, the next group of decoy conformations is considered to further break the degeneracies. This procedure is repeated until a unique assignment of the interaction energies to the contacts becomes possible and the S^+ sequence is obtained. A similar procedure in which the energies of the excited states are lowered as much as possible leads to the S^- sequence.

Operationally, the contacts occurring in the first group have the biggest overlap and are given the highest weight. The contacts occurring in the second group have the second biggest overlap and the weights of each successive group are scaled down by a sufficiently large numerical factor (typically 100) to maintain the hierarchical process. The total weight of a contact comes from summing the contributions from subsequent groups. The number of groups must be sufficiently large to remove degeneracies in the weights of the contacts and the set of the decoy structures must be extensive enough for each contact to receive a non-zero weight. Once all the contacts receive distinct weights, the native and nonnative contacts are sorted according to the assigned weight. In order to generate S^+, we assign values of the native contacts in such a way that the contact with the largest weight (that occurs most significantly among the putative excited states) receives the least attractive B_{ij} and the contact with the smallest weight corresponds to the most attractive native contact. At the same time, the assignment of the nonnative contacts proceeds according to a similar prescription.

Figure 1 (arrows) shows the gaps of the S^+ and S^- sequences obtained in our three-dimensional calculation. As shown, the structure-based design procedure leads roughly to end members in terms of the stability gap. (Monte Carlo sampling in sequence space starting from the S^+ sequence occasionally does lead to larger but quite comparable gap values.) The corresponding values of T_f are shown in the inset of Figure 1. We confirmed that the design procedure works in both two and three dimensions and underscores the importance of the role of nonnative contacts in promoting or decreasing thermodynamic stability. The S^+ sequence may be thought of as corresponding to one with minimal frustration (Bryngelson and Wolynes, 1987).

The design procedure outlined above can also be applied to situations not necessarily generated by rank-ordering of the bonds, in which there is a division of the bonds into those which are allocated to the native contacts and a separate set allocated to the normative contacts. However, one must impose a consistency requirement that the target conformation remains as a ground state of the sequences studied.

Design of Partially Rank-Ordered Sequences

In order to study cases which are not fully rank-ordered (sequences in which the most attractive contact energies are not all assigned to the native contacts), we consider a model in which the strengths of the native contacts (the nonnative contacts are left unaffected) are controlled by a parameter g that reduces their absolute values, originally obtained through the rank ordering, according to

$$B_{ij}(g)\big|_{\text{native}} = B_{ij}(g = 0)\big|_{\text{native}} + g. \qquad (2)$$

Thus, for $g = 0$ the contact energies in the native contacts correspond to the largest attraction available (the fully rank-ordered case) but when g increases in strength, the native contacts become more and more comparable to the normative ones, with the latter remaining unchanged. In particular, there is a critical value of g at which the sequence S^- ceases to have the target conformation as its ground state (Fig. 2). S^+, however, continues to have the target conformation as its ground state until a higher critical value of g is reached. We constructed the end member sequences and studied their properties.

In real proteins, the energy difference between the folded and unfolded states is small (about 5 kcal/mol), resulting in native states which are only marginally stable (Yang et al., 1992). Furthermore, a fully rank-ordered situation seems to be quite exceptional as far as its probability of occurrence is concerned. Furthermore, functional proteins that have emerged from evolutionary processes are not likely to evolve further significantly after they have attained marginal stability and become fully func-

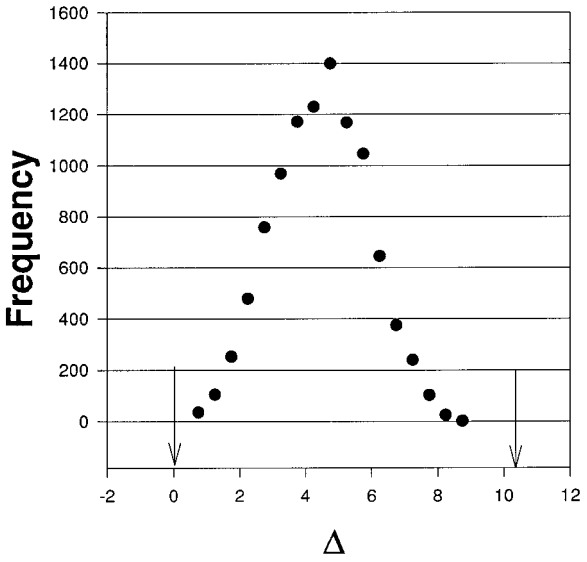

Fig. 2. Histogram of the stability gap (the energy difference between the maximally compact native state and the first excited maximally compact conformation) for the same bare contact energies as in Figure 1 but with $g = g_c$ as derived from a randomly chosen starting sequence followed by 2,000 random swaps of native contact energies and 8,000 random swaps of nonnative contact energies for the three-dimensional random-bond model. Arrows indicate the stability gap of the structure-based designed sequences S^- and S^+.

