Understanding protein folding cooperativity based on topological consideration

L. Wu, W. F. Li, F. Liu, J. Zhang, J. Wang,[a] and W. Wang[b]
Department of Physics and National Laboratory of Solid State Microstructure, Nanjing University, Nanjing 210093, China

(Received 14 January 2009; accepted 12 July 2009; published online 14 August 2009)

The folding cooperativity is an important issue of protein folding dynamics. Since the native topology plays a significant role in determining the folding behavior of proteins, we believe that it also has close relationship with the folding cooperativity. In the present work, we perform simulations on proteins Naf-BBL, QNND-BBL, CI2, and SH3 with the Gō model and compare their different folding behaviors. By analyzing the weak cooperative folding of protein Naf-BBL in detail, we found that the folding of Naf-BBL shows relatively weak thermodynamic coupling between residues, and such weak coupling is found mainly between the nonlocal native contacts. This finding complements our understandings on the source of barrierless folding of Naf-BBL and promotes us to analyze the topological origins of the poor thermodynamic coupling of Naf-BBL. Then, we further extend our analysis to other two-state and multistate proteins. Based on the considerations of the thermodynamic coupling and kinetic coupling, we conclude that the fraction of scattered native contacts, the difference in loop entropy of contacts, and the long range relative contact order are the major topological factors that influence the folding cooperativity. The combination of these three tertiary structural features shows significant correlations with the folding types of proteins. Moreover, we also discuss the topological factors related to downhill folding. Finally, the generic role of tertiary structure in determining the folding cooperativity is summarized. © *2009 American Institute of Physics*. [DOI: 10.1063/1.3200952]

I. INTRODUCTIOIN

The theory of protein folding has been developed greatly during the past two decades, and a number of theoretical models have been proposed. The energy landscape theory makes substantial contributions to our understanding of the general mechanism of protein folding.[1,2] According to this theory, protein folding is considered to be a diffusion process along the globally funneled energy landscape. Such folding scenario allows the proteins to fold into their native conformations or states through any of available pathways on the energy landscape. In recent years, due to progresses in experimental techniques and theoretical models, the folding behaviors of more and more proteins have been well characterized, giving substantial support to the energy landscape theory.

It is well known that the folding behaviors of proteins are determined by both the interaction energies of the various interactions and the conformational entropies of the proteins. The combination of these two factors determines the free energy landscape for folding. It was found that for those small and fast folding proteins, the competition of energy reduction and entropy lost usually results in a free energy barrier separating the unfolded states and the folded state. Because there is no stable folding intermediate, the folding or unfolding transitions are too fast to be directly observed experimentally. Therefore, the two-state folding behaves an "all-or-none" behavior and is a typical folding behavior among the various folding scenarios of proteins. For such a two-state folding, different parts of the native conformation of a protein form almost simultaneously when crossing the free energy barrier, that is, the folding cooperativity is high.

Based on the energy landscape theory, a special case has been argued to exist that if the free energy barrier is small enough, the folding process will not experience a barrier, and therefore exhibits a downhill behavior.[3–5] This important prediction from the energy landscape theory was strongly supported by the observation of the global downhill folding behavior of Naf-BBL.[6,7] However, there are also controversial issues in whether the folding of BBL is globally downhill or not. In several previously experimental and theoretical researches,[8–10] the QNND-BBL, which has an extended N-terminal tail comparing to Naf-BBL, has been reported to have different folding behavior compared with Naf-BBL, but the methods used to interpret the experimental data and the origin of the different folding behaviors of the two variants of BBL in simulations are under debate,[11,12] and there are accumulating experimental evidences suggesting that the two variants should have similar folding behaviors.[11,13–15] Therefore, the problem of downhill folding is still an ongoing research issue, but it is unquestionable that the folding cooperativity of BBL is substantially smaller than those of typical two-state proteins. In our opinion, the differences between downhill and two-state folding are primarily originated from

