Protein Sequence Design by Energy Landscaping

Marcos R. Betancourt*

Laboratory of Computational Genomics, The Donald Danforth Plant Science Center, 893 North Warson Road, Creve Coeur, Missouri 63141

D. Thirumalai

Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742

Received: May 24, 2001; In Final Form: October 2, 2001

We present a novel approach to the protein sequence design problem. Given a target native structure and a target temperature (T_D), our method generates sequences for which the target structure is most likely to be the lowest energy structure. Furthermore, the target structure is rapidly reached at T_D. Thus, the simultaneous *requirements of stability and kinetic accessibility* are satisfied in our design algorithm. The method consists of optimizing a function that captures the energy-landscape features responsible for the native state stability and folding kinetics of protein-like heteropolymers. The efficacy of our method is demonstrated by applications to lattice models (with and without side chains) in which the interaction energies involve pair potentials. The optimization process is computationally efficient. Optimal sequences satisfy the relation $T_D \approx T_\theta$, where T_θ is the collapse transition temperature.

I. Introduction

The problem of protein design consists of finding the sequence (or sequences) of amino acids for a linear polypeptide that self-assembles into a structure with desired functional properties. The solution of this problem is of paramount importance, with applications that include the design of new proteins (de novo designs) and the prescribed modification of existing proteins in order to alter their function, structure, and folding properties.

The problem of *de novo* design can be divided into the selection of a target structure and the determination of the sequence (or sequences) that produces the target structure. Selecting the target structure involves choosing from a pool of large number of protein-like structures that are compatible with the structural constraints (e.g., binding site) needed for a particular application. The problem is complicated by the designability of the structure, as it is possible that no sequence exists for which the selected structure is its ground state. Once a designable target structure is selected, the protein design problem is reduced to finding the sequence that efficiently folds into that structure. This problem is known as the inverse protein folding problem.[1]

One approach to *de novo* design consists of strategic placement of the amino acids in the target structure based on their observed compatibility with known secondary and tertiary structure interactions. The first successful de novo design using this approach was achieved by Reagan and De Grado for a simple 4-helix bundle.[2] An alternate approach (reminiscent of combinatorial chemistry) consists of a "fully automated" search of sequence space, generally done by residue mutations, which is used to evaluate a function (such as the native state stability) whose optimal value generates a sequence with the native state that coincides with the target structure. This is the approach we adopt here. Most automated methods differ in their mutation process, such as different versions of the Metropolis Monte Carlo method,[3−6] genetic algorithms,[7] neural networks,[8,9] and detailed analysis of sequences space.[10,11] They also differ in the choice of scoring functions such as the native state energy gap,[5,12−14] the *Z*-score[15] and the native state probability.[16−20] The energy calculations range from detailed molecular potentials[6] to statistical pair potentials arising from the observed contact frequencies between residues in known protein structures.[7,12−14] Approximations to the denaturated state energy have also been carried out by full enumeration of compact structures,[5,10,16,17] by a Monte Carlo approach,[18] and by threading the sequence into a selected group of compact or known protein structures.[14,16,17] Most design strategies are for a fixed target structure, with some exceptions in which variations of the side chains geometry are allowed.[6] Despite all the effort, only a limited number of *de novo* proteins with "good" properties have been designed. Automated methods are still limited by the inaccuracy of the energy models and the computational cost of calculating the optimization function. Nevertheless, the methods have had some success due in part to the flexibility provided by sequence degeneracy, which allows one to obtain the same structure from many sequences.

We consider two criteria that must be met for a successful *in vitro* design scheme. First the native state must be stable. In most cases, it should be the lowest free energy state consisting of the target structure (nondegenerate), and perhaps a few closely related structures. Second, the folding process should be fast enough to produce significant yield of the folded state in physiological times. Current *de novo* design methods directly consider the native state stability as their design criterion. The kinetic requirements are considered indirectly by the assumption that stable native structures are also fast folders.

In this work, we introduce an optimization (or fitness) function that explicitly considers the kinetic and stability requirements and is constructed from some of the properties of the energy landscape that are responsible for nondegenerate, stable, and fast folding proteins. The optimization of this fitness

10.1021/jp012014c CCC: $22.00
Published on Web 12/21/2001

function leads to fast folding and stable sequences and is relatively easy to compute. Additionally, the resulting sequences are optimized for a particular target temperature. We study the properties of the algorithm for lattice protein models in 3D, with and without side chains. The current version is limited to protein models involving nonbonded contact energies.

We describe the sequence optimization method by giving an overview of the protein energy landscapes, a derivation of the fitness function and a description of the optimization algorithm. Simulations are carried out to test the properties of the designed sequences as a function of the target temperature and the optimization parameter. The robustness of the results is checked by designing sequences for a variety of structures and for lattice models with side chains. We end by discussing the novelty of our results and how they differ from other approaches which are largely based on the maximum stability criteria.[19]

II. Sequence Optimization Method

A. Energy Landscapes. The protein energy landscape consists of the energies of all allowed conformations and the topological ability of one conformation to change into another (its connectivity). Here, we give a brief description of the landscape factors affecting folding based on observations from simple protein models.[21−25] These form the basis of our sequence design method.

The factors that determine the thermodynamics and kinetics of folding, as summarized in a previous work, can be grouped into two categories: a free-energy gradient along the folding pathways and the landscape roughness.[21] The free energy gradient consists of an energy and an entropy gradient and is a function of some suitable "reaction coordinate". The energy gradient is defined as the change in native energy along a folding pathway or a reaction coordinate. Here, native energy means the energy of native contacts. This gradient describes the stabilizing effects of the native contacts as they are being assembled. The entropy gradient, on the other hand, is defined as the change in connectivity entropy along the folding pathways, where the term connectivity entropy refers to the number of available pathways in going from a less folded to a more folded state. In particular, the entropy gradient near the native conformation is related to its geometric accessibility, or the connectivity of the native conformation to its neighbors. This "entropy gap" is important when selecting a target native structure because some structures (such as knotted folds) can be kinetically inaccessible. The other factor that affects folding is the landscape roughness, which is generated by local energy minima, resulting in kinetic traps. It has an energetic part caused by attractive non-native contacts that have to be broken in order for folding to proceed. The landscape roughness can also arise from the connectivity restrictions caused by steric constraints. The geometric (or steric) roughness is caused by native motifs of residues that are relatively oriented in such a way that further folding is strongly restricted, sometimes requiring the unraveling of some of the native motifs. This generates preference in some folding pathways over others.

These factors cannot be independently optimized. Sequence mutations that affect roughness can lead to changes in native stability as well.[26] Therefore, the factors have to be simultaneously optimized to obtain a sequence with the desired landscape properties.

B. Fitness Function. We only consider two-body (nonbonded) contact interactions for which the interactions with and within the solvent are zero. If for a chain with N residues the absolute contact potential is E_{ij}, then the relative contact potential is

$$\epsilon_{ij} = E_{ij} + E_{00} - E_{i0} - E_{0j} \tag{1}$$

where i and j indicate the position of the residues along the chain (1, ..., N), or a solvent molecule (0).

Optimal sequences are obtained by minimizing a heuristic fitness function, $f(T)$, for a specified target structure and a target temperature, $T = T_D$. T_D is the temperature at which the optimal balance between stability and folding rate is desired for the chosen target structure. $f(T)$ is constructed by accounting for the landscape energetic factors that affect folding (i.e., the energy gradient and roughness) and is expressed as

$$f(T) \equiv g(T) - b(T) \tag{2}$$

where $g(T)$ and $b(T)$ are functions that depend on the energy gradient and the energy roughness, respectively. This function is designed by focusing on the space of residue contacts rather than on the conformational space. The basic assumption is that by suitably selecting the native and non-native contact interactions, the native structure stability and kinetics can be optimized.

The only source of energetic roughness arises from non-native contacts. We construct the energy roughness function, $b(T)$, by averaging the Boltzmann factor of each non-native contact over the ensemble of non-native contacts, i.e.,

$$b(T) = -T \ln \langle e^{-\epsilon_{ij}/T} \rangle_{nn} \tag{3}$$

where the average is taken over the ensemble of non-native (nn) contacts. Both T and ϵ_{ij} are in units of $k_B T_o$, where T_o is a reference temperature in Kelvin. The non-native contacts include residue−residue, residue−solvent, and solvent−solvent contributions. The energy roughness function can be written explicitly by separating each contribution in the above order as

$$b(T) = -T \ln \left[\frac{\sum_{i<j}^{nn} P_{ij}(T) e^{-\epsilon_{ij}/T} + \sum_{i=1}^{N} \eta_i \Psi_i + S_{nn}}{\sum_{i<j}^{nn} P_{ij}(T) + \sum_{i=1}^{N} \eta_i \Psi_i + S_{nn}} \right] \tag{4}$$

where the denominator is a normalization factor. Notice that in the ideal case in which all the non-native energies ϵ_{ij} are zero (such as for the Gō model[27]), $b(T) = 0$.

