Folding behavior of model proteins with weak energetic frustration

C. Rebecca Locker and Rigoberto Hernandez

Citation: J. Chem. Phys. **120**, 11292 (2004); doi: 10.1063/1.1751394
View online: http://dx.doi.org/10.1063/1.1751394
View Table of Contents: http://jcp.aip.org/resource/1/JCPSA6/v120/i23
Published by the American Institute of Physics.

Folding behavior of model proteins with weak energetic frustration

C. Rebecca Locker and Rigoberto Hernandez[a)]
Center for Computational Molecular Science and Technology, School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332-0400

(Received 2 December 2002; accepted 26 March 2004)

The native structure of fast-folding proteins, albeit a deep local free-energy minimum, may involve a relatively small energetic penalty due to nonoptimal, though favorable, contacts between amino acid residues. The weak energetic frustration that such contacts represent varies among different proteins and may account for folding behavior not seen in unfrustrated models. Minimalist model proteins with heterogeneous contacts—as represented by lattice heteropolymers consisting of three types of monomers—also give rise to weak energetic frustration in their corresponding native structures, and the present study of their equilibrium and nonequilibrium properties reveals some of the breadth in their behavior. In order to capture this range within a detailed study of only a few proteins, four candidate protein structures (with their cognate sequences) have been selected according to a figure of merit called the winding index—a characteristic of the number of turns the protein winds about an axis. The temperature-dependent heat capacities reveal a high-temperature collapse transition, and an infrequently observed low-temperature rearrangement transition that arises because of the presence of weak energetic frustration. Simulation results motivate the definition of a new measure of *folding affinity* as a sequence-dependent free energy—a function of both a reduced stability gap and high accessibility to non-native structures—that correlates strongly with folding rates. © *2004 American Institute of Physics.* [DOI: 10.1063/1.1751394]

I. INTRODUCTION

Minimalist model proteins have been studied extensively in order to understand the thermodynamics and kinetics manifest in protein folding dynamics. Both analytical theory[1] and computer simulation[2–4] of minimalist model proteins have helped define relationships between the primary sequence of a model protein and its corresponding native structure. These studies have been useful in discovering which inherent features of the primary sequence can determine or predict whether a sequence will fold. The native structures of fast-folding proteins are most likely the global free-energy minimum structures for the system, i.e., all of the native contacts are more favorable than contacts with the solvent environment. However, not all of the intramolecular contacts in protein native structures necessarily involve a given amino acid with its most preferred contact partner. Such nonoptimal, though favorable, contacts can give rise to a weak local energetic frustration that varies among different proteins and may in turn lead to the wider range of folding behavior observed in real proteins than the range expected from entirely unfrustrated models.

In order to determine if a sequence will fold to a native structure, one must first define when the sequence is folded. In lattice models of protein folding, one first determines the native structure of a particular sequence, and subsequently identifies nonfolding sequences as those that do not fold to their native structure within a predetermined simulation time.[5] Since it would be unrealistic to simulate the folding of all possible sequences with a particular polymer length, a smaller subset of representative sequences are sampled. In many studies, the representative sequences are chosen either from a pool of random sequences, or from sequences that are designed to fold to randomly selected target structures.[2–4] Although not all of the random sequences fold, the foldable sequences have generally exhibited similar behavior because they have been constructed with monomers whose contact energies are highly degenerate, e.g., the HP model.[6] In the present work, a larger number of monomer types with only moderately degenerate contacts define the minimalist model so as to allow for the possibility of weakly frustrated target structures and domains.[1]

From both theoretical studies and minimalist model dynamic simulations, some general measures have been identified that can assess the ability of a given sequence to fold to a native structure. These figures of merit can be divided into two major groups: The first group involves those measures that emphasize the native state energy or structure, and the second group involves thermodynamic measures of phase transitions. Intuitively, the energy of the native state must be much lower than the energy of other accessible structures if the sequence is to fold. This has been observed in measurements of the stability gap in several simulations, as defined by $\delta E_s = E(Q_{\min}) - E_N$,[1] where $E(Q_{\min})$ is the average energy of the unfolded ensemble, and E_N is the energy of the sequence in the native state. The simulation results are consistent with the random energy model of heteropolymer collapse when the native structure is minimally frustrated.[1] Additional lattice simulations of minimalist model proteins have used other measures of the energetic stability, e.g., energy gap[4] or Z-score,[7] to show that the energies of non-native

a)Author to whom all correspondence should be addressed. Electronic mail: hernandez@chemistry.gatech.edu

structures are much higher than that of native structures for sequences that fold. When comparing sequences with different native state energies, the sequences with lower native state energies were observed to fold faster in general, but Socci and Onuchic[5] also noted that there is no *direct* correlation between the native state energy and the rate.

Other thermodynamic measures have been developed to identify which sequences are likely to fold as a function of their observed transition temperatures. For instance, sequences with a large temperature range between their folding and glass transitions, i.e., when the ratio T_f/T_g is much greater than one, are good folders as seen in the random energy model calculation,[1] and in computer simulations.[5] Inspired by minimalist lattice protein simulations, Thirumalai and co-workers[3] have suggested that a large temperature range between a collapse transition and a folding transition,

$$\sigma = \frac{T_\theta - T_f}{T_\theta}, \tag{1}$$

is a signature of model sequences with a high ''kinetic foldability'' *vis-a-vis* the ''folding rate'' in Ref. 3. All of these criteria are useful in selecting sequences that are likely to fold from a pool of random sequences, and they also identify some physical parameters that are important for protein folding in general. However, most protein sequences have been optimized by nature not only to fold, but to fold quickly and uniquely. In this work, we define the term, *folding affinity*, to characterize the latter concept—namely, the thermodynamics of the system—so as to create a connection to the former concept—namely, the kinetics. It includes a reduced stability gap that is related, but different than the standard stability gap that measures the energy difference between the native structure and the average energy of all structures. It also includes a measure of the number of accessible states which is related to the entropy of a putative bottleneck. The definition of this new metric is motivated by the need to better characterize proteins—whether they be designed or found[8–10] in nature—that misfold to non-native long-lived structures either because they fold slowly, or because their native state is not bound by high free energy barriers. One objective of this work can thus be surmised as the development of a thermodynamic measure, viz. the folding affinity, that can be correlated with the kinetics of protein folding.

In order to study diverse folding behavior, a weakly frustrated minimalist protein model is developed in which some of the native state contacts are not optimal as described in Sec. II. Cognate sequences corresponding to target folded structures are obtained through an inverse design procedure similar to that in Ref. 11. The target structures have been selected according to the value of their winding index,[12] as justified in Sec. II B. The design procedure, the constructed sequences, and verification of their native structures are presented in Sec. II C. A Monte Carlo pseudodynamical algorithm is used to propagate the model proteins, as discussed in Sec. III A. Phase transitions (in Sec. III B) and folding rates (in Sec. III C) are computed from the pseudodynamics in order to compare the folding behavior of the representative sequences. In addition to the heteropolymer collapse transition, the weak frustration inherent within the target native structures leads to a second lower-temperature phase transition that is clearly distinct. In Sec. IV, a measure of the folding affinity of a given sequence is defined in terms of the thermodynamics of the system and is related to its ability to fold quickly and uniquely. Microscopic measures of the folding pathway, such as the stability gap and the accessibility of non-native structures, are used in Sec. IV A to determine the folding affinity. The kinetics of representative sequences is assessed using this metric in Sec. IV B