[a]Electronic mail: wangj@nju.edu.cn.
[b]Electronic mail: wangwei@nju.edu.cn.

their differences in folding cooperativity, which is relevant to the complicated characteristics of the various interactions between the residues in these proteins.[16]

Since the folding cooperativity is an important factor describing the folding of proteins, it was intensively studied during the past decades.[17–23] To understand its origin and biophysical significance, theoretical models and methods were developed. Among them, Chan and co-workers made a significant contribution.[20–22] They developed a method to quantitatively evaluate the folding cooperativity for proteins. By introducing a factor κ_2, the folding cooperativity can be calculated from the ratio of the van't Hoff enthalpy to the calorimetric enthalpy, and such a method for calculating the folding cooperativity is successfully used for the two-state or rough two-state proteins. A new approach to analyze the folding of proteins at atomic level was proposed recently in a study on the folding of Naf-BBL.[24] By calculating the mean thermodynamic coupling index (MTCI), the degree of coupling during folding between different parts of proteins can be evaluated. Here, the thermodynamic coupling is defined by the accordance of native probabilities of local structures, $P^N(i,T)$ of the ith residue, under different temperatures T. The distances between curves of $P^N(i,T) \sim T$ for different residues i are related to the degree of the coupling, i.e., the MTCI value. For example, if the distances between the curves $P^N(i,T) \sim T$ and $P^N(j,T) \sim T$ for all pairs of residues i and j are small, the degree of coupling index MTCI will be high. Since the native probabilities $P^N(i,T)$ can be determined through thermodynamic measurements in the nuclear magnetic resonance (NMR) relaxation experiments, the coupling between different curves is considered as the thermodynamic coupling.[24] Generally, higher degree of thermodynamic coupling means more cooperative folding. Therefore, the experimentally determined MTCI value provides a measurable quantity of the folding cooperativity as for Naf-BBL.[24] Moreover, the thermal unfolding curves required for calculating the MTCI value can also be obtained through dynamic simulations. Thus, the folding cooperativity of proteins can be evaluated both experimentally and theoretically.

The physical origin of the folding cooperativity is clearly relevant to the interactions between the residues in the native state of a protein. Theoretical simulation works have illustrated that the topology-based Gō-like models not only can well describe the two-state and multistate foldings[25] but also can successfully reproduce the downhill behavior of Naf-BBL,[16] indicating that the information of the folding mechanism or cooperativity is partially encoded in the native topology of proteins. In order to clarify the relationship between native topology and folding cooperativity, experimental data describing the folding cooperativity of proteins are needed. However, there is still lack of such quantitative data except for protein Naf-BBL. Thus, more experiments should be done or intensive simulations should be performed. Fortunately, the cooperativity of a downhill or multistate folding is much lower than that of the two-state folding in general.[16,24] This implies that the folding types, namely, a rough classification of the two-state, multistate, and downhill folding behaviors, can be used as a qualitative indicator of folding cooperativity.

In the present work, the folding of two variants of BBL (Naf-BBL and QNND-BBL) and two representative two-state proteins, i.e., CI2 and SH3, is simulated using the topology-based Gō-like model. The MTCI values are calculated and compared for these proteins. Our results show that the MTCI values correctly reflect the different folding cooperativities of the four proteins, which indicates that the degree of thermodynamic coupling well characterizes the folding cooperativity. Then, the detailed folding behavior of the representative noncooperative folding of Naf-BBL is analyzed, and both the fraction of local native contacts with respect to the total contacts and the weak coupling of nonlocal native contacts are found to dominate the weak folding cooperativity of Naf-BBL. However, the relatively weak thermodynamic coupling more likely results from the nonlocal native contacts. Combining with the analysis of the tertiary structural features of Naf-BBL, several topological factors, namely, the fraction of scattered native contacts, the difference in loop entropy of contacts, and the long range relative contact order, that are responsible for the weak thermodynamic coupling of the nonlocal contacts are proposed. The roles of these tertiary structural features in determining the folding cooperativity are thoroughly discussed based on the consideration of thermodynamic coupling. By combining the three topological factors, a substantial correlation between the tertiary structural features and the folding types of proteins is found. Moreover, we also find some possible downhill folders in the category of two-state proteins according to the folding rate, and most of them have very scattered native contacts. Finally, the generic roles of topology in determining the folding cooperativity are summarized. Our study provides an insight into the mechanism of the protein folding and the physical origin of the folding cooperativity.