tional. In the context of lattice models, it is therefore relevant to determine whether naturally evolved sequences can be modeled as ones that have a combination of a bit of rank-ordering and some structure-based design. Our procedure, which involves the g-shift in the contact energies, allows the generation of a whole spectrum of sequences in which the stability gap varies between 0 (for $g = g_c$ and S^-) and a maximum value corresponding to sequence S^+ with $g = 0$. At $g = g_c$, one has the edge of stability. The extreme sensitivity of the stability of sequences on perturbing influences such as temperature, solvent properties, and mutations in the vicinity of such a point may have resulted in the stability edge being a prime candidate for the existence and evolution of naturally occurring proteins.

We turn now to a discussion of the two-dimensional model in which, initially, $\tilde{B}_{ij}$ are taken from the upper right part of Table I in the article by Dinner et al. (1994) (we shall denote this set as DSKS) and the target native state shown in Figure 4a. There are altogether 802075 possible conformations in this model (Lau and Dill, 1989; Dinner et al., 1994), of which 69 are maximally compact. We use all conformations to calculate the partition function and T_f but restrict the number of conformations that are taken into account when designing the assignment of the contact energies.

We consider two sets of decoys: one which involves all conformations (maximally compact and otherwise) which have at least three common contacts with the target native state, and another in which only the maximally compact states are considered. Figure 5 shows T_f for the end member sequences as a function of g for the two sets of decoys. Clearly, the larger set of the decoy conformations gives a much broader spread between the T_f's for S^+ and S^- (this calculation is nearly exact) than when one restricts the decoys to the maximally compact states. The spread in the values of T_f of typical sequences is substantial at all values of g which are less than g_c. At $g = 0$, the range is between 0.954 and 1.278.

TABLE I. Histogram of Maximally Compact Conformations in Three Dimensions Classified by the Number of Closest Siblings and the Number of Common Contacts Between Each of These Siblings and the Original Conformation*

	24	22	21	20	19	18	17
1	46,083	10,087	1,380	552	235	7	0
2	12,662	12,049	289	156	219	23	2
3	703	8,654	15	29	151	24	0
4	8	5,074	0	4	84	19	0
5	0	2,740	0	0	35	2	0
6	0	1,458	0	0	11	7	1
7	0	428	0	0	0	4	0
8	0	127	0	0	0	0	0
9	0	20	0	0	0	0	0
10	0	4	0	0	0	0	0

*The total number of maximally compact conformations for the $N = 27$ model is 103,346. Column 1 represents the number of nearest siblings while row 1 represents the maximum number of common contacts that the nearest siblings have. For example, 46,083 of the 103,346 maximally compact conformations have exactly one closest sibling with 24 common contacts.

Folding Dynamics

While the stability gap and the thermodynamics are controlled significantly by our design procedure, a key feature of real proteins is that they fold rapidly and reproducibly to the native state. We carried out extensive studies of the folding dynamics within the framework of the two-dimensional model. We found that, while the placement of the contacts affects the T_f in a significant way, the dynamics of folding is affected by it to a much lesser extent. This can be assessed by studying the values of the glass temperature, T_g, for S^+ and S^-. A median folding time, as determined by considering the folding of 200 starting conformations, has a U-shape dependence on temperature (Socci and Onuchic, 1994). T_g is defined as a temperature at which the low temperature branch of the U-shaped curve becomes steep and signifies freezing of the kinetic processes. For the two-dimensional model, it is convenient to take T_g to be the temperature at which the median folding time crosses 300,000 Monte Carlo steps (Cieplak et al., 1996; Cieplak and Banavar, 1997). We find that not only the spread between the values of T_g at $g = 0$ is small, between 0.61 and 0.72, but also T_g is essentially insensitive to g, as long as $g < g_c$. Furthermore, for most of the range of the allowed values of g, the sequences remain good folders in the sense that

$T_g < T_f$ (Sasai and Wolynes, 1990; Goldstein et al., 1992; Socci and Onuchic, 1994, 1995; Onuchic et al., 1995). As one may expect, the best folding takes place at full rank ordering.

In order to assess the dependence of the results on the structure of the target native state, we considered 13 different maximally compact conformations as the target native state. When one compares the dynamics of these target conformations with the same set of couplings strengths, one finds variations which are larger than those exhibited by T_f. For instance, at full rank-ordering, i.e., at $g = 0$, and for the most stable sequence S^+, T_g varies between 0.55 and 0.98. Strikingly, for each of the target native states there is little variation in the value of T_g as a function of g. At g_c, the range of T_g remains essentially the same as the $g = 0$ case—it varies between 0.56 and 0.99.