II. THE MINIMALIST MODEL

A. Protein model

The minimalist protein model is represented as a heteropolymer on a cubic lattice, as used widely in the literature, and the specific notation used in this work may be found in Ref. 12. All sequences are composed of 27 monomers with three monomer types. Each monomer generically represents a set of about three amino acid residues,[13] and they are restricted to a three-dimensional cubic lattice. The energy for a given conformation of the minimalist protein is

$$\mathcal{H}(\{\alpha_i, \mathbf{z}_i\}) = \frac{1}{2}\sum_{i,j}{}' E_{\alpha_i,\alpha_j}, \tag{2}$$

where $\{\alpha_i\}$ is the ordered sequence of monomer types, $E_{\alpha,\beta}$ is the contact energy between the α-type monomer and β-type monomer, $\{\mathbf{z}_i\}$ is the sequence of the positions of ith monomers in a given structure, and the prime restricts j to include only those values that correspond to nonbonded nearest neighbors to i. For simplicity, the monomer types α are restricted to a set of three values, H, P and N, corresponding to the hydrophobic, polar, and neutral subregions of the protein, respectively. All of the contact energies between monomers are favorable, i.e., they are lower in energy than the contact to the uniform solvent. Thus, the native state is maximally compact,[5,14] as long as the range of contact energies is small enough so that the native (ground state) structure is not exclusively determined by the interactions between the monomers with the lowest contact energies. In this paper, the range of contact energies are chosen to be small enough compared to the average conformational energy so that the minimum energy structure is maximally compact. Specifically, these energies are $E_{H,H}=-4$, $E_{H,P}=-1$, $E_{H,N}=E_{P,N}=-2$, and $E_{P,P}=E_{N,N}=-3$. (Note that all energies and inverse temperatures reported throughout this paper are reported in dimensionless units relative to a standard temperature T_0.) The specified values for the contact energies also ensure that an increase in the number of hydrophobic contacts is the dominant driving force for folding.

B. Representing the folded space

It is not necessarily clear—in fact, unlikely—that structural metrics can be used to subdivide the space of all sequences into subspaces which can be characterized by a common value of the folding affinity. Several previous studies of the thermodynamics and kinetics of foldable and nonfoldable sequences have randomly selected sequences from the full sequence space of various representations of a 27-mer model protein. In the particular case of the model in this

work, in which equal concentrations of hydrophobic, polar, and neutral monomers are included, there are a total of 1.14×10^{11} $[=1/2(\binom{27}{9})(\binom{18}{9})]$ possible 27-mer sequences. Generating a random pool of sequences from this total space has the advantage that no previous knowledge of how the sequence will behave dynamically (i.e., its accessibility to low energy conformations) is required before studying the thermodynamic properties of the sequence. However, the disadvantage is that one does not necessarily know which type of native fold a particular sequence will assume, or if the sequence will fold at all. In order to assess the trends in the folding affinity and the folding behavior of the model sequences it is consequently useful to determine a figure of merit restricted to the space of foldable sequences that can further divide this space into smaller subspaces that might exhibit a stronger correlation between their folding affinity and their folding behavior.

1. Projected variables

Projected variables have been used routinely to identify the folding pathways, and hence are a natural choice for the identification of representative native structures among different classes of foldable model proteins. Since sequences may be designed to fold to target structures as outlined in Sec. II C, target native structures can be identified in conformation space, and the sequence that folds to the selected (nondegenerate up to mirror symmetries) native structure can be obtained. The advantages of this approach are that: (i) fewer sequences need to be studied because the dimensionality of the space of compact structures is much smaller than that of sequences, and (ii) if there exists a projected variable that effectively divides the space of foldable proteins into classes of foldable model proteins with similar characteristic behavior, then only a few representatives from each such class need to be studied in order to characterize the range of behavior. The latter assumption is unfortunately difficult to prove for any given candidate projected variable, although it is likely that a projected variable that would satisfy this condition would also serve as a reasonable folding coordinate in the spirit of the energy landscape perspective. That is, many order parameters that have been used to map the folded to unfolded pathway of lattice model proteins are structural parameters that give information about the geometry of the protein as it unfolds.[4,12,15] These structural parameters can also be used to identify candidate native structures if the parameters divide the conformation space into classes of native structures with similar behavior. For example, Fig. 1 shows the distribution of three possible order parameters for all of the maximally compact conformations, also referred to as the cube spectrum, for sequence DS1. (The DS1 sequence is given in Table I.) Each of these are discussed in the remainder of the section and we conclude by suggesting the usefulness of the winding index in the selection of characteristic structures.

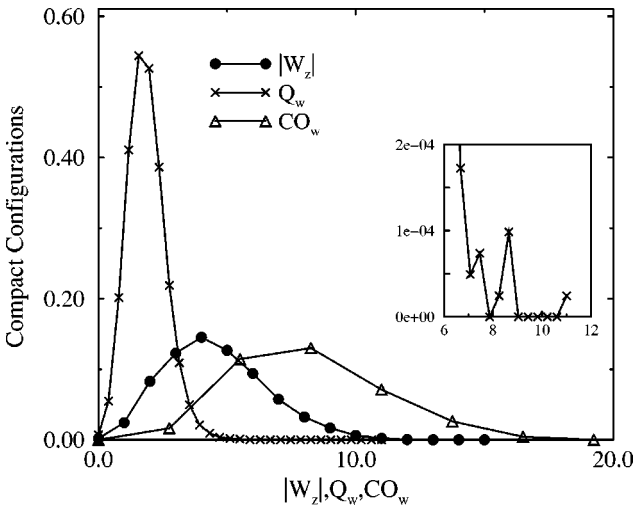

FIG. 1. The number of compact structures normalized by the total number of compact structures for each value of the winding index, W_z (defined in Sec. II B), the number of native contacts, Q, and the contact order, CO, are displayed for DS1. The points on each curve represent the only possible values of Q and W_z, but since CO can be noninteger, its distribution curve is actually continuous. The number of native contacts and the contact order are renormalized so that the values of all three measures are equal when evaluated for the native structure and the completely unfolded state, i.e., $Q_w=(w_{ns}/Q_{ns})\,Q$, and $CO_w=(11/4)CO-77/4$.

The number of native contacts, Q, is defined here as in other studies (e.g., Ref. 16) as

$$Q(\{\mathbf{z}_i\})=\frac{1}{2}\sum_{i,j}{}'\ \delta(i-n(i,j)), \tag{3}$$

where the prime restricts the sum to the nonbonded nearest neighbors in $\{\mathbf{z}_i\}$, and

$$n(i,j)=\begin{cases} i & \text{if } j \text{ non-bonded nearest neighbor of } i \text{ in the native structure,} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Since Q is a function of the native structure, it is necessarily sequence dependent. Thus, Q may not be used to designate different native structures. However, as a structural parameter, non-native structures from the same sequence with different values of Q could, in principle, be used to design other sequences. The Poisson-like distribution of Q in Fig. 1 shows that there are many compact structures with low—and similar—values of Q, but very few with high values of Q. In fact, there are some high values of Q, as shown in the insert of Fig. 1, that do not represent any compact structure. This is not surprising. Q is primarily a measure of whether a structure is or is not the native structure, and evidently does not distinguish other possibly significant metrics well even among the class of compact structures. Thus, Q is not used in this work to choose structures that represent different possible classes of foldable proteins.

TABLE I. The structural quantities, W_z, S_z, and CO used to select four maximally compact structures to represent the full conformation space of native structures. The designed sequences, DS1–DS4, are chosen arbitrarily from the group of sequences that fold to the corresponding target native structure in Fig. 2 with the minimal energy of E_{ns}.

Native structure	W_z	S_z	CO	Designed sequence	E_{ns}
NS1	11	8	10.93	*PHPNPNNNHPHPHPNNNHPHPHPNHHN*	−89
NS2	12	12	8.50	*NHPNPPNHPHPNHNHPNPNPHHNHPNH*	−90
NS3	5	5	8.64	*PHNPHNNNHPHHPNPHPHNNHPNPNHP*	−89
NS4	12	2	10.14	*PHPHNHPNHHNPHPHPNPHNNHNNPNP*	−89

The contact order, CO, is defined similarly to that in Ref. 15 as

$$\mathrm{CO}=\frac{1}{N}\sum_{i,j}^{N}{}' \Delta S_{i,j}, \tag{5}$$

where N is the total number of contacts, the prime restricts the sum to nonbonded nearest neighbors, and $\Delta S_{i,j}$ is the distance between monomer i and monomer j along the chain. The contact order is sequence independent. Thus, the distribution shown in Fig. 1 is the same for all sequences. The CO distribution is much more Gaussian-like than Q, which implies that the CO more evenly projects onto the full maximally compact conformation space than Q does. For proteins and long peptides, there is potential for a wide range of tertiary interactions, with a small value of CO representing a less folded structure than structures with a large value of CO. However, for smaller peptides, and lattice model heteropolymers with only 27 monomer units, the value of CO will always be relatively small. Since a small change in CO could correspond to a large change in topology of a lattice model protein, the contact order does not contain enough information to definitively select structures that are far apart in conformation space. In other words, conformations with different COs could be in the same folding class. In addition, CO is a real number, which is more computationally expensive than alternative structural parameters that take on only integer values. For these reasons, CO is not used to to select lattice model structures for this study.