II. MODELS AND METHODS

A. Gō model

Since our work mainly focuses on the topological features that influence the folding cooperativity of proteins, the topology-based Gō-like model[25–30] is suitable for this study. In the Gō model, residues are represented by identical beads located at the position of their C_α atoms and are connected by virtual bonds. The bond lengths and bond angles are restricted by harmonic potentials. A periodic trigonometric function is used to characterize the potentials for dihedral angles formed by four subsequent residues. The nonbond interactions are classified into two types, i.e., the native contacts and the non-native contacts. In this work, the native contact is defined between two residues if the distance between the closest atom pairs from these two residues is shorter than 5 Å. Each native contact is assigned with a 12-10 Lennard-Jones potential, and each non-native contact is assigned with a repulsive potential. The Gō-like potential is given as

$$E(C,C_0) = \sum_{\text{bonds}} K_r(r-r_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedral}} K_\phi^{(n)}[1+\cos(n\times(\phi-\phi_0))] + \sum_{i<j-3}^{\text{native}} \varepsilon_0\left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10}\right] + \sum_{i<j-3}^{\text{non-native}} \varepsilon_0\left(\frac{\sigma_0}{r_{ij}}\right)^{12}, \quad (1)$$

where $E(C,C_0)$ stands for the total potential energy of conformation C with C_0 as its native conformation. r, θ, and ϕ represent the bond length, bond angle, and dihedral angle, respectively. The r_0, θ_0, and ϕ_0 refer to the values of r, θ, and ϕ in the native conformation C_0. r_{ij} and σ_{ij} represent the distance between two residues i and j in conformations C and C_0. If the residues i and j form a native contact, the 12-10 Lennard-Jones potential is used; otherwise the repulsive potential is used. The parameters in the above equation are chosen as $K_r=100\varepsilon_0$, $K_\theta=20\varepsilon_0$, and $K_\phi=\varepsilon_0$ for $n=1$ or $0.5\varepsilon_0$ for $n=3$, $\sigma_0=4$ Å.[25]

B. Data analysis

To obtain the thermal unfolding curves for each residue, long time dynamic simulations are performed at a series of temperatures ranging from temperatures at which the proteins are highly folded to temperatures where the proteins are highly denatured. The native probabilities of each native contact are recorded from the simulation at each temperature. Therefore, the average native probability $P^N(i,T)$ of the residue i at the temperature T can be calculated from

$$P^N(i,T) = \sum_k \langle q_{ik}\rangle_T / n_i, \quad (2)$$

where k and n_i are the index and the total number of all the native contacts formed by residue i, and q_{ik} equal to 1 if the kth native contact of residue i is formed and 0 if it is not formed. $\langle q_{ik}\rangle_T$ is the arithmetic average of the q_{ik} values at temperature T. A series of $P^N(i,T)$ values calculated from different temperatures composes a vector p_i,

$$p_i = \{P^N(i,T_1), P^N(i,T_2), P^N(i,T_3), \ldots, P^N(i,T_n)\}. \quad (3)$$

Here, n is the total number of temperatures at which the simulations are performed. The calculation of the theoretical MTCI value is based on these vectors for different residues and is similar to that used for the experimental cases.[24] The MTCI value can be expressed as

$$\text{MTCI} = \ln(\langle\sqrt{\sigma(u_i-u_j)(u_i-u_j)^T}\rangle / \langle\sqrt{\sigma(p_i-p_j)(p_i-p_j)^T}\rangle), \quad (4)$$