Designability

The designability of a structure is measured by the number of unique amino acid sequences that fold into this structure. It has been argued that, in the evolutionary process, nature selects structures that are more designable over others (Li et al., 1996). The difference in designability of various structures has been demonstrated within the HP model (Li et al., 1996). An analysis of the protein databank and of the sequence databank reveals that the structure of a protein is much better conserved than the amino acid sequence it encodes (Orengo, 1994). It has been suggested that the number of unique three-dimensional folds that can possibly exist is only around 1,000 (Chothia, 1992), thus implying that a huge number of sequences ought to fit into a limited number of three-dimensional folds. These would be the designable structures.

The stability parameter, g_c, can also be interpreted as a measure of the designability of native conformations, since the structures with larger g, (for the same set of contact energies at $g = 0$) are precisely those for which a larger set of sequences is stable (Fig. 3). We suggest, therefore, that nature may select structures with a large stability with respect to the shifts of the contact energies over those in which such stability is small.

We may gain insight into the geometry-based design of proteins by studying numbers of "siblings" of native conformations. The nearest sibling of a conformation is defined as that maximally compact conformation which has the biggest number of common contacts with the conformation. Each conformation can be characterized by two numbers, the number of maximum contacts that the nearest sibling has and the number of such siblings. These characteristics are listed in Table I for the three-dimensional model. For example, there are eight conformations whose nearest sibling has 24 common contacts (there are four such siblings). Likewise, only one conformation has six nearest siblings with each of them having 17 common contacts with it. It would be useful to test the hypothesis that stable conformations are those whose nearest siblings have a relatively smaller number of common contacts as well as having a smaller number of such siblings, i.e., a conformation whose nearest sibling has 17 common contacts and two such siblings might be expected to be more stable than one whose nearest sibling has 17 common contacts and six such siblings, which, in turn, is stabler than a conformation whose closest sibling has 24 common contacts and one such sibling. For example, the highly designable native state structure reported by Li et al. (1996) which accommodates $N_s = 3{,}794$ sequences (Fig. 1 of their article) has two nearest siblings with 18 common contacts with the native state. These observations suggest that an alternative way to pick out highly designable structures may proceed by the exploration of the

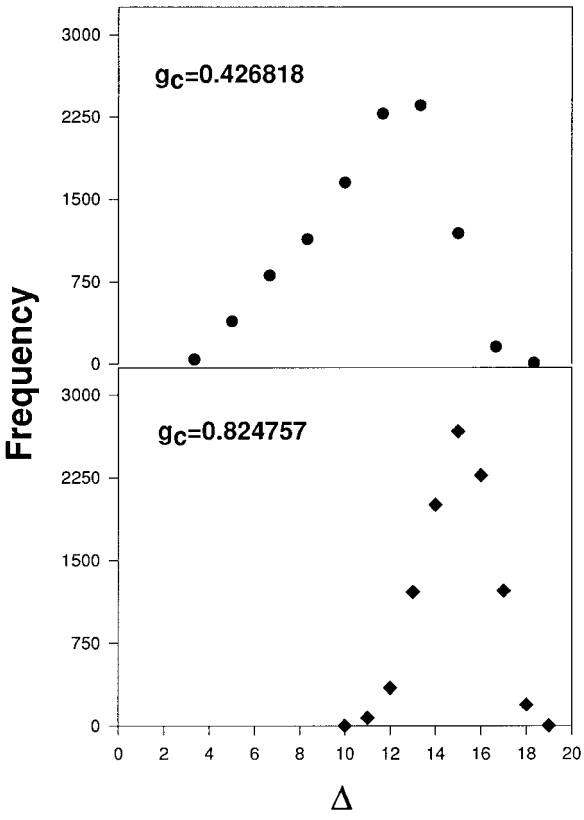

Fig. 3. Histograms of the stability gap (the energy difference between the maximally compact native state and the first excited maximally compact conformation) as derived from two randomly chosen, fully rank-ordered starting sequences followed by 2,000 random swaps of native contact energies and 8,000 random swaps of nonnative contact energies for the three-dimensional, random-bond model. The same set of B_{ij} contact energies were employed in both panels of the figure. The only difference is that two distinct target native state conformations were used with differing values of g_c. The conformation chosen in the lower panel is more designable than the one in the upper panel in that many more sequences exist that could have the lower panel conformation as a native state than the upper one. The higher g_c value in the lower panel nicely correlates with much higher thermodynamic stability as measured by a larger stability gap.

space of conformations and overlaps with the native state without a brute-force enumeration of sequences that design a particular native state.