2. The winding index

The winding index, W, about an axis generically represents the number of times the structure makes a $\pi/2$ turn about the axis. The definition of W used in Ref. 12 provided values which were not strictly invariant to rotation, and could lead to errors in very large structures. Such erroneous cases are infrequent in 27-mers and the prior results would not be affected by a corrected definition. Nonetheless, in the present work the winding index is defined as

$$W_q \equiv \sum_i \left[(P^{(q)}\Delta \mathbf{z}_{s_q(i)})\times(P^{(q)}\Delta \mathbf{z}_{s_q(i+1)})\right]_q \cdot [\Delta \mathbf{z}_{r_q(i)}]_q, \tag{6}$$

where $\Delta \mathbf{z}_i \equiv (\mathbf{z}_{i+1}-\mathbf{z}_i)$, $P^{(q)}$ projects onto the plane orthogonal to the q axis, $\{s_q(i)\}$ is an ordered sequence of the indices for which $P^{(q)}\Delta Z_{s_q(i)}$ is nonzero, the q subscript in the bracketed summand denotes the q component of the cross product, $\{r_q(i)\}$ is an ordered sequence of the indices for which $P^{(q)}\Delta Z_{s_q(i)}$ is zero, and i runs across the members of the $\{s_q(i)\}$ sequence. The scalar winding index is chosen as $W=W_z$ such that the maximal degree of winding (either positive or negative) occurs around the z axis. In cases where $|W_\alpha|=|W_\beta|$, the axis with positive winding is arbitrarily chosen as W_z. The winding index shown in Eq. (6) includes the dot product with $\Delta \mathbf{z}_{r_q(i)}$ in order to set the directionality of the winding to be concomitant with the traversal of the chain along the axis projection. For example, if the heteropolymer makes a $\pi/2$ counterclockwise turn as it moves up the axis, then it gives a *positive* contribution to the winding index. This new convention continues to ensure that mirror images will have opposite values of the winding index. However, in the present paper, the structural mirror symmetries which were central to Ref. 12 are ignored, and consequently all winding indices are reported as an absolute (positive) value.

The winding index shown in Fig. 1 smoothly maps the conformation space with a Gaussian-like distribution of values for all maximally compact structures. The winding index varies from 0 to 14 for all maximally compact conformations, so more structural information is available from W than from CO for the maximally compact conformations. In addition, the determination of W involves integer calculations that require less computational time that the real calculations required for CO. Since W evenly projects onto the space of maximally compact structures while maintaining a range wide enough to differentiate between many different structures, it is assumed that choosing structures with varying values of the winding index will give rise to foldable proteins from different folding classes.

C. Protein structures and design

The four target native structures studied in this work are shown in Fig. 2. These structures were chosen as representatives of the full conformation space of maximally compact native structures because they vary in winding indices and selectivity for winding, as shown in Table I. The selectivity for winding about the axis of maximal winding, S_z, is defined as

$$S_z \equiv |W_z|-|W_y|, \tag{7}$$

where $|W_z|\geq|W_y|\geq|W_x|$. In this sense, each native structure is far apart from the others in conformation space: the first native structure, NS1, has a moderate degree of winding and selectivity, NS2 has a high degree of winding and selectivity,

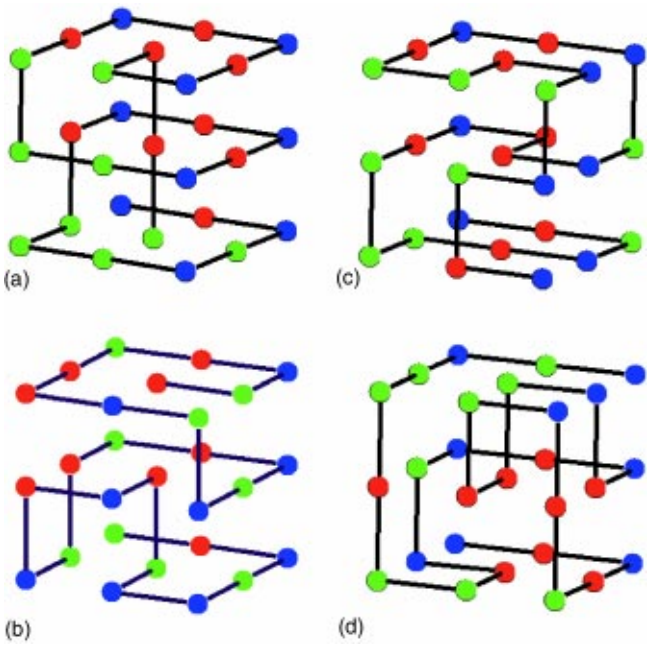

FIG. 2. The four target structures, NS1 (top left), NS2 (top right), NS3 (bottom left), and NS4 (bottom right) are displayed with their cognate sequences—DS1, DS2, DS3, DS4, respectively—distinguished by the coloring of the monomers. The three monomer types, *H*, *P*, and *N* appear respectively as red, blue, and green spheres in color prints, and as light grey, dark grey, and off white spheres in greyscale prints.

NS3 has a low degree of winding and a high selectivity, and NS4 has a high degree of winding and a low selectivity.

Several algorithms have been developed to design fast-folding model proteins with Hamiltonians similar to Eq. (2).[5,11,17–21] The energy landscape perspective of protein folding suggests that if a sequence gives rise to a fast-folding protein, then the energy for the corresponding native structure will be a minimum.[1] Given a compact target structure and fixed sequence composition, a monomer sequence that folds to the target structure may be designed by minimizing the energy of the contacts in the target structure.[5,11] In this work, the energy minimization of the sequence identity is obtained using simulated annealing by the Metropolis Monte Carlo algorithm. Trial moves correspond to the exchange of monomers in the sequence and the trial energy is obtained as a function of the contacts in the target structure for the given sequence.

All four of the designed sequences contain some energetic frustration (unlike contacts) in their corresponding compact target structure. The possibility therefore exists that any or all of these sequences would lead to a structure of lower free energy. Such an alternative structure is not compact because an exhaustive enumeration of the energies for all possible compact structures with a given sequence leads to the target structure as the energy minimum with no degeneracies in all cases. Noncompact alternative energy minimum structures are unlikely because all the contact energies in the model are favorable and differ by only small amounts. Nonetheless, noncompact alternative free energy minimum structures have also been ruled out computationally through the use of simulated annealing within a Metropolis Monte Carlo dynamics algorithm with respect to structural trial moves—*vide infra*—of all possible structures. Once it became clear that the present algorithm provided sequences with nondegenerate ground states for this class of contacts, then it was unnecessary to implement more rigorous design algorithms such as the one of Deutsch and Kurosky.[21]

The primary sequences of all four structures in Fig. 2 are designed for 27-monomer compact structures with equal concentrations of *H*, *P*, and *N* type monomers. The DS1 sequence was chosen arbitrarily from three sequences that folded to NS1 with an energy of $E=-89$. DS1 is the same sequence studied in Ref. 12 in order to illustrate the possible dynamics that would be seen if the energy landscape were dominated by not one, but two distinct energy funnels. This is seen in DS1 because there is a large enthalpic barrier between the funnel corresponding to NS1 and that of its mirror image. In the present work, however, the emphasis is on the properties of a given minimalist protein within the energy funnel, and as in much of the literature, structures that are equivalent up to mirror symmetry are not distinguished. The DS2 sequence was found to be the only sequence that folded to NS2 with an energy of $E=-90$. The DS3 sequence was chosen arbitrarily from three sequences that folded to NS3 with an energy of $E=-89$. The DS4 sequence was chosen arbitrarily from 72 sequences that folded to NS4 with an energy of $E=-89$. All four of these structures follow a general trend in which the higher the selectivity of the native structure, the lower the number of admissible sequences. Although not shown, this observation is consistent with the distribution of selectivity for all maximally compact conformations.