The summation involving residue−residue interactions contains the function $P_{ij}(T)$, which is proportional to the probability of contact formation between monomers i and j. To obtain $P_{ij}(T)$, we use the quasi-chemical approximation, i.e.,

$$P_{ij}(T) = P_{ij}^{o} e^{-\epsilon_{ij}/T} \tag{5}$$

where P_{ij}^{o} is the equilibrium probability of forming a contact for $\epsilon_{ij} = 0$ (infinite temperature limit). The residue−solvent summation involves the non-native residue−solvent contact probability at residue i, Ψ_i, times the number of native residue−residue contacts at i, η_i, summed over all residues; Ψ_i is defined as

$$\Psi_i = \frac{\langle \max(\eta_i - c_i, 0) \rangle}{\eta_i} \tag{6}$$

where c_i is the total number of residue−residue contacts of residue i in a given conformation, and the average is taken over

the ensemble of all conformations that are equally weighted. Notice that the maximum number of non-native residue−solvent contacts is η_i. If in a given conformation c_i is smaller than η_i, then residue i has more residue−solvent contacts than in the native state and these additional residue−solvent contacts are considered non-native. The solvent−solvent term is defined as

$$S_{nn} = \langle \max(C - C_n, 0) \rangle \tag{7}$$

where $C = {}^1/_2\sum_{i=1}^{N} c_i$ is the total number of contacts in a particular conformation and $C_n = {}^1/_2\sum_{i=1}^{N}\eta_i$ is the number of native contacts. For maximally compact native conformations $S_{nn} = 0$, and for compact ones it is approximately zero. Both terms involving the solvent are constant independent of sequence because $\epsilon_{00} = 0$ and $\epsilon_{i0} = 0$ for all i.

The energy gradient function, $g(T)$, which is a measure of the "free" energy of native contacts, is

$$g(T) = T \ln\langle e^{\epsilon_{ij}/T}\rangle_{\text{nat}} \tag{8}$$

with the average being taken over the ensemble of native (nat) contacts. A difference between $g(T)$ and $b(T)$ is that in $g(T)$ the signs in front of the logarithm and in the exponential are positive. The result is that the average in $g(T)$ gives a larger weight to high-energy contacts in contrast to $b(T)$, which gives a higher weight to low energy contacts. The explicit form of $g(T)$ (separating the solvent contributions) is

$$g(T) = T \ln\left[\frac{\sum_{i<j}^{\text{nat}} \hat{P}_{ij}(T) e^{\epsilon_{ij}/T} + \sum_{i=1}^{N}\omega_i\hat{\Theta}_i + \hat{\Phi}C_n}{\sum_{i<j}^{\text{nat}} \hat{P}_{ij}(T) + \sum_{i=1}^{N}\omega_i\hat{\Theta}_i + \hat{\Phi}C_n}\right] \tag{9}$$

The first term in the numerator of eq 9, which corresponds to the native residue−residue contacts, is weighted by the probability $\hat{P}_{ij}(T)$, which is defined as

$$\hat{P}_{ij}(T) = 1 - \frac{P_{ij}^{\text{o}} e^{-\epsilon_{ij}/T}}{(1 - P_{ij}^{\text{o}}) + P_{ij}^{\text{o}} e^{-\epsilon_{ij}/T}} \tag{10}$$

This function modifies the energy gradient by stabilizing them according to their formation probability. The first term on the right side represents the probability of finding a native contact (i, j) in the native conformation (i.e., 1). The second term is proportional to the probability of finding a native contact (i, j) in the unfolded conformations (i.e., $P_{ij}^{\text{o}}e^{-\epsilon_{ij}/T}$). This term is normalized so that it is P_{ij}^{o} as $\epsilon_{ij}/T \rightarrow 0$ and 1 when $\epsilon_{ij}/T \rightarrow -\infty$. The latter condition implies that for $\epsilon_{ij} < 0$, the probability of finding the native contact (i, j) among all states approaches 1 as T goes to zero.

The definition given by eq 10 has the following effects. The smaller the probability P_{ij}^{o}, the larger the function $\hat{P}_{ij}(T)$, and their contribution to the average in eq 9. This has consequences on the stability and uniqueness of the native state because the design priority is given to native contacts that are more unique to the native state and less likely to be in other non-native conformations. This argument is in agreement with calculations of the intrachain loop probability.[28] There are also kinetic consequences that follow from eq 10. When the formation of a native contact with low probability P_{ij}^{o} is essential for folding, the premature formation of other native contacts with higher P_{ij}^{o} could lead to geometric roughness effects.[26] This could arise even in the Gō model. By giving higher priority to the low probability native contacts, the effects of such kinetic traps are diminished.

The residue−solvent summation involves $\hat{\Theta}_i$, which is defined as the probability that a native residue−solvent contact at residue i is formed minus a contribution from the unfolded states. Note that a residue−solvent contact is defined as native when the total number of solvent contacts with that particular residue is equal to or less than the number found in the native state. Explicitly,

$$\hat{\Theta}_i = 1 - \frac{\Theta_i e^{-\epsilon_{i0}/T}}{(1 - \Theta_i) + \Theta_i e^{-\epsilon_{i0}/T}} = 1 - \Theta_i \tag{11}$$

where

$$\Theta_i = 1 - \frac{\langle \max(c_i - \eta_i, 0)\rangle}{\omega_i} \tag{12}$$

is the random ensemble probability that a native residue−solvent contact at residue i is formed and ω_i is the number of native residue−solvent contacts at i. Note that if $z_i = \eta_i + \omega_i$ is the available contact number of residue i, the native and non-native residue−solvent probabilities are related by

$$\frac{\sum_{j=1}^{N} P_{ij}^{\text{o}}}{z_i} = 1 - \left(\frac{\omega_i}{z_i}\Theta_i + \frac{\eta_i}{z_i}\Psi_i\right) \tag{13}$$

which is the total probability that a particular residue−residue contact in i is formed. $\hat{\Phi}$ is the probability that a native solvent−solvent contact is formed in the native state minus the contribution from the unfolded states. The maximum number of native solvent−solvent contacts is equal to the number of native contacts, i.e., solvent−solvent contacts are native when the protein is in its native state. Therefore,

$$\hat{\Phi} = 1 - \frac{\Phi e^{-\epsilon_{00}/T}}{(1 - \Phi) + \Phi e^{-\epsilon_{00}/T}} = 1 - \Phi \tag{14}$$

where

$$\Phi = 1 - \frac{\langle \max(C_n - C, 0)\rangle}{C_n} \tag{15}$$

is the random ensemble probability that a native solvent−solvent contact is formed. The solvent−solvent terms satisfy the relation

$$\sum_{i<j}^{N} P_{ij}^{\text{o}} = C_n\Phi + S_{nn} \tag{16}$$

which gives $\Phi = \sum_{i<j}^{N} P_{ij}^{\text{o}} C_n$ when $S_{nn} = 0$.

The difference between the weighting factors in $g(T)$ (eq 9) and $b(T)$ (eq 4) originates from the contributions of the native and non-native contacts to the native and non-native states. The function $b(T)$ accounts for the probability of making non-native contacts in the non-native states. Since native contacts are present in both the native and non-native states, the native contact contribution to the non-native states is subtracted in $g(T)$. As a result, we can think of $b(T)$ as containing the net effect of the non-native contacts in the non-native states and $g(T)$ as containing the net effect of the native contacts in the native state. The solvation contribution to $g(T)$ and $b(T)$ prevents

the native residue−residue contact energies from becoming positive and dissolved by non-native solvent contacts. It also discourages the non-native residue−residue contact energies from becoming much lower than the native solvent contact energies. The minimization of eq 2 should produce a sequence with the optimal balance between stability and folding kinetics at target temperature $T = T_D$. That is, at high T_D it is more important to have interactions that stabilize the native state despite a rougher landscape, while at low T_D the opposite is true.

For the simplified case in which the fluctuations of the native and non-native contact energies are neglected, the low and high temperature limits of eq 2 can be readily evaluated. In particular, let $\epsilon_{ij} = -\epsilon_g \leq 0$ for the native contacts and $\epsilon_{ij} = \epsilon_b \geq 0$ for the non-native ones. This is just a generalization of the Gō model. At low temperatures, the fitness function is approximated by

$$f(T) \simeq -T\frac{\sum_{i<j}^{\text{nat}} \hat{P}_{ij}^{o}/P_{ij}^{o}}{\sum_{i=1}^{N}\omega_i\hat{\Theta}_i + \hat{\Phi}C_n}e^{-\epsilon_g/T} - T\frac{\sum_{i<j}^{\text{n.n.}} P_{ij}^{o}}{\sum_{i=1}^{N}\eta_i\Psi_i + S_{nn}}e^{-\epsilon_b/T} \tag{17}$$

where $\hat{P}_{ij}^{o} = 1 - P_{ij}^{o}$. This equation is valid for $-\epsilon_g/T \ll \ln P_{ij}^{o} < 0$ and $\epsilon_b/T \gg 0$ and shows that the values of ϵ_g and ϵ_b that minimize $f(T)$ rapidly approaches zero as $T \rightarrow 0$. The effect of lowering the temperature is to increase the effective roughness of the energy landscape. Therefore, the vanishing values of ϵ_g and ϵ_b counteract by maintaining the landscape as smooth as possible without sacrificing stability, which should be maintained by the condition $-\epsilon_g/T \ll 0$. At high temperatures,

$$f(T) \simeq -\frac{\sum_{i<j}^{\text{nat}} \hat{P}_{ij}^{o}}{\sum_{i<j}^{\text{nat}} \hat{P}_{ij}^{o} + \sum_{i=1}^{N}\omega_i\hat{\Theta}_i + \hat{\Phi}C_n}\epsilon_g - \frac{\sum_{i<j}^{\text{nn}} P_{ij}^{o}}{\sum_{i<j}^{\text{n.n.}} P_{ij}^{o} + \sum_{i=1}^{N}\eta_i\Psi_i + S_{\text{nn}}}\epsilon_b \tag{18}$$

is valid for $\epsilon_g/T \ll 1$ and $\epsilon_b/T \ll 1$. In this limit, $f(T)$ is minimized by maximizing ϵ_g and ϵ_b, i.e., maximizing the native state stability. The landscape roughness has little effect on the kinetics at these temperatures.