Local vs. Nonlocal Interactions

The importance of the degree of local interactions vs. nonlocal interactions in a rapidly folding protein has been a subject of great interest. Within the framework of a lattice model of heteropolymers, it has been shown (Unger and Moult, 1996; Abkevich et al., 1995) that native states with predominantly nonlocal contacts show much faster folding kinetics than those with many local contacts. A possible criticism of the above result stems from the fact that there is no simple way of comparing native structures that have quite different ground state energies. All our designed native structures have the same energies, E_0, thus alleviating the problem of different native states having different energies. With our design procedure, we may choose target conformations from among the 103,346 (69) possibilities in three (two) dimensions, which include a sizeable number of short-, medium-, or long-range contacts. Each target native state conformation is characterized by either a locality index, l_I (inspired by Unger and Moult (1996) but with the difference that we have replaced N — $|i - j|$ in the Unger and Moult definition by $|i - j|$) or a structure locality index, l_i. The locality index is defined through

$$l_I = \sum_{i<j} (-B_{ij})|i - j|\Delta_{ij}, \tag{3}$$

where $\Delta_{ij} = 1$ if i and j are monomers which make a contact on the lattice and 0 otherwise. l_I incorporates the coupling strengths and distances along a sequence that are involved in the contacts. In the definition of the structure locality index, B_{ij} is replaced by -1. Thus, l_i provides some characterization of the geometry independent of the actual coupling strengths and is thus somewhat "absolute" in nature. Nevertheless, it assumes that the relevant couplings are attractive. A small value for l_i should indicate a conformation having a large number of short-range contacts, while a large l_i represents a conformation with many long-range contacts.

The target conformation shown in Figure 4a has the maximum number of long-range contacts among the 69 maximally compact states: it has three contacts of the type i, $i + 13$, where i labels consecutive beads along the sequence. The corresponding l_i is equal to 71. On the other hand, the target structure shown in Figure 4b excels in the number of short-range contacts: it has six contacts of the type i, $i + 3$ and $l_i = 49$. In this case, in order to break the degeneracies and deduce a unique rank-ordering it is necessary to include, in the set of decoy states, conformations which have at least two common

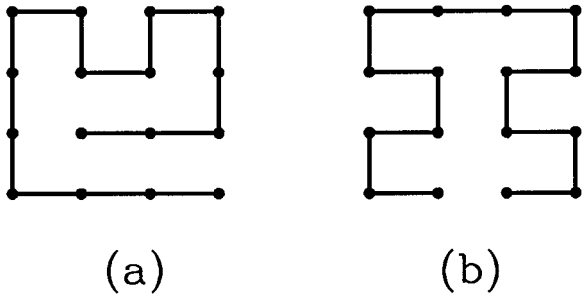

Fig. 4. Two target native conformations for the two-dimensional random-bond model.

contacts with the target state. The resulting g_c is 0.61, which is significantly lower than g_c of 1.09, characterizing the target of Figure 4a with the DSKS couplings. The lower value of g_c signifies a reduced stability of the structure against manipulations of the strengths of the native couplings. The value of T_f corresponding to S^+ at g_c acquires about 68% of its $g = 0$ value of 1.133 for the conformation. This simply reflects different stabilities against the g-shift, even though the $g = 0$ values of T_f are close to each other.

We have found that similar relationships hold for other choices of the couplings strengths than the DSKS. For instance, g_c for the conformation of Figure 4b is always lower (by about 0.5) than for the conformation of Figure 4a. Qualitatively similar conclusions are reached when the decoy structures are the maximally compact conformations.

When considering a set of 11 more target conformations, however, we were unable to correlate the values of T_f or g_c either with the locality index, l_I or with l_i. For instance, the most unstable two-dimensional target, with the lowest value of g_c that we found, corresponded to a conformation with the l_i of 59, i.e., midway between the two values characterizing the conformations shown in Figure 4. We conclude that in the two-dimensional model we see no correlation of the stability either against the g-shift or against temperature with l_I or l_i.

A similar conclusion can be reached for the three-dimensional model. This is demonstrated in Table II, which shows results on T_f and Δ for 12 randomly selected target conformations together with their l_i index. While the data of Table II may suggest an overall increase of g_c with l_i, there are many deviations. Furthermore, this trend appears to be the opposite of that shown by the two-dimensional data for the conformations of Figure 4. Thus, we do not observe any clear influence of the degree of the local vs. nonlocal interactions in the native state on its thermodynamic stability.

RANDOM-SITE MODEL

We now turn to the site-random problem of heteropolymer design within the framework of the two-