III. RESULTS

A. Numerical methods

The ensemble space and energetics of the possible structures on a cubic lattice corresponding to all four designed sequences have been explored using the Metropolis Monte Carlo algorithm.[11,5] Trial moves for each structure are chosen from a set of three possible types—a one-monomer corner flip, a one-monomer end pivot, and a two-monomer crankshaft move.[5] As alluded to earlier, the global free energy minimum for each of these sequences leads to the cognate compact target structure in all four cases using a slowly quenched simulated annealing procedure—not shown. All thermodynamic averages are obtained at constant temperature once the simulation has equilibrated. In Sec. III C, the pseudodynamic rates from an initial to final subspace of protein structures are obtained by averaging the results of 100 Monte Carlo trajectories originating from a random sampling of the initial space. (When the initial space contains only one structure, e.g., a straight chain, each trajectory is nevertheless distinct because different random numbers are used to propagate the Monte Carlo simulation.) The expected error of 0.1 in the natural logarithm of the rates is visible in the noise seen in the data but is small enough to support the claims to be made.

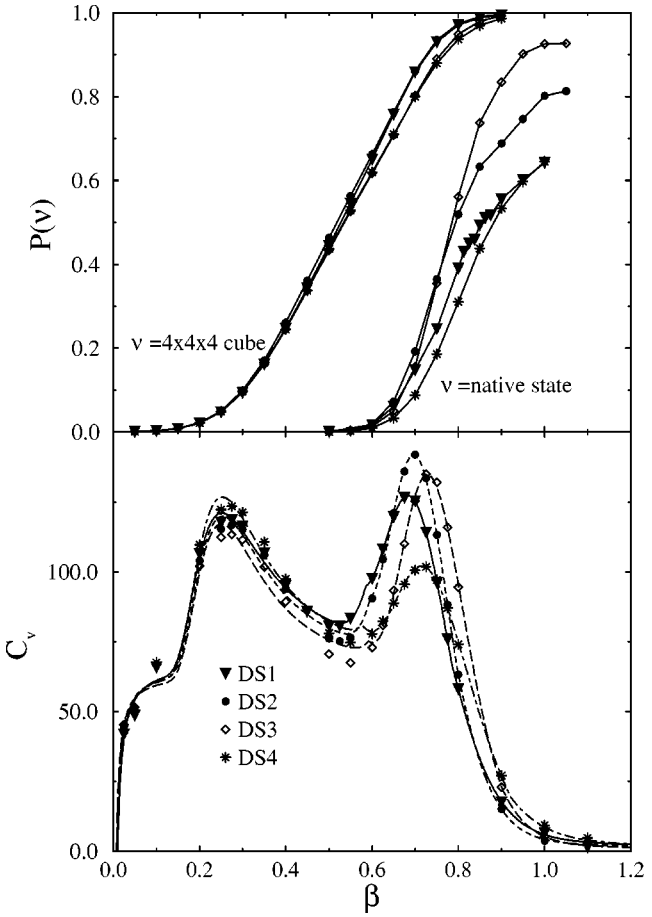

FIG. 3. The normalized probability of the sequence collapsing to a $4\times4\times4$ cube (on the left) and the probability of the sequence assuming the native state structure (on the right) is shown at various inverse temperatures in the top panel. The lower panel displays the heat capacity of DS1 (triangles), DS2 (circles), DS3 (diamonds), and DS4 (stars) at various inverse temperatures as well as the best fits of Eq. (A1) using smooth, short-dashed, long-dashed, and dot-dashed curves, respectively.

B. Phase transitions in the model proteins

The heat capacity for a canonical ensemble can be calculated from the energy fluctuations in a trajectory propagated through the Metropolis Monte Carlo algorithm by the use of[22]

$$C_v = \frac{\langle (\delta E)^2 \rangle}{k_B T^2}. \tag{8}$$

The heat capacity for all four sequences as a function of $\beta \equiv 1/k_B T$ is shown in Fig. 3, in arbitrary units (i.e., $k_B T_0 = 1$). Sigmoids and peaks in the heat capacity of this discrete model are signatures of first and second-order phase transitions, respectively. In particular, the presence of two features in Fig. 3 suggests that the model is undergoing two distinct phase transitions. The variance—viz. widths—of these peaks is related to the cooperativity of the phase transition. A first-order transition is a highly cooperative process that occurs over a small temperature range, whereas a second-order transition is a continuous, non-cooperative process that occurs over a large temperature range. Thus, one would expect a broad peak (large variance) in a plot of heat capacity versus β to correspond to a second-order transition and a narrow feature to correspond to a first-order phase transition.

The high-temperature (low β) transition has been observed in analytical protein models[1] and minimalist model simulations[2] and therein described as a second-order phase transition. The similarity in low β transition in Fig. 3 for all four sequences can likewise be attributed to a broad, second-order phase transition which appears to be sequence-independent. It should be noted that the high-temperature transition has been seen to depend on sequence among a set of fast-folding sequences with only two monomer types.[2] This seemingly contradictory case, however, did not exhibit a low-temperature phase transition. It must therefore contain features of both the heteropolymer collapse—which is sequence independent[6]—and the partial organization of the monomer contacts, whereas the high-temperature transition in the present work need only correspond to the sequence-independent collapse.

The low-temperature (high β) transition is a narrow sequence-dependent transition, though its order is not readily discernible from the heat capacity data. The low-temperature transition is most likely seen in the present model because there are four—not two or three—different values of interactions between the three distinct monomers types. The increased heterogeneity in the contact energies increases the likelihood that the structure is at least weakly frustrated energetically. This, in turn, implicates the low-temperature transition as arising from a cooperativity associated with the rearrangement of an ensemble of compact (collapsed) conformations to the native structure. If this heuristic argument is true, then the transition is first-order.

Why does this system exhibit two phase transitions? To answer this question, one might consider the dynamics in and out of the ''molten globule'' ensemble, but unfortunately it does not lend itself to a simple geometric definition. The ensemble of all structures that lie inside a $4\times4\times4$ cube does offer an alternative to study the possibility of the sequence-independent collapse. The time-averaged probability for which a sequence collapses into a $4\times4\times4$ cube, $P(4\times4\times4)$, is shown in the top panel of Fig. 3. There is a clear structural transition from high energy, extended structures at high temperatures to the compact $4\times4\times4$ structures. This transition is seen in the figure to be essentially independent of sequence with only small changes in the slope of the sigmoidal function of $P(4\times4\times4)$. As shown in Sec. IV, the small changes in the slopes of the curves can be attributed to the fact that sequences with a lower stability gap (DS1 and DS2) between their native structure and the other accessible structures spend more time in the collapsed ensemble than do the other sequences (DS3 and DS4). This structural transition is independent of the sequence as is the low β (high-temperature) phase transition, but the transition to a $4\times4\times4$ cube occurs at a slightly higher β because the confinement to a $4\times4\times4$ cube is less severe than the collapse that is presumably occurring in the low β phase transition. In summary, the data suggest that the high-temperature phase transition seen in the heat capacity is a geometric, i.e., topological, transition from an unfolded en-

semble to a collapsed ensemble that is independent of both sequence and cognate native structure.