where the superscript T denotes vector transpose, and σ denotes summation over the following scalar product of vectors. Therefore, if all of the contacts form simultaneously at each temperature, the difference between p_i and p_j will be very small; thus the value of the denominator of Eq. (4) will be small. Otherwise, if the folding is totally uncoupled, the difference between each pair of p_i and p_j will be large, and the denominator will be much larger. However, for different proteins, the number of residues and the average transition temperatures are different. The transitions at higher temperatures are usually sharper. These will lead to the dependence of the MTCI value on the protein size and transition temperature. Therefore, u_i and u_j are introduced to normalize the MTCI value to the broadness of the transition temperatures of proteins. The u_i and u_j are vectors similar with p_i and p_j, but the native probabilities of residues i and j are randomly chosen from the theoretically obtained thermal unfolding curves within the temperature scope of the global unfolding transition,

$$u_i = \{P^N(i,T_{r1}), P^N(i,T_{r2}), P^N(i,T_{r3}), \ldots, P^N(i,T_{rn})\}, \quad (5)$$

where the ri $(i=1,2,3,\ldots,n)$ are random numbers. The randomly chosen parameters tend to produce maximally uncoupled state and allows the MTCI value to be independent of protein size and transition temperature.[24] In this way, the thermodynamic coupling between residues can be scaled. The value of MTCI will be 0 for a totally uncoupled folding and tends to increase with the degree of coupling. Different from the experimental determination of the MTCI value, the data uncertainty is neglected in our theoretical calculations because there is little error in native probabilities when the simulation time is long enough.

The MTCI directly measures the coupling during folding between different parts of the native structure. Theoretically, the detailed folding kinetic of each residue in any protein can be explicitly obtained by dynamic simulations. This guarantees the generality of using this method to evaluate folding cooperativity. However, the computational feasibility and the accuracy of dynamic simulations become the bottleneck of this method.

C. Dynamics simulation

The off-lattice molecular dynamics simulations are carried out using the Langevin equation

$$m\ddot{r} = -\nabla_r E(C,C_0) - \gamma\dot{r} + \Gamma, \quad (6)$$

where γ is the friction coefficient and Γ is the random force which is Gaussian distributed and depends on temperature. The dynamic simulations are performed at 20 temperatures to determine each unfolding curve, i.e., the row vectors p_i and p_j contain 20 elements. Accordingly, the row vectors u_i and u_j also contain 20 elements corresponding to 20 temperatures spanning the global unfolding transition with uniform separations. At each temperature, the dynamic simulation is performed long enough to ensure the desirable convergence of native probability.

III. RESULTS

A. Weak cooperative folding of Naf-BBL

To display the differences in folding behaviors between Naf-BBL, QNND-BBL, and other two-state proteins, one dimensional free energy profiles and the thermal unfolding curves of residues for the four proteins are simulated using the Gō model, and the results are shown in Figs. 1 and 2, respectively. For protein Naf-BBL, there is no apparent free

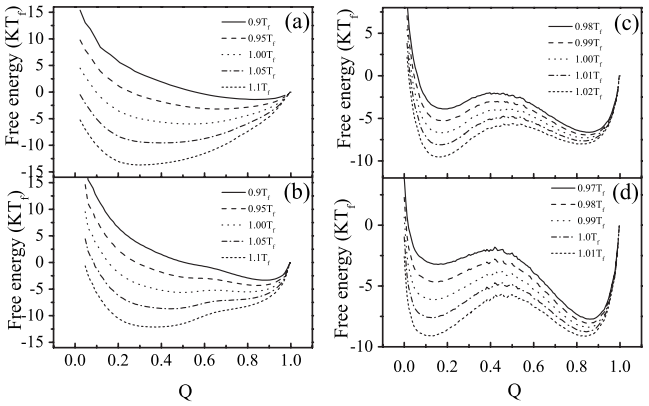

FIG. 1. Comparison of the free energy profiles for (a) Naf-BBL (PDB: 2CYU), (b) QNND-BBL (PDB: 1W4H), (c) CI2 (PDB: 2CI2), and (d) SH3 (PDB: 1NLO). The free energy profiles are calculated using the weighted histogram analysis method (WHAM) (Ref. 31) method and are plotted as a function of the reaction coordinate Q.