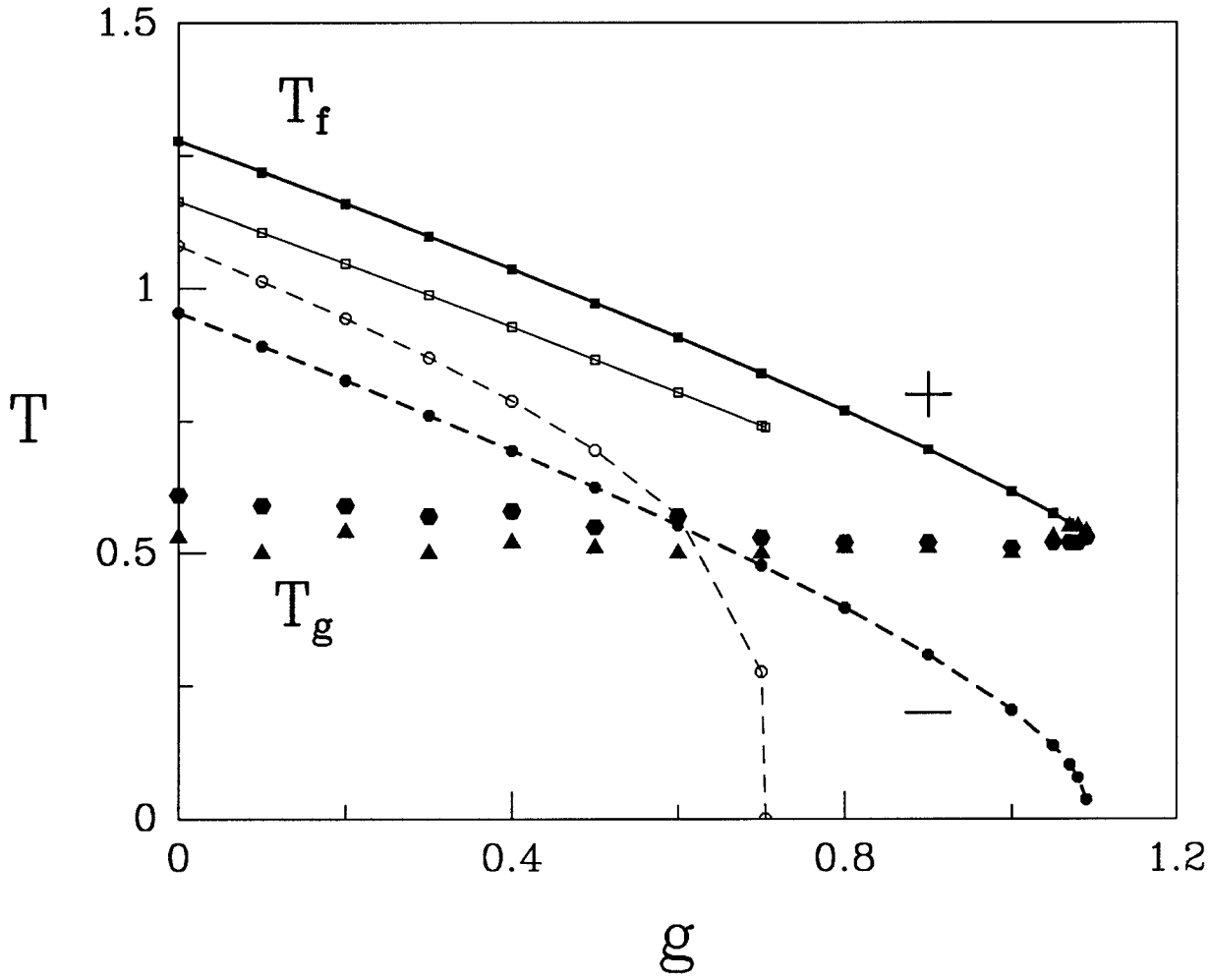

Fig. 5. The dependence of the folding temperature and the glass temperature on parameter g for the two-dimensional conformation shown in Figure 4a. The thick lines connect the data points for T_f obtained by considering the set of decoy structures, which consist of all conformations (not necessarily maximally compact) which have at least three common contacts with the target state. In this case $g_c = 1.09$. The thin lines correspond to the situation in which the decoy states are all 68 maximally compact conformations which are different from the native conformation; $g_c = 0.70$. The solid (broken) lines correspond to the construction of S^+ (S^-). The data points not connected by lines are the values of T_g for S^+ (black hexagons) and S^- (black triangles) as determined from the procedure with the larger set of decoy structures.

dimensional HP model, an extensively studied simple lattice model with only two kinds of amino acids, H and P (Lau and Dill, 1989, 1990; Chan and Dill, 1991). The Hamiltonian is

$$H_s(\Gamma) = \sum_{ij} u(S_i, S_j)\Delta(r_i - r_j). \qquad (4)$$

The i-th amino acid S_i in conformation Γ sits on the lattice site at position $\mathbf{r}_i$. The contact matrix $\Delta(\mathbf{r}_i - \mathbf{r}_j)$ is 1 if $\mathbf{r}_i$ and $\mathbf{r}_j$ are nearest neighbor sites that are not occupied by consecutive amino acids along the chain, and zero otherwise. Finally, $u(H,H) = -\epsilon$ (attractive interaction) whereas $u(H,P) = u(P,P) = 0$.