The time-averaged probability for which the sequence assumes the native structure, $P(\mathrm{ns})$, is also shown in Fig. 3 as a function of β for all four sequences. At high temperatures, all of the sequences have a low value of $P(\mathrm{ns})$, indicating that the proteins are mostly unfolded. As the temperature of the system drops, the behavior becomes sequence-specific and some sequences spend more time in the native state than others at the same temperature. In the limit of very low temperature, it is clear that DS3 spends the greatest percentage of time in the native structure in comparison with the other three structures. Perhaps interestingly, all four studied sequences have different low-temperature limits of $P(\mathrm{ns})$. Although the low-temperature phase transition between unfolded proteins and those folded to their native structures seen in the heat capacity does not occur at precisely the same temperature, this transition is sequence-dependent like the $P(\mathrm{ns})$. This suggests that the low-temperature phase transition found in the heat capacity is a sequence dependent transition that is caused by the energetic rearrangements in folding to a specific native structure from a generic collapsed ensemble as was claimed earlier.

It is worth noting the absence of a third peak in the heat capacity curves at still lower temperatures. Such a peak would presumably arise because of a glass transition, as has previously been observed via the random energy model.[23] This transition is partially due to increased structural correlation times between the protein and its solvent. Therefore a glass transition will not be seen in the heat capacity of these minimalist model proteins because no solvent correlations—*vis-à-vis* a nonstationary frictional response—are introduced in its dynamics. Of course, one may define the glass transition to occur at some chosen point when the dynamics has slowed sufficiently so that the protein will not fold within the simulation time,[5] as was done for DS1 in Ref. 12, but this will not cause a discontinuous change in the time-averaged fluctuations of the energy of the system. Thus, the heat capacity does not reveal an invariant glass transition.

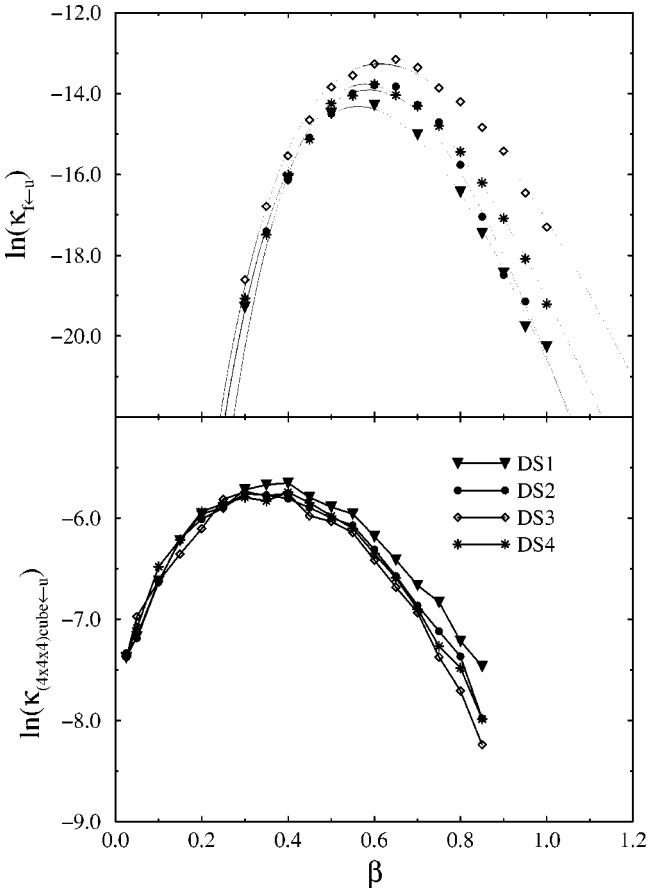

FIG. 4. The logarithms of the folding rates for a Monte Carlo dynamics simulation of DS1–DS4 are displayed at various inverse temperatures, β. The labels of the symbols are defined as in Fig. 3. The folding rates in the top panel are calculated as the inverse of the mean first passage time (MFPT) for 100 initially unfolded configurations u to reach the folded protein f. These points are overlaid with the corresponding optimal fit of the rate expression, Eq. (9), and the parameters of the fit are listed in Table II. The lower panel displays the logarithm of the collapse rates from the unfolded configurations to a $4\times4\times4$ cube. Throughout this temperature regime, the equilibration time within either the ''reactant'' or ''product'' basins is faster than the reported MFPTs, and consequently this choice of initial and final product state does not adversely affect the results. Although the error bars are about 10% of the actual rate, the error in the logarithm is significantly smaller, and hence 100 trajectories are sufficient to provide accuracy within the visible resolution.

C. Transition rates from folding dynamics

The mean first passage time (MFPT) approach, as discussed in Ref. 12, is used to calculate rates of various processes for the four designed sequences, DS1–DS4. Any pairing of the region of interest along the folding pathway will give rise to a rate process, such as the folding rate $k_{f\leftarrow u}$ or the collapse rate $k_{4\times4\times4\leftarrow u}$ between the unfolded and compact state. In the present work, there exist roughly three regions of interest along the folding pathway of each sequence. These are associated with the completely unfolded state (u), the folded—native—state (f), and the ''molten globule'' states that may perhaps be differentiated somewhere in between the two extremes. Although it is not obvious how one should represent the latter, it has been common to describe the ''molten globule'' states with respect to a geometric collapse akin to the coil–globule polymer transition. In this spirit, the collapse to a $4\times4\times4$ cube is used herein to provide a qualitative bound of the molten globule region.

The MFPT rates for folding to a $4\times4\times4$ cube from a straight-chain configuration are shown in the bottom panel of Fig. 4. In all four sequences, these β-dependent rates are independent of sequence. Similar sequence independent collapse rates have been found in other minimalist models,[2] and are consistent with the sequence-independent collapse transition seen in Fig. 3. That is, the maximal rates to the $4\times4\times4$ compact cubes occur at a temperature just below the collapse transition temperature suggested by the heat-capacity curves. The MFPT rates for folding to the native structure from a straight-chain configuration are shown in the top panel of Fig. 4. At low β, the rate curves are weakly dependent on sequence, as expected in a high temperature region where the chain can easily explore many conformations, albeit short-lived for any one conformation. The maximum rate in the turnover region of the curves is different for each sequence, which is consistent with a sequence-

TABLE II. The rates as a function of inverse temperature displayed in the top panel of Fig. 4 were also overlaid with optimal fits of the rate expression, Eq. (9). The latter can be characterized by three parameters, M, $a=\ln(\text{prefactor})$, and $b=E^{\ddagger}-E_0$, which are listed here with respect to each of the designed sequences.

Sequence	M	a	b
DS1	31.0	34.6	55.2
DS2	37.2	43.9	64.4
DS3	27.1	27.2	44.3
DS4	30.0	32.1	51.2

dependent rate at the highest rate of folding. Moreover, these observations *in toto* suggest that whatever the ''molten globule'' region actually is, it is a subspace of the $4\times4\times4$ space, and it is distinct from the native states.

The folding rates in the top panel of Fig. 4 can be described using transition state theory (TST).[24–28] In the usual TST rate formula, the prefactor is proportional to the temperature. This result follows from the general rate expression—a ratio of the flux through the activated complex $Q^{\ddagger}$ over the reactant population Q_0—which can be integrated to yield[26]

$$k=\frac{Q^{\ddagger}}{Q_0}\propto\frac{\left(\frac{h}{k_BT}\right)^{N-M}e^{-\beta E^{\ddagger}}}{\left(\frac{h}{k_BT}\right)^{N}e^{-\beta E_0}}\propto(h\beta)^{-M}e^{-\beta(E^{\ddagger}-E_0)}, \qquad (9)$$

where h is Plank's constant, N is the total number of quasibound degrees of freedom available to the reactants, $N-M$ is the number of quasibound degrees of freedom available to the activated complex, $E^{\ddagger}$ is the energy of the activated complex, and E_0 is the ground state energy. In the usual TST case, $M=1$, but in the present context of the minimalist model, a larger value of M is conceivable and could represent a decrease in the dimensionality of the activated complex because its compact structure constrains several otherwise-free modes. The optimal fits of Eq. (9) to the folding rates are overlaid over the points in Fig. 4, and their corresponding parameters are listed in Table II. The unfolded structures for all four designed sequences contain approximately 49 degrees of freedom, but the rate fits suggest that approximately 30 of these are constrained in the activated complex. (The 49 arises from counting $3N-6$ internal coordinates and subsequent subtraction of the $N-1$ fixed bonds.) Thus the transition state bottleneck for the proteins to fold is very narrow for all four sequences. Although further study is required, it may also be notable that DS2 corresponds to the largest values of W_z and S_z and leads to the largest value of M, while DS3 corresponds to the smallest values of W_z and S_z and leads to the smallest value of M. This naively suggests that greater geometric constraints (i.e., larger values of W_z and S_z) in the target structure are correlated with a more constrained activated complex, viz. a bottleneck of lower dimensionality. Note that this statement is not exactly equivalent to the statements that will be described in Sec. IV A concerning the reduced number of accessible states, N_r, though obviously they are not unrelated.