In the preceding equations, the minimal $f(T)$ value depends on several sequence independent factors. These factors are correlated, as can be shown in part by noticing that

$$\sum_{i<j}^{\text{nat}} \hat{P}_{ij}^{o} + \sum_{i=1}^{N}\omega_i\hat{\Theta}_i + \hat{\Phi}C_n = \sum_{i<j}^{\text{nn}} P_{ij}^{o} + \sum_{i=1}^{N}\eta_i\Psi_i + S_{\text{nn}} \tag{19}$$

which follows from eq 4 and eq 9. A parameter that describes the geometrical design fitness of a structure may be defined as

$$s \equiv -\ln\left(\frac{\sum_{i<j}^{\text{nn}} P_{ij}^{o}}{\sum_{i<j}^{\text{nn}} P_{ij}^{o} + \sum_{i=1}^{N}\eta_i\Psi_i + S_{\text{nn}}}\right) \tag{20}$$

which depends only on the relative number and geometric arrangements of native contacts. Compact target structures with lower random probabilities (P_{ij}^{o}) of making native contacts have smaller s values. Typically, smaller s values yield sequences with better folding properties.

C. Optimization Algorithm. The optimization procedure is carried out by Monte Carlo simulations in sequence space. A slight variation of the simulated annealing method[4] is used. An arbitrarily selected sequence is subject to random mutations, and $f(T)$ is calculated before and after the mutation. The new sequence is accepted with probability

$$\rho = e^{-\max[\Delta f(T),0]/\theta} \tag{21}$$

where θ is the optimization temperature. Initially, a group of random sequences are optimized at a high θ until a stable minimal value for $f(T)$ is reached for each sequence. Then, the temperature θ is lowered by a small increment and the sequences are optimized once again. The starting sequence at the new temperature is the optimal sequence at the previous high temperature. This process is repeated until $f(T)$ converges to a minimum value.

To test our sequence design method, we utilize a single bead representation of the polypeptide chain residues. The beads are constrained to a cubic lattice in 3-D. Nonbonded residues interact when they are separated by a single lattice space (a contact). While the choice of the protein contact potential ϵ_{ij} is crucial in the protein design problem, we are more concerned with the principles of the design process itself so that the precise potential used is not critical. However, the ability of the algorithm to design sequences with the desired properties will depend on the variety of interactions and the amino acid "alphabet" size.

The residue interactions are modeled by a knowledge-based pair-potential described elsewhere.[29] It represents the effective energies between the 20 common amino acids. Here, we scale this potential by a factor of 7.75 to increase the correspondence to the experimental hydrophobicities.[30]

The probabilities P_{ij}^{o}, Ψ_i, Θ_i, and Φ are computed by performing Monte Carlo simulations for an N residue chain after setting $\epsilon_{ij} = 0$. We use either the Rosenbluth scheme[31] or full enumeration whenever possible. Although there are some theoretical estimates for P_{ij}^{o}, such as the ones given by Jacobson and Stockmayer[32] ($P_{ij}^{o} \propto [d/2\pi|i-j|]^{d/2}$ in d dimensions) or by others,[28,33] the Monte Carlo calculation is preferred because it considers all lattice and chain end effects. The solvent contact probability Ψ_i is a function of the native state, therefore, to avoid calculating it for every native conformation, the quantity $\langle\max(\eta - c_i, 0)\rangle$ is calculated for all values of η. Θ_i can be calculated in a similar manner or from eq 13. Given these parameters and the target structure, $f(T)$ can be readily calculated.

III. Simulation Results

A. 15-mer. We first consider chains with $N = 15$ so that all conformations can be examined in a reasonable time, allowing

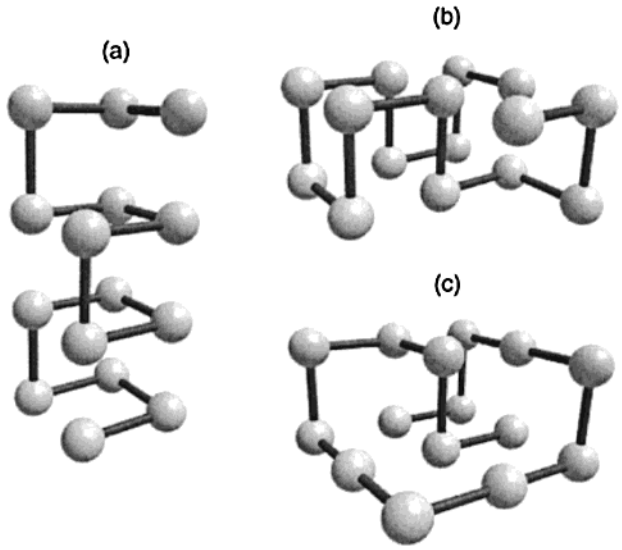

Figure 1. Examples of three compact conformations for $N = 15$. Structure (a) has $C_n = 11$ contacts and $s = 2.75$. This is the conformation with the highest s. Notice that this is the square lattice version of a helix where the $(i, i + 3)$ contacts are favored. (b) $C_n = 11$ contacts and $s = 2.58$. (c) $C_n = 10$ contacts and $s = 2.29$. This is the conformation with the lowest s among all compact conformations with 10 or 11 contacts. Even though this conformation appears more compact, it has no $(i, i + 3)$ interactions and less contacts than (a).

TABLE 1: Designed Sequences for a Range of Target Temperatures T_D[a]

no.	sequence	T_D	T_θ	$T_{1/2}$	E_{ns}	g_{ns}
1	APSHNYRDNQQKDRC	0.6	0.72	0.49	-13.33	1
2	VFSHGYKGGQQGDKH	0.8	0.99	0.73	-18.60	1
3	DPSAGHKGGQHGIFF	1.0	1.13	0.64	-20.15	1
4	RASCHYQHDKKHAQI	1.2	1.45	0.98	-26.89	1
5	RASIHYQHDKKHAQI	1.4−1.6	1.57	0.98	-28.52	1
6	MPEKKMGDKGVDDAY	1.8	2.32	1.48	-44.02	1
7	MPQRKIGDKGIDDAW	2.0−2.2	2.53	1.77	-48.98	1
8	MPQRKIGDKGVDDAW	2.4	2.57	1.78	-49.83	1
9	HMSLDWKKDDPKAWF	2.6−2.8	3.01	2.09	-58.13	2
10	HMSLEWKRDDPKVWF	3.0	3.34	2.04	-63.32	2
11	HMSLEWKKDEPKVWF	3.2−3.4	3.46	2.06	-65.64	2
12	PMSLEWKKDEPKVWF	3.6	3.47	2.44	-67.74	2
13	PCSCEWKKDEMKVFC	3.8−4.0	3.85	2.39	-75.87	1

[a] T_θ is the energy collapse temperature, $T_{1/2}$ is the folding temperature for which the probability of being in the native state is $^1/_2$, E_{ns} is the native state energy, and g_{ns} is the native state degeneracy. The native structure for all these sequences is the one shown in Figure 1b.

exact thermodynamic calculations.[25] The target structures are selected from the set of all compact structures with $C_n = 10$ and $C_n = 11$ native contacts. The structural factor s defined by eq 20 is near minimal for these structures, making them ideal candidates from a design perspective. Among the compact structures, the two conformations with the largest and smallest s values are shown in Figure 1a,c, respectively.