A conformation Γ is defined to be "good," or encodable, if there is at least one sequence, out of the possible 2^N, that has Γ as its nondegenerate ground state. For the square lattice and $N = 16$, there are 456 good conformations (out of the total of 802,075) and 1,539 sequences with a unique ground state in one of these good conformations—different sequences can have the same good conformation as their native state.

Our design procedure starts with a target structure from one of the 456 good conformations. The key new aspect of our negative design strategy is that in addition to an exploration of sequences, we consider an ensemble of decoy conformations that one hopes are the significant competitors of the target structure in being the native state for the sequences being considered. The procedure consists of considering, for any given sequence, the energy in the target structure and in all the decoy structures. In a

TABLE II. Energy Gap and T_f of the S^+ and S^- Sequences at $g = 0$ and the Critical Values of g for 12 Selected Three-Dimensional Native States*

		S^+		S^-	
l_i	g_c	Δ	T_f	Δ	T_f
232	0.493978	15.380	3.160	1.939	2.089
232	0.426818	14.507	3.174	1.683	2.020
250	0.755568	13.421	3.183	4.517	2.507
252	0.506740	13.147	3.147	2.020	2.007
252	0.676224	9.119	3.080	4.862	2.700
254	0.623939	14.254	3.248	3.736	2.137
274	0.600423	15.903	3.201	3.546	2.339
274	0.428557	16.429	3.222	2.107	1.736
288	0.712084	15.840	3.188	2.447	2.014
302	0.612420	14.835	3.221	2.440	2.014
320	0.656889	16.614	3.296	5.882	2.664
320	0.824757	17.672	3.296	7.329	2.711

*The corresponding values of the structure locality index, l_i are also displayed.

TABLE III. Percentage of Cases in Which the Design Procedure was Successful in the Two-Dimensional HP Model for Various Sets of Decoy Conformations*

Decoy conformations	Percentage of success
Maximally compact	64.47
Good conformations	60.08
$\geq$6 common contacts with target	45.81
$\geq$5 common contacts with target	90.13
$\geq$4 common contacts with target	99.34
$\geq$3 common contacts with target	99.56

*The decoy conformations do not include the target structure itself. An exact enumeration of all conformations and sequences was carried out for this simple model.

screening phase, only those sequences are chosen as candidates for the design process for which the target structure energy is lower than the energy in any of the decoy conformations. If there are many candidate sequences that satisfy this criterion, the tie is broken by requiring that a cost function, defined as the sum over all the decoys of the energy difference between the decoy structure energy and the target structure energy, be maximized. Our choice of the cost function is not unique, but was picked to ensure that the decoy structure energies (viewed as excited states or energies of misfolded states) were as high as possible relative to the native state energy. Unlike real proteins, we note that the HP model is not highly encodable and the overall thermodynamic stability of HP sequences is not very high. Thus, the HP model provides a particularly stringent test of the design procedure.

We carried out the following tests of the new design procedure: For a given set of decoy conformations, we seek to design in each of the 456 good conformations. Using the exact enumeration technique, we then check what percentage of the cases the design procedure is successful in. We repeat this for various sets of decoy conformations. Our results are summarized in Table III.

In our tests, we do not work with just a single composition of the H and P monomers. Unlike the simplified design procedure (Shakhnovich, 1994), we are able to discriminate between differing compositions using our procedure. Also, because we are working with a fixed set of decoy conformations, the calculational scheme is much simpler than the rigorous design process for which the partition function needs to be determined for each sequence.

In a recent work, Deutsch and Kurosky (1996) approximated the free energy associated with Z_s (defined in Eq. 1) for the HP model ($F_s = (-T \log(Z_s))$ to be temperature independent and given by

$$F_s = \sum_{ij} u(S_i, S_j)\langle\Delta(r_i - r_j)\rangle, \qquad (5)$$

where the average is performed over all conformations having seven or more contacts (maximally compact conformations have nine contacts for the $N = 16$ case on the square lattice). It is interesting to note that their approach is an approximate high temperature cumulant expansion of F_s and leads to a 50–70% success rate for the HP model. The average in the above equation may be considered to be over decoy conformations which have seven or more contacts and thus embodies the spirit of the structure-based design strategy we have presented here.

SUMMARY AND CONCLUSIONS

Earlier protein design studies focused on strategies for lowering the energy of the putative native state. In this article, we present a new design strategy which entails not only the study of sequences, but also the conformations which significantly compete with the target structure. (It should be stressed that there are earlier important studies, referred to in the Introduction, that have considered alternative design schemes, some of which do consider the role played by nonnative contacts.) Our approach embodies the idea of negative design in that not only is the native state energy stabilized, but also the low-lying excited states are destabilized, leading to an overall higher thermodynamic stability and a larger stability gap. These low-lying decoy conformations can be obtained either with an exact enumeration, as in the present studies, or by an importance Monte Carlo sampling procedure (Seno et al., 1996). The latter is effective for sampling both maximally compact as well as noncompact conformations in situations in which exact enumeration is impractical.