IV. FOLDING AFFINITY OF THE MODEL PROTEINS

An optimally designed sequence is one that both folds quickly to the native structure (fast dynamics) and remains there once it does (stability). The latter property will be satisfied if there is a high barrier between the native state ensemble and all other accessible conformations on the energy landscape. In principle, one could calculate a projected free energy landscape (potential of mean force) for each sequence in order to quantify the barrier height between the native state ensemble and the other accessible conformations. This procedure would only ensure an accurate assessment of the stability of the native structure if the projected variable(s) were to smoothly map the entire conformation space of the model protein, and it would be computationally expensive. The stability of the native structure measured through the low-temperature probability of being in the native structure —shown in the top panel of Fig. 3—follows the ordering, DS3$>$DS2$>$DS1$\geqslant$DS4. On the other hand, the folding rates given in Fig. 4 show the maximum rates and the broadness of the rate curves follow a different ordering, DS3 $>$DS4$\geqslant$DS2$>$DS1. Since both the rate and the stability of the native structure are important in determining whether the sequence is optimally designed, there appears to be an inconsistency in these trends. This suggests that model proteins with a high folding affinity be defined as those sequences with a high stability gap and with a large number of on-pathway conformations. These two components of the folding affinity may be obtained by an analysis of the low-temperature behavior of the folding probability as described in Sec. IV A. A further discussion describing how the folding affinity and the folding rates play a role in assessing the optimal sequence measured from thermodynamic observables is described in Sec. IV B.

A. Microscopic properties that define folding affinity

The folding affinity will be defined in the following in terms of the stability gap and the accessibility of non-native structures. These latter quantities, in turn, may be found from an analysis of the probability that the protein can be found in the native structure. This statistical probability depends on temperature through the Boltzmann distribution,

$$P_0(\beta)=\frac{e^{-\beta E_0}}{\sum_{i=0}^{N}e^{-\beta E_i}}, \qquad (10)$$

where N is the total number of states, and 0 denotes the nondegenerate native state. The probability of being in the native state can be rearranged to

$$P_0(\beta)=\frac{1}{1+e^{\beta E_0}Q^*}, \qquad (11)$$

where $Q^*=\sum_{i=1}^{N}e^{-\beta E_i}$. A microcanonical reference ensemble, $\mathcal{E}^*(\text{ns})$, consisting of all structures accessible to the native state is now introduced such that

$$e^{\beta E_0}Q^*\approx\sum_{i\epsilon\mathcal{E}^*(\text{ns})}e^{-\beta\Delta E_i}=N_r\langle e^{-\beta\Delta E_i}\rangle_*, \qquad (12)$$

where $N_r(\equiv\sum_{i\epsilon\mathcal{E}^*(\text{ns})}1)$ is the number of accessible states, $\Delta E_i\equiv(E_i-E_0)$, and $\langle\cdot\rangle_*$ is the microcanonical average

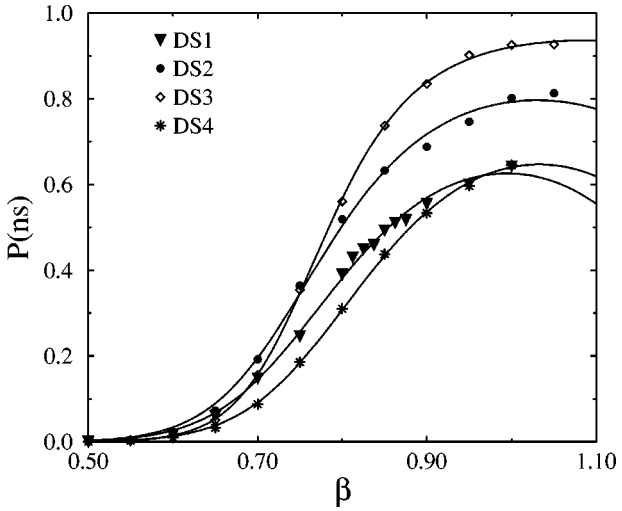

FIG. 5. The normalized probability that the sequence assumes the native structure is shown at various inverse temperatures. The best fits of the data for DS1 (triangles), DS2 (circles), DS3 (diamonds), and DS4 (stars) to Eq. (13) are also shown.

over the $\mathcal{E}^*(\mathrm{ns})$ ensemble. In this microcanonical ensemble, $k_B \ln N_r$ is also the entropy of the accessible states and hence a measure of the entropic bottleneck to folding. Expanding the average in Eq. (12) using a generalized cumulant expansion to second order leads to the following useful result:

$$P_{\mathrm{ns}}(\beta) = \left(1 + N_r \exp\left(-\beta \langle \Delta E_i \rangle_* + \frac{\beta^2}{2} \langle (\Delta E_i)^2 \rangle_* \right) \right)^{-1}. \tag{13}$$

Note that Eq. (12) is accurate at low to intermediate temperatures when only the states in $\mathcal{E}^*(\mathrm{ns})$ dominate the partition function, whereas the cumulant expansion is accurate only at high to intermediate temperatures when the higher-order terms may be safely ignored. Equation (13) should therefore be accurate at intermediate temperatures. At high temperatures, $P_{\mathrm{ns}}(\beta)$ is expected to be near zero, and agrees with Eq. (13) well within the error that is expected in the simulation data.

The probability that the sequence assumes the native structure is shown for all four sequences as a function of inverse temperature in Fig. 5. The form in Eq. (12) can be fit to the observed $P_{\mathrm{ns}}(\beta)$ in order to obtain values for the parameters in Eq. (13), i.e., estimates of N_r, $\langle (\Delta E_i)^2 \rangle_*$, and $\langle \Delta E_i \rangle_*$. The primary features in the observed $P_{\mathrm{ns}}(\beta)$ that determine the fit parameters lie in the intermediate temperature region when Eq. (13) is expected to be a reasonable approximation. The primary discrepancies are seen only at low-temperature where the third- and higher-order terms in the cumulant expansion contribute. As the temperature of the system decreases, DS3 spends more time in the native state for a given temperature (i.e., the slope of the DS3 curve is greater than the slope of the other three curves). Similarly, in the limit of low temperature, DS3 is seen in Fig. 5 to spend more time in its native state than that spent by any of the other three sequences. These observations can be quantitatively understood on a microscopic level by a fit of Eq. (13) to the data points. The number of states, N_r, available to the microcanonical reference state is orders of magnitude greater for DS3 than for any of the other three sequences, as shown in Table III. This larger number of states corresponds to an effectively wider funnel that focuses misfolded structures into the native structure of DS3, i.e., DS3 is surrounded by fewer entropic bottlenecks. This interpretation is corroborated by the maximum folding rates in Fig. 4 which are approximately correlated with N_r. Note that the zero-temperature limit of $P(\mathrm{ns})$ should equal one for all four sequences because the native structures are the minimum energy structures with no degeneracies. However, this limit is not visible in the apparent plateaus of the figure because the lowest temperature displayed is still not low compared to the excitation energy of the nearest non-native states, and hence is far from the correct limit.

TABLE III. The native state energies and the parameters extracted from fits of Eq. (13) to the data presented in Fig. 5 for all four sequences, DS1–DS4. For comparison with N_r, there are 1.14×10^{11} maximally compact structures and $6\times5^{25}=1.8\times10^{18}$ phantom chain 27-mers.