Consider the dependence of the designed sequences on the target temperature. For this we use the maximally compact target structure (Figure 1b) with a typical s value. The minimization of $f(T)$ was carried out at 18 different target temperatures ranging from $T_D = 0.6$ to $T_D = 4.0$ and leading to 13 different sequences, which are listed in Table 1. The sequences start from the nearest end residue in Figure 1b. Also shown are the collapse temperatures, T_θ, as defined by the maximum in the specific heat,[24] the folding temperature, $T_{1/2}$, for which the probability of being in the native state is $^1/_2$, the native state energy, E_{ns}, and the native state degeneracy, g_{ns}. All the native states produced by

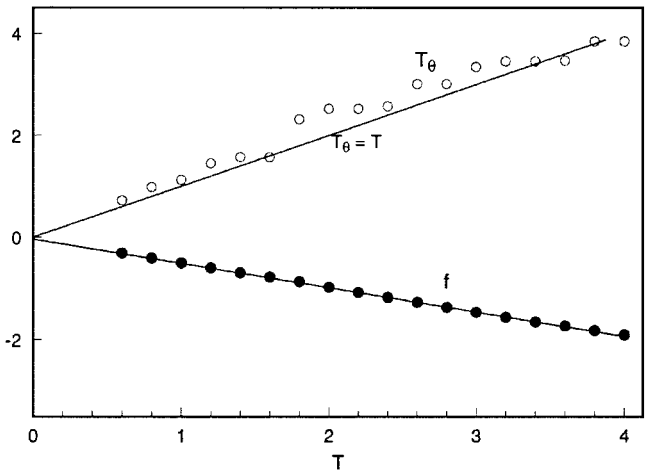

Figure 2. Optimal fitness function values for the sequences obtained at several target temperatures. All sequences were designed for the same native structure. f and T_D are linearly correlated, with a least-squares fit of $-0.47T_D$ and a correlation coefficient of $r \simeq 1$. Also shown is the correlation between T_θ and T_D. The linear fit is $T_\theta \simeq 0.33 + 0.94T_D$, with $r = 0.943$.

the sequence design method correspond to the target structures, and most are nondegenerate (nine are nondegenerate and four are doubly degenerate). The collapse and folding temperatures generally increase as the target temperature is increased, maintaining optimal fitness conditions. In particular, T_θ and T_D are well correlated, with T_θ near and usually above T_D. Also note that the sequences designed at high temperatures ($T_D = 2.6-4.0$) contain stronger interacting residues, while these are less frequent at lower temperatures ($T_D = 0.6-0.8$). This is also evident by the decrease in native energy as the temperature increases.

The values of $f_{min} = \min f(T_D)$ (Figure 2) show a remarkably strong correlation to T_D, i.e., $f \simeq -0.47T_D$ (correlation coefficient of $r \simeq 1$). This relation is satisfied for optimal sequences (those that minimize f), and the slope depends on the native structure. We expect deviations from linearity at very high and low temperatures, where optimal sequences cannot be obtained. Figure 2 also shows the correlation between T_θ and T_D. Not only is this correlation good, but also the linear-fit slope is near unity. The latter result is strongly dependent on the solvation terms. If the solvation terms are neglected from $f(T)$, the T_θ and T_D correlation is still reasonably good, but the linear-fit slope largely deviates from unity. Therefore, the solvation terms are important in maintaining a delicate balance between stability and the folding kinetics at a given T_D.

The thermodynamic stability of the designed sequences are obtained from the native state probability as a function of temperature, $P_{ns}(T)$. Figure 3 shows $P_{ns}(T)$ for each of the sequences in Table 1. Above $T > 4$, the optimal probability rapidly drops, indicating that there is a high-temperature limit for which good (stable) sequences can be obtained. As the target temperature decreases, sequences with high $P_{ns}(T_D)$ become easier to generate. The folding transition becomes sharper for sequences designed at lower temperatures (Figure 3). In Figure 4, we show the kinetic effects of the design method by plotting the folding time of each sequence at and around their target temperature. The folding time, or τ_f, is defined by the mean first passage time to the native state, averaged over a large number (typically 400) of random initial conformations. It is calculated using a dynamic Monte Carlo method,[26] which is summarized in Appendix A. Time is measured in units of

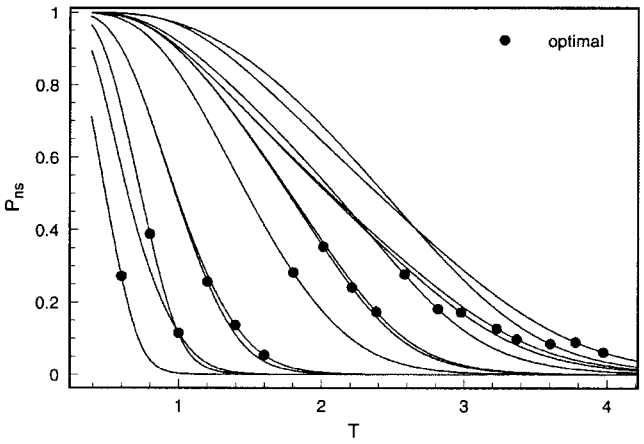

Figure 3. Native state stability for sequences designed at different temperatures. Each line shows the native state probability for each sequence, while the black circles correspond to the values for the optimal sequence at their target temperatures.

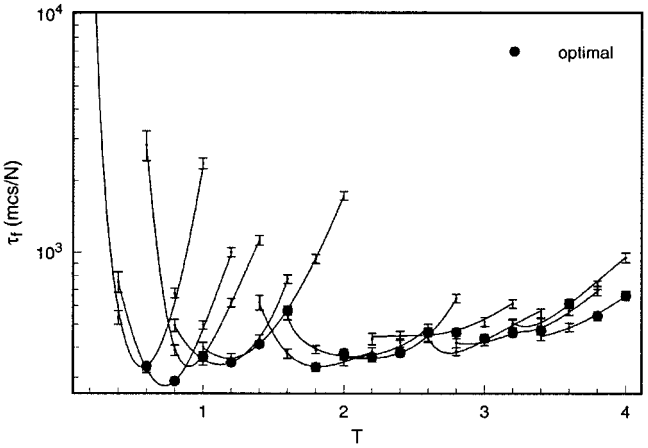

Figure 4. Folding time for designed sequences at different temperatures. Each line shows the mean folding time for each sequence, while the black circles correspond to the values for the optimal sequence at their target temperatures.

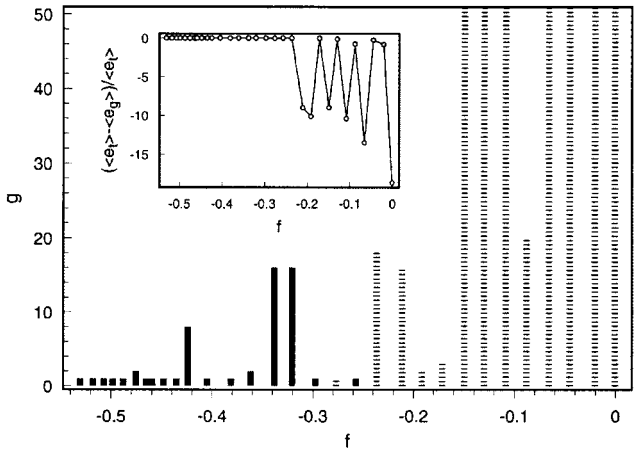

Figure 5. Degeneracy g_{ns} for a group of sequences with negative f values. Black bars indicate sequences that folded to the correct native structure, while dashed bars indicate those that failed. The inset picture shows the relative difference in energy between the target native state and the actual ground state. $T_D = 1.06$.

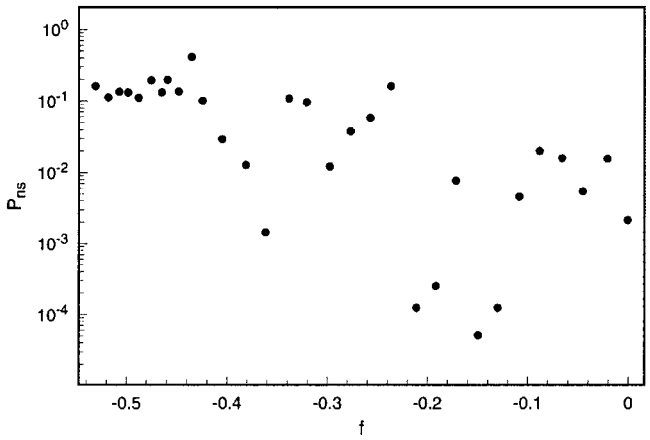

Figure 6. Correlation between the native state stability and f for the same sequences as in Figure 5. The native state probability is calculated at $T_D = 1.06$.

Monte Carlo steps per sequence length (mcs/N). As the target temperature changes, the selected sequences retain the smallest values of τ_f. As the temperature decreases and the folding time increases, a faster folding sequence with a somewhat lower stability is replaced. It seems that a lower bound for designing good (fast folding) sequences cannot be too far below $T_D \approx 0.5$ (Figure 4). For sequences designed at relatively high temperatures, the folding time remains optimal over a wider temperature range, which is in accord with the findings of Abkevich et al.[34]

We have selected as the optimal sequence the one with the lowest f at a given target structure and temperature. However, we expect that a set of sequences with low f values should also be adequate sequences. This set corresponds to the sequence degeneracy group. Sequence degeneracy is examined by looking at sequences with various f values obtained during the optimization process for the target structure shown in Figure 1b at $T_D = 1.06$. In Figure 5, the ground-state degeneracy is plotted for negative f values. For a significant range of low f values (-0.53, -0.25) there are many sequences that have this particular target structure, or a very similar structure, as their unique ground state. The size of this degeneracy group depends on the target structure and on the variety of the interaction model.