The design procedures are carried out on the random-bond model in both three and two dimen-

sions (we are limited to studying only the maximally compact conformations in the former case, but the latter allows for an exploration of all conformations and studies of the kinetics of folding) and the random site HP model in two dimensions, each offering unique challenges and providing complementary results.

In the random-bond model, for a maximally compact target conformation with K native and M nonnative rank-ordered contacts, there are $K!\ M!$ sequences with the same native state energy, but yet different thermodynamic stability as measured by the folding transition temperature (T_f) and stability gap (Δ). Our structure-based design strategy allows us to identify the S^+ and S^- sequences which maximally stabilize or destabilize the target native state with respect to competing nonnative structures (decoy structures).

Our detailed studies of the random-bond model for various target structures do not reveal any simple correlation between thermodynamic stability and the degree of local or nonlocal interactions. In order to ensure the maximal stability of the designed native state, one may assign the contact interactions in a rank-ordered way to the native contacts. However, real proteins are only marginally stable and we extended our design strategy for the study of such cases by the introduction of a critical energy shift parameter (g_c) for which the S^- sequence is marginally stable. In the two-dimensional, random-bond model, we have shown that although the thermodynamic stability varies significantly between the S^+ and S^- sequences, the folding dynamics is affected to a much lesser extent.

The question of designability is addressed in connection with g_c and overlap of the candidates for being the low-lying excited states with the target native state. The number of siblings of each maximally compact structure with the highest common contacts is evaluated in three dimensions. The highly designable structures are predicted to be those with high g_c and those with fewer numbers of siblings with a smaller number of common contacts. These observations suggest that highly designable structures can be identified by an exploration of conformational space and their overlaps instead of an enumeration of sequences that design a particular native state.

The random-site model is more realistic because mutations of the amino acids can be effected at given sites. Our studies have been on the HP model which, unlike real proteins, is not highly encodable and provides a particularly stringent test of the design procedure. A high degree of success is achieved for the structure-based design by considering a sufficient number of decoy structures in order to ensure that the native state energy is lower than the energy in these competing conformations.

REFERENCES

Abkevich, V.I., Gutin, A.M., Shakhnovich, E.I. Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. J. Mol. Biol. 252:460–471, 1995.

Bowie, J.U., Luthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170, 1991.

Bryngelson, J.D., Onuchic, J.N., Socci, N.D., Wolynes, P.C. Funnels, pathways and the energy landscape of protein folding: A synthesis. Proteins 21:167–195, 1995.

Bryngelson, J.D., Wolynes, P.G. Spin glasses and the statistical mechanics of protein folding. Proc. Natl. Acad. Sci. USA 84:7524–7528, 1987.

Bryngelson, J.D., Wolynes, P.G. Intermediates and barrier crossing in a random-energy model (with applications to protein folding). J. Phys. Chem. 93:6902–6915, 1989.

Camacho, C.J., Thirumalai, D. Kinetics and thermodynamics of folding in model proteins. Proc. Natl. Acad. Sci. USA 90:6369–6372, 1993.

Chan, H.S., Dill, K.A. Sequence space soup of proteins and copolymers. J. Chem. Phys. 95:3775–3787, 1991.

Chothia, C. Proteins — 1,000 families for the molecular biologist Nature 357:543–544, 1992.

Cieplak, M., Banavar, J.R. Cell dynamics of folding in two dimensional model proteins. Folding Des. 2:235–245, 1997.

Cieplak, M., Vishveshwara, S., Banavar, J.R. Cell dynamics of model proteins. Phys. Rev. Lett. 77:3681–3684, 1996.

Cordes, M.H.J., Davidson, A.R., Sauer, R.T. Sequence space, folding and protein design. Curr. Opin. Struct. Biol. 6:3–10, 1996.

Deutsch, J.M., Kurosky, T. New algorithm for protein design. Phys. Rev. Lett. 76:323–326, 1996.

Dill, K.A., Bromberg, S., Yue, S., Fiebig, K., Yee, K.M., Thomas, P.D., Chan, H.S. Principles of protein folding — A perspective from simple exact models. Protein Sci. 4:561–602, 1995.

Dinner A., Sali A., Karplus M., Shakhnovich E. Phase diagram of a model protein derived by exhaustive enumeration of the conformations. J. Chem. Phys. 101:1444–1451, 1994.

Go, N., Abe, H. Non-interacting local-structure model of folding and unfolding transition in globular proteins. Biopolymers 20:1013–1031, 1981.

Goldstein, R.A., Luthey-Schulten, Z.A., Wolynes, P.G. Optimal protein-folding codes from spin-glass theory. Proc. Natl. Acad. USA 89:4918–4922, 1992.

Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M., Hecht, M.H. Protein design by binary patterning of polar and nonpolar amino-acids. Science 262:1680–1685, 1993.