Sequence	E_{ns}	N_r	$\langle E_i \rangle_*$	$\langle (\Delta E_i)^2 \rangle_*$
DS1	−89	1.29×10^{11}	−36.5	53.0
DS2	−90	1.43×10^{11}	−37.5	51.0
DS3	−89	4.20×10^{13}	−26.6	57.2
DS4	−89	7.58×10^{11}	−34.8	52.5

Table III also lists the average energy difference of the accessible states, $\langle E_i \rangle_*$, for all of the structures. DS3 is once again unlike the others as it exhibits the largest energy difference. It should be noted that this energy difference is related to the stability gap which measures the energy difference of the native structure as compared to all other structures. But it is not exactly the same because as measured here, $\langle E_i \rangle_*$ measures the energy difference compared to the subensemble of accessible states. As such, $\langle E_i \rangle_*$ can be interpreted as a reduced stability gap. These reduced stability gaps are shown pictorially in Fig. 6, and highlight the fact that DS3 has a much larger stability gap than the other sequences. The results are generally consistent with the energy landscape theory of protein folding, which suggests that proteins with a larger stability gap are more optimally designed.

The new insight, though, is that large values in *both* N_r and $\langle E_i \rangle_*$ lead to a protein that has increased stability in its native structure—*vis-à-vis* increased folding affinity.

B. Macroscopic properties that predict folding affinity

Topological parameters such as the contact order have been shown to correlate well with the folding rate of proteins.[29] Indeed, as shown in Table I, the fastest folding sequence, DS3 has a low degree of winding, and a low contact order. The latter is presumably correlated with the lower topological frustration found in NS3 in comparison with the other native structures. That is, the bottleneck between the unfolded structures to the native structure is wider in DS3 as was noted at the end of Sec. III C.

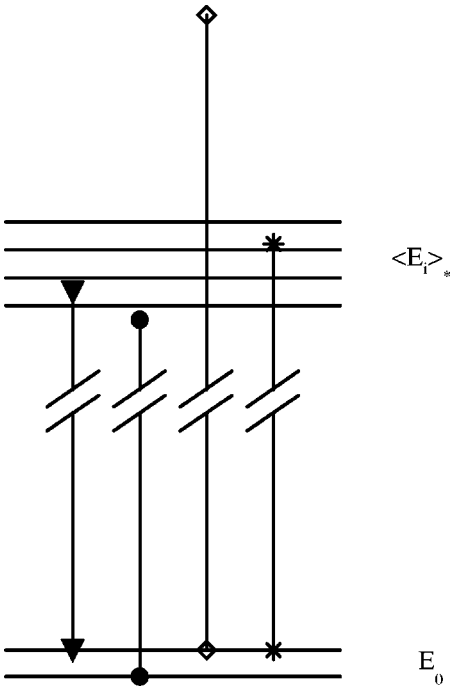

FIG. 6. The stability gaps listed in Table III as extracted from fits of Eq. (13) to the data presented in Fig. 5, are shown for DS1 (triangle), DS2 (circle), DS3 (diamond), and DS4 (star).

Thermodynamic measures that predict whether or not a sequence will fold are not all appropriate measures for determining the rates for foldable sequences. This can be illustrated using two key measures, the ratio of the folding transition temperature to the glass transition temperature, T_f/T_g, and the σ parameter [Eq. (1)]. Both of these quantities are essentially sequence-independent among foldable sequences because they depend on the glass and collapse transitions alone which are, in turn, known to be sequence-independent among foldable sequences.[1] However, it is clear that not all such sequences fold equally well.

The heat capacities in Fig. 3, however, do suggest that the behavior of the proteins at the transitions are correlated with a ''folding affinity'' via the heat of reaction. The question, though, is how to uniquely define such a folding affinity. In a two-state model, the Van't Hoff enthalpy of reaction is given by[30]

$$\Delta H = \frac{4k_B N_A T'^2 \Delta C_p}{Q_r}, \tag{14}$$

where N_A is Avogadro's number, ΔC_p is the heat capacity at the peak maximum, and Q_r is the heat of the reaction. By assuming that the collapse transition is a two-state reaction from unfolded states to collapsed states, and similarly that the folding transition is a two-state reaction from reactant collapsed states to a product native state, one can calculate the heat of collapse, ΔH_c, and the heat of folding, ΔH_f.[31] In the present case, the two transitions are overlapping, but are discernible through the fitting procedure described in detail in the Appendix.

The collapse and folding enthalpies for each of the four sequences are shown in Table IV. These enthalpies do not seem to exhibit any obvious trends. Meanwhile DS3 should be at an extreme because it folds at the fastest rate for all temperatures shown in Fig. 4 and with the greatest folding affinity as discussed earlier. The reduced stability gap, $\langle E_i \rangle_*$, alone is also not sufficient to determine the trends in the rates; there is little difference in the stability gaps of DS1, DS2, and DS4 shown in Table III, though DS1 is a substantially slower folder. The collapse and maximal folding rate nonetheless exhibit some sequence-dependence because the sequence identity defines the subspace of accessible collapsed structures.

TABLE IV. The collapse temperatures, β_c, and folding temperatures, β_f, as determined by the maximum value in the peaks of C_V in Fig. 3. The collapse and folding enthalpies are calculated using Eq. (14). Note that there is seemingly little correlation between the two enthalpies, though, as shown in Fig. 7, the logarithm of the folding rate is seemingly correlated with the collapse transition enthalpy.

Sequence	β_c	β_f	ΔH_c	ΔH_f
DS1	0.251	0.705	20.1	20.5
DS2	0.253	0.711	20.8	22.7
DS3	0.250	0.739	21.9	21.8
DS4	0.252	0.738	20.6	19.3

A possible alternative for the folding affinity may be defined through a Gibb's free energy of folding,

$$\Delta G_f = \Delta H_f - T_f \Delta S. \tag{15}$$

The enthalpy change is approximated using the value listed in Table IV. The entropy change may be calculated for the model proteins through an approximation,

$$\Delta S \approx k_B \ln \Omega \approx k_B \ln N_r, \tag{16}$$

where the number of accessible states is taken to be the value of N_r given in Table III and determined from the folding probability at the inverse folding temperature, β_f, which was in turn found through the heat capacity curves. These two thermodynamic measures combine to form a folding affinity that correlates well (for this admittedly small number of proteins) with the folding rates as shown in Fig. 7. Note that the folding affinity is clearly not the transition state barrier free energy (*vis-à-vis* $k_{f\leftarrow u} = Ae^{-\beta \Delta G}$) because if it were the slope in this figure would be -1. Nonetheless, the notion of the folding affinity of a protein seems to emerge through a relation to both the reduced stability gap and the accessibility of non-native structures. Further work is required to see if the definition provided by Eq. (15) is in fact universal.

V. CONCLUDING REMARKS

Some aspects of the folding mechanisms for a simple model of protein folding can be elucidated by the four sequences studied in this work to represent the space of sequences that fold to unique ground states with weak energetic frustration. For example, the heat capacity reveals that all sequences that fold undergo a sequence-independent transition from an unfolded to a collapsed ensemble. This is followed, at lower temperatures, by a second sequence-dependent transition from the collapsed ensemble to the native structure. A transition characterized by a structural rear-

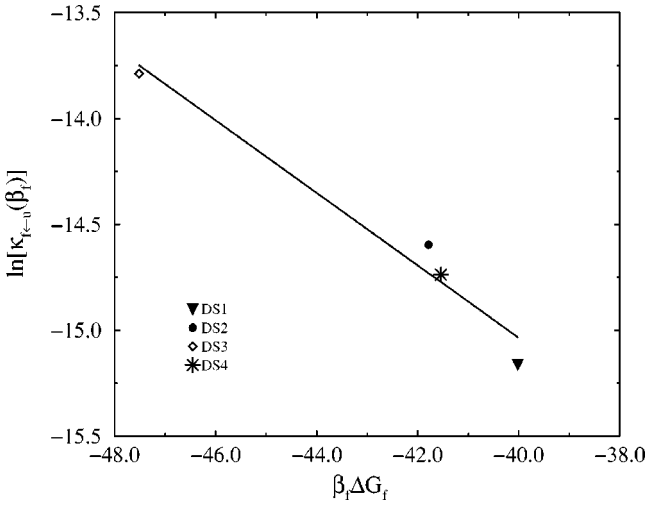

FIG. 7. The folding rate for each designed sequence at the folding temperature (shown in Table V) is extracted from the best-fit curves to the MFPT data in Fig. 4 and displayed here as a function of the folding affinity, ΔG_f, defined by Eq. (15). The best fit line has a slope of -0.171 and an intercept at -21.9.

rangement is consistent with the random energy model for protein folding, but had not been observed in previous minimalist lattice models that gave rise to unfrustrated native structures. It is most likely seen in the present model because of the weak frustration arising due to the presence of three monomer types—as opposed to two in the HP model—and the subtle differences in the contact energies assigned to each pair. The mean first passage time rates show that the slow step is the sequence-dependent rearrangement from the collapsed ensemble to the native structure. Thus, the folding rates of foldable model proteins depend on the primary sequence.