The native state probabilities, P_{ns}, calculated at $T_D = 1.06$, for the sequences used to generate the plot in Figure 5, are shown in Figure 6. The probability correlates in general with f, showing a destabilizing tendency for the target structure as f increases. The degree of correlation can increase with target temperature because a higher importance is given to the minimization of $b(T_D)$ as T_D increases. Note that most of the optimal sequences are not hyper stable at the target temperature. This is because maximal stability does not generally imply fast folding, and our algorithm is designed to balance stability and kinetics.

The correlation of the folding time τ_f with f is shown in Figure 7. Sequences with low f ($f \lesssim -0.25$) are also fast folding, with similar folding times near $\tau_f \simeq 10^3$ mcs/N. Above these values, and as the ground state structures begin to differ significantly from the target native structure, the folding time increases by many orders of magnitude. This and the previous two plots show that the present sequence-design method leads to fast folding, stable and low degenerate native states.

We tested our results by optimizing sequences for a diverse group of compact structures (with 10 and 11 contacts). A total of 50 structures were selected with the condition that each set of native contacts correspond to a single structure. After optimizing the sequences at $T_D = 1.06$ and $T_D = 1.94$ it was found that all of them led to the correct native structure. The average thermodynamic and kinetic properties are summarized in Table 2.

TABLE 2: Average Folding Properties of Designed Sequences for 15-mers and 36-mers[a]

N	T_D	$\langle g_{ns}\rangle$	$\langle T_\theta\rangle \pm \sigma_T$	$\langle T_{1/2}\rangle \pm \sigma_T$	$\langle P_{ns}\rangle \pm \sigma_p$	$\langle t_f\rangle \pm \sigma_\tau$
15	1.06	1.04	1.13 ± 0.12	0.84 ± 0.16	0.25 ± 0.17	$(7.27 \pm 3.00) \times 10^2$
15	1.94	1.00	2.17 ± 0.24	1.61 ± 0.40	0.32 ± 0.20	$(7.00 \pm 2.53) \times 10^2$
36	1.06	1.05	1.30 ± 0.07	1.17 ± 0.10	0.67 ± 0.15	$(6.74 \pm 19.2) \times 10^5$
36	1.94	1.00	2.25 ± 0.14	2.04 ± 0.19	0.64 ± 0.19	$(2.29 \pm 1.26) \times 10^4$

[a] The results were obtained at two target temperatures. For the 15-mer and the 36-mer, 50 and 20 different structures were used as target structures. The ensemble average is denoted by the brackets and their respective standard deviation by σ. The collapse temperature, T_θ, is obtained from the maximum in the specific heat. The native state probabilities P_{ns} and the folding times τ_f are calculated at $T_D = 1.94$. The folding time is given in Monte Carlo steps divided by the sequence length (mcs/N).

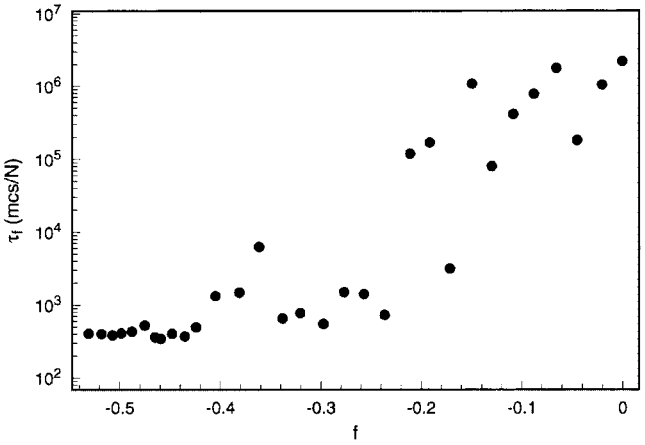

Figure 7. Correlation of folding time with f for the same sequences as in Figure 5. All folding times are calculated at $T_D = 1.06$.

B. 36-mers. The chain length dependence of our results is examined by using a 36-mer. We have randomly selected 20 different maximally compact structures (an example of which is shown in Figure 8a) as target native structures. For 36-mers we are unable to fully enumerate all the conformations. Therefore, restricting to maximally compact conformations increases the chances that no other structure has the same set of contacts. To obtain the minimum energy structure generated by the 20 chains, we have performed a Monte Carlo simulation in conformation space using simulated annealing.[4] The native state stability was calculated using an implementation (see the AppendixB for details) of the multiple histogram method.[35]

The average thermodynamic and kinetic properties for the 36-mers are also shown in Table 2. We designed sequences at $T_D = 1.06$ and $T_D = 1.94$. As for the 15-mer sequences, all the 36-mer sequences yield the correct native structure. Also, all native states are unique for the sequences designed at $T_D = 1.94$, and only one was doubly degenerate for $T_D = 1.06$. The average values of T_θ are also close and slightly above the target temperature with very similar values, as shown by the relatively small dispersion. The results for the 36-mer differ from the 15-mer in the values for the stability and folding times. At the target temperature, the 36-mer sequences produce native states with a higher stability than the 15-mers. T_θ is similar but slightly higher than in the 15-mer case, which when combined with a sharp transition, explains a rapid increase in stability. The folding time is near optimal for the sequences designed at $T_D = 1.94$ and for most of the sequences at $T_D = 1.06$ with values as low as 8.6×10^3 mcs/N at $T_D = 1.94$. At $T_D = 1.06$, two sequences produce slower folding times and are responsible for the higher average value and the large dispersion. Without considering those two, the average folding time is $\tau_f = (6 \pm 2) \times 10^4$ mcs/N. The sharpness of the transitions at $T_D = 1.06$ creates a narrow temperature region for generating fast folding times. For temperatures outside the optimal range, the folding times increase rapidly.

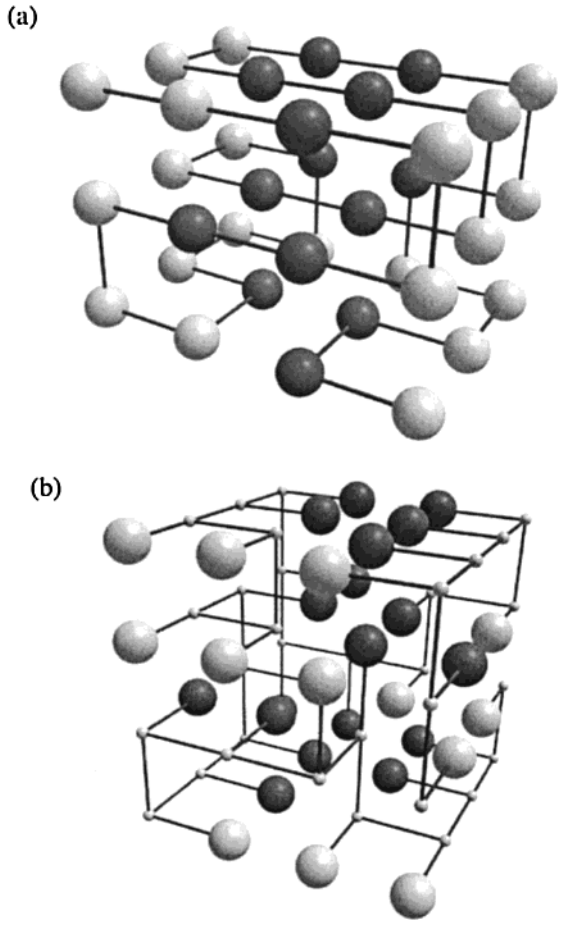

Figure 8. Examples of target conformations for the 36-mer and the 32-mer with side chain. The light and dark residues represent hydrophilic and hydrophobic residues, respectively. (a) 36-mer without side chain. The sequence obtained at a target temperature of $T_D = 1.06$ is G P V K Q A C G G P H G G D H G G H L K Q A C G G H A K N Y R K N Y C K starting from the top-left end residue. (b) 32-mer with side chains. The backbone beads are reduced in size for clarity. The sequence for the side chain beads at $T_D = 1.94$ is R Y D H H H K Y K I A H H H Q K D K H Y M A G A I R Y G G Y I E starting from the front-end residue. We have used a one-letter code for amino acids.

We argued that the minimal value of the fitness function depends on the structural parameter s given by eq 20. This can be demonstrated from the correlation between the optimal f values and s. By combining the results of all 15-mer and 36-mer sequences, designed at previously mentioned temperatures, we obtained an approximate linear fit of $f/T_D \simeq -1.10 + 0.25s$ with a correlation coefficient of $r = -0.9$ and s values ranging from 2.3 to 2.8. By comparing s to the native state stability calculated at T_D, we observe a slight tendency for structures with lower s values to be more stable. This is in general agreement with our previous arguments.