Kuroda, Y., Nakai, T., Ohkubo, T. Solution structure of a de-novo helical protein by 2D-NMR spectroscopy. J. Mol. Biol. 236:862–868, 1994.

Kurosky, T., Deutsch, J.M. Design of co-polymeric materials. J. Phys. A 28:L387–L393, 1995.

Lau, K.F., Dill, K.A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules 22:3986–3997, 1989.

Lau, K.F., Dill, K.A. Theory for protein mutability and biogenesis. Proc. Natl. Acad. Sci. USA 87:638–642, 1990.

Li, H., Helling, R., Tang, C., Wingreen, N. Emergence of preferred structures in a simple model of protein folding. Science 273:666–669, 1996.

Lombardi, A., Bryson, J.W., DeGrado W.F. De novo design of heterotrimeric coiled coils. Biopolymers 40:495–504, 1996.

Miyazawa S., Jernigan R.L. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 18:534–552, 1985.

Morrissey, M.P., Shakhnovich, E.I. Design of proteins with selected thermal properties. Folding Des. 1:391–405, 1996.

Onuchic, J.N., Wolynes, P.G., Luthey-Schulten, Z., Socci, N. Toward an outline of the topography of a realistic protein folding funnel. Proc. Natl. Acad. Sci. USA 92:3626–3630, 1995.

Orengo, C. Classification of protein folds. Curr. Opin. Struct. Biol. 4:429–440, 1994.

Pabo, C. Molecular technology — Designing proteins and peptides. Nature 301:200, 1983.

Ponder, J.W., Richards, F.M. Tertiary templates for proteins — Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.

Quinn, T.P., Tweedy, N.B., Williams R.W., Richardson, J.S., Richardson, D.C. Betadoublet — De-novo design, synthesis and characterization of a beta-sandwich protein. Proc. Natl. Acad. Sci USA 91:8747–8751, 1994.

Richardson, J.S., Richardson, D.C., Tweedy, N.B., Gernet, K.M., Quinn, T.P., Hecht, M.H., Erickson, B.W., Yan, Y., McClain, R.D., Donlan, M.E., Surles, M.C. Looking at proteins — Representations, folding, packing, and design. Biophysical Society National Lecture, 1992. Biophys. J. 63:1186–2109, 1992.

Sali, A., Shakhnovich, E.I., Karplus, M. Kinetics of protein folding — A lattice model study of the requirements for folding to the native state. J. Mol. Biol. 235:1614–1636, 1994.

Sasai, M., Wolynes, P.G. Molecular theory of associated memory Hamiltonian models. Phys. Rev. Lett. 65:2740–2743, 1990.

Seno, F., Vendruscolo, M., Maritan, A., Banavar, J.R. Optimal protein design procedure. Phys. Rev. Lett. 77:1901–1904, 1996.

Shakhnovich, E.I. Proteins with selected sequences fold to unique native conformation. Phys. Rev. Lett. 72:3907–3910, 1994.

Shakhnovich, E.I., Gutin, A.M. Engineering of stable and fast-folding sequences of model proteins. Proc. Natl. Acad. Sci. USA 90:7195–7199, 1993.

Shrivastava, I., Vishveshwara, S., Cieplak, M., Maritan, A., Banavar, J.R. Lattice model for rapidly folding protein-like heteropolymers. Proc. Natl. Acad. Sci. USA 92:9206–9209, 1995.

Socci N.D., Onuchic, J.N. Folding kinetics of proteinlike heteropolymers. J. Chem. Phys. 101:1519–1528, 1994.

Socci N.D., Onuchic, J.N. Kinetic and thermodynamic analysis of protein-like heteropolymers: Monte-Carlo histogram technique. J. Chem. Phys. 103:4732–4744, 1995.

Sun, S., Brem, R., Chan, H.S., Dill, K.A. Designing amino-acid sequences to fold with good hydrophobic cores. Protein Eng. 8:1205–1213, 1995.

Unger, R., Moult, J. Local interactions dominate folding in a simple protein model. J. Mol. Biol. 259:988–994, 1996.

Yang, A.S., Sharp, K.A., Honig, B. Analysis of heat capacity dependence of protein folding J. Mol. Biol. 227:889–900, 1992.

Yue, K., Dill, K.A. Inverse protein folding problem: Designing polymer sequences. Proc. Natl. Acad. Sci. USA 89:4163–4167, 1992.

Yue, K., Fiebig, K.M., Thomas, P.D., Chan, H.S., Shakhnovich, E.I., Dill, K.A. A test of lattice protein folding algorithms. Proc. Natl. Acad. Sci. USA 92:325–329, 1995.

Vendruscolo M., Maritan A., Banavar, J.R. Stability threshold as a selection principle for protein design. Phys. Rev. Lett. 78:3967–3970, 1997.