The stability gap and the conformational entropy of the native structure for each sequence has been inferred from an analysis of the probability that the corresponding model protein will be found in its native structure. (The conformational entropy of the native structure is the logarithm of the number of states accessible to the native structure.) These quantities combine to define a folding affinity and they have been shown to be correlated with more optimal protein sequence design in agreement with the energy landscape perspective of protein folding.

ACKNOWLEDGMENTS

This work has been partially supported by National Science Foundation Grant No. NSF 0213223. R.H. is a Cottrell Scholar of the Research Corporation, an Alfred P. Sloan Fellow, and the Goizueta Foundation Junior Professor. R.H. greatly acknowledges a fruitful discussion with Dr. John Tully.

APPENDIX: HEAT CAPACITY CURVES

The heat capacities measured in the numerical simulations have been fit to a functional form whose sum consists of two Gaussians—corresponding to the phase transitions—and a quasilinear baseline—corresponding to the "background" heat capacity due to internal motion. Specifically, the functional form takes the form,

$$C_{p,\mathrm{fit}}(t) \equiv G_c(T) + G_f(T) + y_{\mathrm{bl}}(T), \tag{A1}$$

where the Gaussians are parametrized as

$$G_i(T) \equiv \Upsilon_i e^{-\alpha_i (T - T_i^*)^2}, \tag{A2}$$

for $i \in (c,f)$ corresponding to the respective collapse and folding transition, and the baseline is determined by

$$y_{\mathrm{bl}}(T) = \begin{cases} 0 & \text{if } T \leqslant T_0, \\ y_1 \dfrac{(T - T_0)}{(T_1 - T_0)} & \text{if } T_0 \leqslant T \leqslant T_1, \\ y_2 + (y_1 - y_2) \dfrac{(T_2 - T)}{(T_2 - T_1)} & \text{if } T_1 \leqslant T \leqslant T_2, \\ 0 & \text{if } T_2 \leqslant T. \end{cases} \tag{A3}$$

This provides 11 free parameters; namely: Υ_c, α_c, T_c^*, Υ_f, α_f, T_f^*, T_0, T_1, T_2, y_1, and y_2. Each of the data sets have between 23 and 24 points; and consequently the least-squares fits are overdetermined by the data. The rms deviation in the fits is less than 3 (in the dimensionless units of the data) in all four fits. The fits are displayed in Fig. 3 in comparison with the data as a function of β, and the optimal parameters are listed in Table V. Note that the linear baselines, $y_{\mathrm{bl}}(T)$, are curved when shown versus β.

TABLE V. Parameters for the fits of Eq. (A1) to the data in Fig. 3 as described in the Appendix.

Parameter	DS1	DS2	DS3	DS4
Υ_c	55.8	55.8	54.9	62.8
α_c	0.316	0.347	0.369	0.334
T_c^*	3.98	3.95	4.00	3.97
Υ_f	94.3	110	103	66.0
α_f	20.4	25.8	27.7	19.9
T_f^*	1.42	1.41	1.35	1.36
T_0	1.27	1.25	1.11	1.08
T_1	1.71	1.68	1.69	1.73
T_2	98.7	121	139	114
y_1	153	153	109	101
y_2	−0.689	−0.547	−0.460	−0.578

Given the least-squares fits of Eq. (A1) to the numerical heat capacities, the free energy of each transition may be obtained. Following Refs. 30 and 31, the heat capacity from the baseline is ignored, and only the contribution due to the respective Gaussian function, $G_i(T)$, is used to compute the effective Van't Hoff enthalpy of the corresponding transition. For simplicity, the heat of reaction, Q_r, is obtained by integration of the Gaussian over an infinite domain ignoring the small error that this extended domain might entail. Evaluation of Eq. (14) for each of the transitions leads to the result,

$$\Delta H_i = 4 k_B N_A (T_i^*)^2 \sqrt{\frac{\alpha_i}{\pi}}, \tag{A4}$$

for the effective heat capacity associated with the collapse and folding transitions.

[1] J. Onuchic, P. Wolynes, and Z. Luthey-Schulten, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[2] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **103**, 4732 (1995).
[3] D. K. Klimov and D. Thirumalai, Phys. Rev. E **76**, 4070 (1996).
[4] A. Sali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[5] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **101**, 1519 (1994).
[6] E. I. Shakhnovich, Curr. Opin. Struct. Biol. **7**, 29 (1997).
[7] J. Bowie, R. Luthy, and D. Eisenburg, Science **253**, 164 (1991).
[8] S. B. Prusiner, Proc. Natl. Acad. Sci. U.S.A. **95**, 13363 (1998).
[9] J. F. Sinclair, M. M. Ziegler, and T. O. Baldwin, Nat. Struct. Biol. **1**, 320 (1994).
[10] J. W. Kelly, Curr. Opin. Struct. Biol. **8**, 101 (1998).
[11] E. I. Shakhnovich and A. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).
[12] C. R. Locker and R. Hernandez, Proc. Natl. Acad. Sci. U.S.A. **98**, 9074 (2001).
[13] J. Onuchic, P. Wolynes, Z. Luthey-Schulten, and N. Socci, Proc. Natl. Acad. Sci. U.S.A. **92**, 3626 (1995).
[14] N. D. Socci, J. N. Onuchic, and P. G. Wolynes, J. Chem. Phys. **104**, 5860 (1996).
[15] K. T. S. K. W. Plaxco and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[16] E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[17] S. Ramanathan and E. I. Shakhnovich, Phys. Rev. E **50**, 1303 (1994).
[18] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Proc. Natl. Acad. Sci. U.S.A. **91**, 12976 (1994).
[19] S. Sun, P. Thomas, and K. Dill, Protein Eng. **8**, 769 (1995).
[20] S. Sun, R. Brem, H. Chan, and K. Dill, Protein Eng. **8**, 1205 (1995).
[21] J. Deutsch and T. Kurosky, Phys. Rev. Lett. **76**, 323 (1996).
[22] D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, New York, 1987).
[23] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[24] H. Eyring, J. Chem. Phys. **3**, 107 (1935).
[25] E. P. Wigner, J. Chem. Phys. **5**, 720 (1937).
[26] P. Pechukas, *Modern Theoretical Chemistry* (Plenum, New York, 1976), Vol. 2, pp. 269–322.
[27] D. G. Truhlar and B. C. Garrett, Annu. Rev. Phys. Chem. **35**, 159 (1984).
[28] D. G. Truhlar, A. D. Issacson, and B. C. Garrett, *Theory of Chemical Reaction Dynamics* (CRC, Boca Raton, FL, 1985), Vol. 4, pp. 65–137.
[29] K. W. Plaxco and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **95**, 13591 (1998).
[30] P. L. Privalov, Adv. Protein Chem. **33**, 167 (1979).
[31] A. Bakk, A. Hansen, and K. Sneppen, Physica A **291**, 60 (2001).