TABLE 3: Designed Sequences for 32-mers with Side Chains

no.	sequence	E_{ns}	T_θ	T_f	T'_D	τ_f
1	EPYIAVCYPGGHDGGYPDNHHRHMCHGGRYAV	−147.41	1.15	0.74	1.00	1.62×10^7
2	RYDHHHKYKIAHHHQKDKHYMAGAIRYGGYIE	−172.60	1.27	0.91	1.08	4.22×10^7
3	YHHHREFHIHYPCPDKDAIPCDHGQKGAHYKA	−154.54	1.13	0.88	0.93	8.99×10^7
4	EPYPYHDNGALHHGGGMHKDYKHHAAHDIICG	−150.12	1.11	0.94	1.03	5.86×10^7
5	PYIHVYCDNAHHHRVKNYPALGDGAHHAGIPY	−150.27	1.14	0.84	0.93	7.50×10^7
6	DRRVHIGGGGMMVKGGDHGGGDKHVMGGGYCD	−170.81	1.41	1.06	1.00	2.06×10^8
7	MDHIGGAIAGHHQCGKHHGCKHDYPKDCPFAV	−147.41	1.05	0.94	0.91	4.26×10^7
8	HCIAYKAIDNPGIIGAHHKYGYHDQKYPRYCH	−167.56	1.28	0.92	1.03	5.52×10^7

[a] Also shown are the native state energy (E_{ns}), the collapse temperature (T_θ), the temperature for which the native state probability is $^1/_2$ (T_f), the effective design temperature (T'_D), and the mean folding time (τ_f) in units of mcs/N. The folding time was computed at T'_D, which was obtained from eq 23 with $T_D = 1.94$.

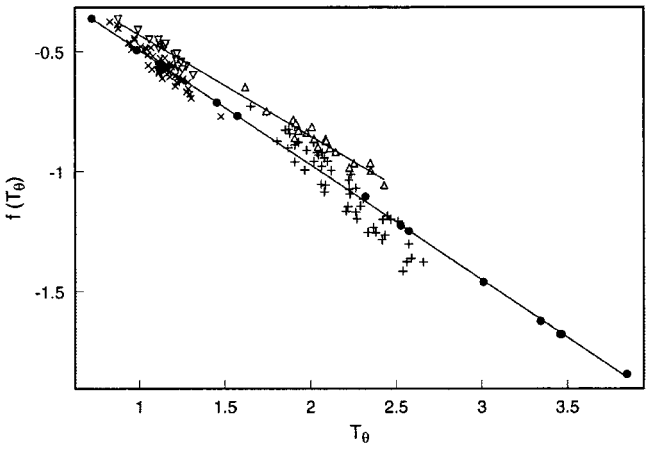

Figure 9. Fitness function evaluated at the energy collapse temperature (T_θ). (●) corresponds to a fixed 15-mer structure and sequences obtained at several temperatures. The (×) and (+) are for 50 different structures of 15-mers with optimal sequences obtained at $T_D = 1.06$ and $T_D = 1.94$, respectively. The (▽) and (△) are for 20 different structures of 36-mers with optimal sequences obtained at $T_D = 1.06$ and $T_D = 1.94$, respectively. The line is a least-squares fit for (●), which is given by $f = -0.478T_\theta$ with a correlation of $r \simeq -1$. The corresponding fit for the other 15-mers (× and +) is $f = -0.504T_\theta$ with a correlation of $r = -0.983$. For the 36-mers (▽ and △) is $f = -0.416T_\theta$ with a correlation of $r = -0.993$.

The fitness function $f(T)$ is an indicator of the folding "free energy" per contact. As we showed, its optimal value is related to the target temperature, which in turn is related to T_θ. The relation between $f(T)$ and T_θ can be observed by calculating $f(T)$ for the optimal sequence at T_θ, i.e., $f(T_\theta)$. A plot of this calculation is shown in Figure 9 for the 15-mer and the 36-mer. For a fixed 15-mer structure and the sequences of Table 1, $f(T_\theta)$ shows a strong correlation with T_θ. The correlation coefficient is $r \simeq -1$ with a fit given by $f(T_\theta) \simeq -0.48T_\theta$, which is similar to the relation between $f(T_D)$ and T_D. For a general structure this fit is also closely satisfied, even though there is a nonrandom dispersion indicating a slight dependence of the slope of the fit with the target structure. This is expected, as f depends on s, which varies with each structure. In fact, T_θ and s are well correlated, with T_θ increasing as s decreases and with a correlation coefficient of $r = 0.6$ (for sequences designed at $T_D = 1.06$). Even though designed sequences have T_θ values near T_D, generally $T_\theta > T_D$ for structures with low s values and $T_\theta \lesssim T_D$ for structures with higher s values. This explains the increase in native state stability at T_D as s decreases.

The results for the 36-mer also show a strong correlation between $f(T_\theta)$ and T_θ. This observation when combined with the expected cooperative folding for optimal sequences is consistent with the concept that these sequences have small values of σ_T. The slope of the linear fit is only slightly different from the slope in the 15-mer case. This shows that $f(T)$ is approximately an intensive quantity in units of energy per contact. The smaller slope for the 36-mers is likely due to the solvent-residue term in the fitness function, which for a globular structure varies as $N^{-1/3}$ relative to the residue−residue and solvent−solvent terms. That is, the average probability of a residue to be in contact with the solvent is larger for a small chain than for a large one. Therefore, we expect that the slope converges as the sequence length increases.

C. Lattice Model with Side Chains. As an additional test of our method, we have designed sequences for a 3-D lattice model with side chains. In this model, each residue now consists of two beads, one representing the backbone and one representing the side chain. The backbone beads are linearly connected as before, and the side chain beads are connected only to their corresponding backbone bead. The residues are geometrically identical and the only difference is in the interactions between the side chains. The side chains interact through the effective contact-pair potentials described earlier. The energy of a conformation can be written as

$$H = \epsilon_b \sum_{i=1,j>i+1}^{N} \delta_{r_{ij}^{bb},a} + \epsilon_b \sum_{i=1,j\neq i}^{N} \delta_{r_{ij}^{bs},a} + \sum_{i=1,j>i}^{N} \epsilon_{ij} \delta_{r_{ij}^{ss},a} \qquad (22)$$

where ϵ_b is the interactions between a side chain and a backbone bead, and between any two backbone beads. We arbitrarily set $\epsilon_b = -0.93$. The quantities r_{ij}^{bb}, r_{ij}^{bs}, and r_{ij}^{ss} are the spatial distance between the ith and jth residues of the backbone−backbone, backbone−side chain, and side chain−side chain, respectively. Using a chain length of $N = 32$, 10 structures were randomly generated (see Figure 8b for a selected target structure), constrained to conformations with the maximum number of contacts and with a relatively high number of side chain to side chain contacts. On average, out of a total of 81 contacts, 41.2 contacts (or about 51%) consisted of side chain to side chain contacts. These structures form 64 monomer cubes and were chosen as our target native state structures. The proportion of contacts between side chains and those between side chains and backbone is consistent with those found in proteins (unpublished). We verified this by analyzing the native structures of several proteins using the coordinates in the protein data bank.

For a target temperature of $T_D = 1.94$ we found that eight sequences resulted in nondegenerate ground states equal to the target structures. Two of them yielded ground state conformations with slightly lower energy (≃0.2% lower) and differing in only a few contacts from the target conformations. The resulting eight sequences and some of their properties are listed in Table 3.

The constraints imposed by this model affect the optimal value of f and the relationship between T_θ and T_D. The optimal value of f is around 66% of the value that could be obtained if

the backbone beads were allowed to vary as well (as the side chains do). Therefore, as more constraints are imposed on the sequence the less likely it is to find a sequence that folds into the target structure. The resulting values of T_θ are on average $\langle T_\theta \rangle = 1.20 \pm 0.11$, which are smaller than T_D. Furthermore, we found a significant linear correlation between T_D and the fraction of designable contacts, f_d. Therefore, the similarity between T_θ and T_D can be approximately recovered by redefining an effective target temperature as

$$T_D' = f_d T_D \tag{23}$$

For the preceding eight sequences and structures, $\langle f_d \rangle = 0.51$ giving $T_D' \approx 1.0$, which is approximately 17% smaller than $\langle T_\theta \rangle$ in agreement with what was found for the 36-mer without side chains.

The folding simulations were carried out by allowing in each Monte Carlo step a side-chain and a backbone move attempt. The backbone move is selected first, just like for chains with no side chain. Then, regardless of whether the backbone move is allowed or not, the side-chain move is selected by choosing a random position adjacent to its backbone residue. If either one, the backbone or the side chain move, is sterically allowed, the move is selected and the Metropolis test is performed. The folding times for all 32-mers with side chains were calculated at T_D' (see Table 3) and found to be not far from minimal ($\approx 7.3 \times 10^7$ mcs/N on average). The fastest sequence folded with a time of 1.6×10^7 mcs/N.

IV. Comparison with Other Methods

Most of the previously proposed design methods[18,19] are based on maximizing the native state stability. Two approaches are examined, one of which utilizes as optimization functions the Z-score and the other the native state probability. The Z-score is defined as $Z = (E_{ns} - \langle E \rangle)/\sigma_E$, where $\langle E \rangle$ is the average energy of all conformations and σ_E is their standard deviation. Lattice simulations have shown that folding times correlate well with Z-scores.[37] However, it is also known that sequences obtained by optimizing the Z-score with respect to the desired target structure can have native states that are different from the target structures.[37] The other popular approach optimizes the native state probability, which corresponds to maximizing $P_{ns}(T) = g_{ns} \exp(-E_{ns}/T)/Z(T)$, where g_{ns} is the native state degeneracy and $Z(T)$ is the partition function. The appeal of these optimization functions is that they produce, in principle, sequences with the desired, unique, and highly stable ground-state structure. However, these methods assume that maximal stability leads to fast folding sequences.

There are differences and limitations in each approach that can affect the properties of the designed sequences. The method based in maximizing $P_{ns}(T)$ is temperature dependent, and its optimization is a direct measure of the native state stability at a given temperature. Precise calculations by detailed exploration of conformation space, either by full enumeration[10] or by Monte Carlo methods,[18] is instructive but computationally expensive. Therefore, approximations are necessary. One of them consists of a cumulant expansion[19] of the partition function. This is a high temperature expansion in terms of the energy moments, whose calculations require the same type of approximations used to calculate the Z-score. If only the first few terms in the cumulant expansion are retained, one obtains $-T \ln P_{ns}(T) \simeq E_{ns} - (\langle E \rangle - \sigma_E^2/2T)$. To obtain a reasonable approximation to the partition function at high temperatures, about 14 terms in the cumulant expansion are needed when using lattice models.[19]

Methods based on the maximum stability criteria do not always optimize the folding kinetics consistently. To demonstrate this, we compare the results between our method, the Z-score, and the P_{ns} method for the sequences shown in Table 1. The Z-score and $P_{ns}(T)$ can be obtained exactly for these sequences by full enumeration. First, because the Z-score is temperature independent, only one sequence is optimal for all temperatures. The sequence is #13 with a Z-score value of -12.2. The other sequences have Z-score values down to -7.3. When using P_{ns} as a design criteria, only two of the 13 sequences are optimal in the given temperature range. In particular, sequence #13 is optimal between $T = 0.6$ and 0.8, #12 between $T = 1.0$ and 2.6, and again #13 between $T = 2.8$ and 4.0. This can be roughly seen in Figure 3. Sequence #13, obtained from the Z-score method, is not the most stable one for all temperatures. More importantly, sequence #13 is an extremely slow folder between $T = 2.8$ and 4.0, with a folding time that is nearly 5 orders of magnitude slower than the folding time of the sequence obtained with our method. A similar result holds for sequence #12, with a τ_f about 4 orders of magnitude larger as $T = 1.0$ is approached. It follows that designing sequences by maximizing the native state stability can lead to very slow folding sequences. This view is additionally supported from observations indicating that sequences obtained by optimizing the folding kinetics fold generally faster than sequences obtained by optimizing the Z-score.[36]

Morrissey and Shakhnovich[19] used the maximum stability criteria to design sequences, which results in optimal folding kinetics around their target temperatures. As an indication of folding rates, they show (see Figure 5 of Morrissey and Shakhnovich[19]) foldicity (% of runs that find the native state in 10^6 mcs) as function of a temperature dependent energy gap. To obtain these results, they ran 10 folding runs. Because there is a broad (Poisson) distribution of first passage time, it is known[38] that 10 trajectories is inadequate to compute τ_f reliably. Thus, foldicity is only an approximate indicator of rapid folding. In general it appears that the kinetic and stability results for sequences designed at various target temperatures, as the ones shown in Figures 3 and 4, are not reproducible by the maximum stability criteria alone.

V. Concluding Remarks

We have developed a method to generate protein sequences that rapidly fold to a unique and stable target structure at a desired temperature. It is based on the optimization of the protein landscape properties that simultaneously affect the stability and kinetics of folding. Given a target native structure and a target temperature, a de novo sequence is generated by starting from a random sequence and performing residue mutations that minimize the fitness function using the simulated annealing method.

The fitness function considers the statistics of contacts and their formation probability, which describe the energy landscape of the protein model used here. In part, minimizing $f(T) = g(T) - b(T)$ increases the probability of forming native contacts as described by $g(T)$ while it decreases the probability of forming non-native ones as described by $b(T)$. The resulting sequences generate an optimal balance between the folding kinetics and the native state stability. Consideration of residue−solvent and solvent−solvent contacts in both $b(T)$ and $g(T)$ is necessary to maintain the optimal balance between native stability and folding kinetics.

Our method leads to sequences with optimal folding properties around the target temperature. The calculations show that,

as the target temperature is varied, this balance is maintained by producing different sequences at different temperatures. Sequences at high temperatures have a lower energy native state and can accommodate stronger non-native contacts, as the landscape roughness is less critical and stability is emphasized. Additionally, native contacts in non-native conformations are important to drive the protein collapse at high temperatures. In contrast, at low temperatures the roughness effects can lead to slow dynamics and a smoother landscape is preferred over a highly stable native state. Achieving a stable native state at low temperatures is easier even with weaker native contacts. The sequences at low temperatures are characterized by higher native energies generated by weaker nonbonded contacts, which helps to avoid strong non-native contacts.

The optimal value of the fitness function is linearly related to the target temperature ($f \sim -1/2T_D$). The correlation is very strong and the slope depends on the native structure. In addition, the energy collapse temperature for optimal sequences is correlated to the target temperature. On the average, T_θ is slightly higher than T_D. This is important for the optimal balance between kinetics and stability as the minimum folding time is achieved near T_θ and the native state is stable near and below T_θ in our designed sequences. The latter is in agreement with the theoretical notion that for optimal sequences, $T_\theta \approx T_f$, where T_f is the folding temperature. We can conclude that our method produces efficient folding sequences at a specified T_θ, as long as T_θ is not too high or low. That is, achieving a very high T_θ is restricted by intrinsic stability limitations, while achieving a very low T_θ is restricted by intrinsic kinetic considerations.

The relative significance of nonlocal (far along the chain) contacts leading to their stability depends on the geometry of the native structure. This can be estimated by the structural parameter *s*, which increases with the locality of native contacts. Even though our method generate sequences with T_θ near the target temperature, the deviations are mainly due to differences in *s*. Target structures with smaller *s* values (many nonlocal native contacts) produce T_θ values typically higher than the target temperature. The relation between the ability to design a sequence for a particular structure and the locality of its contacts has been pointed out in previous simulations[39] and experiments.[40]

The calculation of $f(T)$ requires the estimation of the contact probabilities. The quasi-chemical approximation proved to be a good model, as the results justified. One of the advantages of this method, when applied to the protein model used here, is that the entropic part of the contact probability, P_{ij}^{o}, can be calculated a priori and remains constant during the minimization of $f(T)$. The only sequence dependence terms are the Boltzmann factors, which depend only on the contact energies and temperature. This is in contrast with other methods, many of them requiring the estimation of the unfolded distribution of conformation energies for each sequence. The simplicity of our fitness function makes the sequence design method very fast and efficient.

Acknowledgment. We thank Ruxandra Dima for a careful reading of the manuscript. While this research was conducted M.R.B. was a National Science Foundation Minority Postdoctoral Fellow. This work was supported in part by a grant from the National Science Foundation, grant number NSF CHE99-75150.

Appendix A: Move Selection for the Monte Carlo Method in Conformational Space

In this Appendix, we summarize the selection of the Monte Carlo backbone moves used in our kinetic and thermodynamic analysis, whenever required. The backbones moves are selected with the same protocol for chains with and without side chains. First, a tree array is constructed from the enumeration of all possible nonoverlapping conformations of a linear chain up to a maximum of $r_m + 2$ residues, where r_m is the maximum number of residues that we allow to move in one Monte Carlo step. In our simulations, we set $r_m = 3$. The move selection protocol is as follows.

1. The number of residues to move is selected with an exponentially decaying probability given by

$$P_r = \left(\frac{\gamma - 1}{\gamma^{r_m} - 1}\right)\frac{\gamma^{r_m}}{\gamma^{r}} \tag{A1}$$

where $\gamma > 0$ is an arbitrary parameter that controls the width of the distribution, and $r = 1, ..., r_m$ is the number of residues to move, with $r_m \leq N$. Here, γ is set to 1.35.

2. A chain segment of *r* residues is chosen at random, with equal probability among the $N - r + 1$ possibilities.

3. If the segment is bounded by two other residues, the conformation of the segment and the two bounded residues is searched in the tree array. The array contains a predetermined list of all other possible neighboring conformations of the segment, with the given bounding residues fixed in space. For kinetic reasons, the conformations are defined as neighbors when each individual residue is separated by two lattice sites (in the sense of the lattice or Manhattan metric) from their position in the original conformation (local displacement condition) and that none of the bounding residues are crossed. If the segment is at the end of the chain, the determination of the neighboring conformations is done similarly to the bounded segments, but this time with only one bounding residue (one end fixed).

4. One of the b_r neighboring conformations of the segment is chosen with equal probability as the new conformation. Because b_r can depend on the initial conformation of the segment, the move is accepted with probability $b_r/\max b_r$ in order to satisfy the detailed balance condition. Here, $\max b_r$ is the maximum number of neighbors that a segment of *r* residues can have under our definition of neighbors. The number b_r depends on whether the segment is bounded on one or both sides.

5. The new conformation is checked for steric violations against the rest of the chain, and if steric clashes are found, the Monte Carlo step is terminated.

6. If the move is selected, the Metropolis criteria is applied for chains without side chains. For chains with side chains, the side chain moves are determined prior to Metropolis, as described in the main text.

Appendix B: Multiple Histogram Method Implementation

In the multiple histogram method, several histograms are collected at different temperatures and combined to obtain the final histogram. In our implementation, we collect the histograms at different inverse temperatures $\beta = 1/T$ and let the algorithm select them automatically. The first β is set to zero and a histogram is obtained for an ensemble of initial conditions until it converges to a desired level of precision. From this histogram, the inverse temperature corresponding to the maximum of the estimated specific heat, β'_θ is calculated and the next histogram is obtained at β'_θ. The inverse temperature for the following histograms can be again obtained from refined estimates of the specific heat maximum. The collection of

histograms is repeated until the native state sampling is obtained to a desired degree of precision. In more complicated cases, such as when there are multiple specific heat maxima, a series of histograms can be quickly obtained for a range of inverse temperatures, starting at β'_θ, to estimate the collapse inverse temperature. These histograms are collected at a fixed time equal to the time it took the histogram obtained at β'_θ to converge. For each histogram, the minimum energy state with the prescribed precision is calculated. By fitting a quadratic polynomial through the minimum energies as function of β, its minimum β can be obtained and used to compute the final histogram.

The convergence criterion for an energy histogram at a given inverse temperature is defined by generating r simultaneous histograms $h_r(E,\beta)$, defining the error at each energy level as

$$\sigma(E,\beta) \equiv \frac{\sqrt{\langle h_r{}^2(E,\beta)\rangle - \langle h_r(E,\beta)\rangle^2}}{\sqrt{r-1}} \tag{B1}$$

where the averages are taken over the r histograms. The relative error for the average histogram $H(E,\beta) = \langle h_r(E,\beta)\rangle$ is defined as $Q(E,\beta) = \sigma(E,\beta)/H(E,\beta)$ and the total relative error for a histogram at β is

$$Q(\beta) = \frac{\sum_E Q(E,\beta)\, H(E,\beta)}{\sum_E H(E,\beta)} \tag{B2}$$

A histogram at a given β converges when $Q(\beta)$ dips below some tolerance value. The final β histogram converges when $Q(\beta)$ and $Q(E_{ns},\beta)$ converges, where E_{ns} is the native state energy. Given several histograms $H(E,\beta_0)$, $H(E,\beta_1)$, ..., $H(E,\beta_m)$, they can be combined into the final histogram by transforming them with respect to the same β, i.e., β_0, which we select as zero. If $H_n(E)$ is the combination of the first n histograms with respect to $\beta_0 = 0$ and $\sigma_n(E)$ its corresponding error, then a recursion relation for the combined histograms is given by

$$H_{n+1}(E) = \frac{\dfrac{H_n(E)}{\sigma_n{}^2(E)} + \dfrac{H(E,\beta_{n+1})}{\sigma^2(E,\beta_{n+1})} e^{-c_{n+1}-\beta_{n+1}E}}{\dfrac{1}{\sigma_{n+1}{}^2(E)}} \tag{B3}$$

and for the combined error is

$$\frac{1}{\sigma_{n+1}{}^2(E)} = \frac{1}{\sigma_n{}^2(E)} + \frac{1}{\sigma^2(E,\beta_{n+1})} e^{-2\beta_{n+1}E} \tag{B4}$$

The quantity c_{n+1} is a constant needed for the combined histogram segments to match. It is the average logarithmic difference between the temperature-transformed histograms around their overlapping region. More precisely, if $c_{n+1}(E)$ is defined as

$$c_{n+1}(E) = \ln[H(E,\beta_{n+1})e^{\beta_{n+1}E}] - \ln H_n(E) \tag{B5}$$

where the histograms are transformed in relation to $\beta = 0$, then c_{n+1} is given by

$$c_{n+1} = \frac{\sum_E \dfrac{c_{n+1}(E)}{G_{n+1}{}^2(E)}}{\sum_E \dfrac{1}{G_{n+1}{}^2(E)}} \tag{B6}$$

where $G_{n+1}(E)$ is the estimated error of $c_{n+1}(E)$, or

$$G_{n+1}{}^2(E) = \left[\frac{\sigma_n(E)}{H_n(E)}\right]^2 + \left[\frac{\sigma(E,\beta_{n+1})}{H(E,\beta_{n+1})}\right]^2 = Q_n{}^2(E) + Q^2(E,\beta_{n+1}) \tag{B7}$$

If $H(E)$ is the combination of all m histograms in relation to $\beta_0 = 0$, then the probability density at any β is obtained from $P(E,\beta) = H(E)e^{\beta E}/\sum_E H(E)e^{\beta E}$.

References and Notes

(1) Drexler, K. E. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 5275.
(2) Regan, L.; De Grado, W. F. *Science* **1988**, *241*, 976.
(3) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. *J. Chem. Phys.* **1953**, *21*, 1087.
(4) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. *Science* **1983**, *220*, 671.
(5) Shakhnovich, E. I.; Gutin, A. M. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 7195.
(6) Hellinga, H. W.; Richards, F. M. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 5803.
(7) Jones, D. T. *Protein Sci.* **1994**, *3*, 567.
(8) Friedrichs, M. S.; Goldstein, R. A.; Wolynes, P. G. *J. Mol. Biol.* **1991**, *222*, 1013.
(9) Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4918.
(10) Yue, K.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4163.
(11) Dahiyat, B. I.; Mayo, S. L. *Science* **1997**, *278*, 82.
(12) Shakhnovich, E. I. *Phys. Rev. Lett.* **1994**, *72*, 3907.
(13) Sasai, M. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 8438.
(14) Godzik, A. *Protein Eng.* **1995**, *8*, 409.
(15) Bowie, J. E.; Lüthy, R.; Eisenberg, D. *Science* **1991**, *253*, 164.
(16) Kurosky, T.; Deutsch, J. M. *J. Phys.* **1995**, *A27*, L387.
(17) Deutsch, J. M.; Kurosky, T. *Phys. Rev. Lett.* **1996**, *76*, 323.
(18) Seno, F.; Vendruscolo, M.; Martian, A.; Banavar, J. *Phys. Rev. Lett.* **1996**, *77*, 1901.
(19) Morrissey, M. P.; Shakhnovich, E. *Folding Design* **1996**, *1*, 391.
(20) Iba, Y.; Tokita, K.; Kikuchi, M. Preprint, 1998.
(21) Betancourt, M. R.; Onuchic, J. N. *J. Chem. Phys.* **1995**, *103*, 773.
(22) Bryngelson, J. D.; Wolynes, P. G. *J. Phys. Chem.* **1989**, *93*, 6902.
(23) Dill, K. A.; Chan, H. S. *Nat. Struct. Biol.* **1997**, *4*, 10.
(24) Camacho, C.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6369.
(25) Klimov, D. K.; Thirumalai, D. *Proteins Struct. Funct. Genet.* **1996**, *26*, 411.
(26) Betancourt, M. R. *J. Chem. Phys.* **1998**, *109*, 1545.
(27) Ueda, Y.; Taketomi, H.; Gō, N. *Int. J. Pept. Protein Res.* **1975**, *7*, 445.
(28) Chan, H. S.; Dill, K. A. *J. Chem. Phys.* **1989**, *90*, 492.
(29) Betancourt, M. R.; Thirumalai, D. *Protein Sci.* **1999**, *8*, 1.
(30) Specifically, the potential is defined as $E_{\alpha\beta} = 7.75(M_{\alpha\beta} + M_{tt} - M_{\alpha t} - M_{t\beta})$, where $M_{\alpha\beta}$ are the parameters from Table 3 in: Miyazawa, S.; Yernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623. t is the index for the threonine residue.
(31) Rosenbluth, M. N.; Rosenbluth, A. W. *J. Chem. Phys.* **1955**, *23*, 356.
(32) Jacobson, H.; Stockmayer, W. H. *J. Chem. Phys.* **1950**, *18*, 1600.
(33) Thirumalai, D. *J. Phys. Chem. B* **1999**, *103*, 608.
(34) Abkevich, V.; Mirny, L.; Shakhnovich, E. *Monte Carlo approach to Biopolymers and Protein Folding*; Grassberger, P., Bankana, G. T., Nadler, W., Eds.; World Scientific: Singapore, 1998; pp 1−18.
(35) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195.
(36) Shakhnovich, E. I. *Folding Design* **1998**, *3*, R45.
(37) Klimov, D. K.; Thirumalai, D. *J. Chem. Phys.* **1998**, *109*, 4119.
(38) Thirumalai, D.; Klimov, D. K. *Folding Design* **1998**, *3*, R112.
(39) Govindarajan, S.; Goldstein, R. *Biopolymers* **1995**, *36*, 43.
(40) Muñoz, V.; Serrano, L. *Folding Design* **1996**, *1*, R71.
(41) Lacroix, E.; Viguera, A.; Serrano, L. *Folding Design* **1998**, *3*, 79.