energy barrier [see Fig. 1(a)]. This is in accordance with the experimental findings and previous theoretical simulations that Naf-BBL folds with marginal barrier and exhibits downhill folding behavior.[6,16,24] A small free energy barrier appears in the folding of QNND-BBL [see Fig. 1(b)]. Such a difference between Naf-BBL and QNND-BBL resembles the results obtained by another native structure-based Gō model.[10] Different from the two variants of BBL, the folding of proteins CI2 and SH3 both experience significant barriers [see Figs. 1(c) and 1(d)], exhibiting similar results with previous studies.[25]

Figure 2 shows the unfolding curves which are obtained

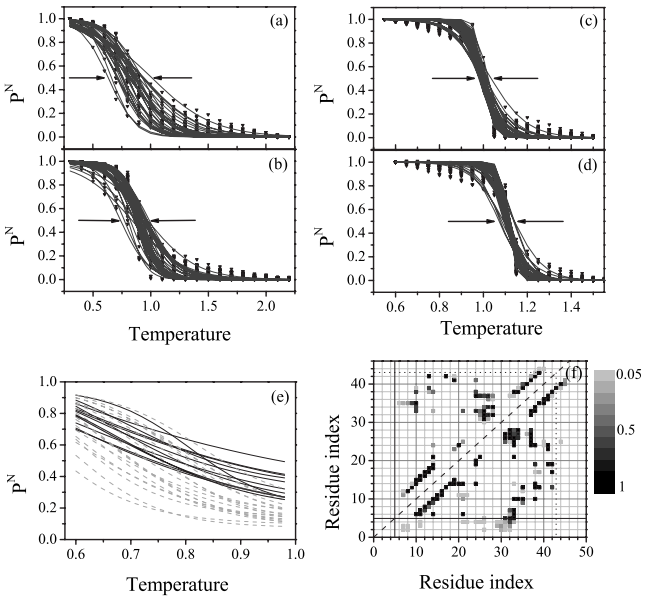

FIG. 2. Different folding behaviors of downhill and two-state proteins. The thermal unfolding curves for (a) Naf-BBL, (b) QNND-BBL, (c) CI2, and (d) SH3 are obtained from Boltzmann fitting of the native probabilities in equilibrium simulations. The global transition region from 0.75 to 1.23 of Naf-BBL is magnified in (e). The contact maps for the Naf-BBL and QNND-BBL is compared in (f). The probability of forming native contacts in the 20 NMR structures of Naf-BBL and QNND-BBL is scaled from light gray (0.05) to black (1). The vertical and horizontal solid lines signify the start position of the sequence of Naf-BBL on the QNND-BBL. The native contacts are defined by a cutoff value of 5 Å for heavy atoms.

through long time simulations at series of temperatures. Each unfolding curve represents the probability of contact formation $P_i(T)$ for residue i at temperature T as a function of the temperature T. For protein Naf-BBL, a broad distribution of the unfolding curves can be seen from Fig. 2(a), and the midpoint temperatures, defined by $P_i(T)=0.5$, have a wide range from $T_m=0.60$ to 1.00. Such a broad distribution of the unfolding curves really indicates weak thermodynamic coupling, and the folding of Naf-BBL with the topology-based model presents similar behavior with the experimental results.[24] Therefore, the weak thermodynamic coupling of Naf-BBL should also be the major source of the barrierless folding in the theoretical simulations as in the experiments.[24] Figure 2(b) shows the unfolding curves for QNND-BBL. The midpoint temperatures for QNND-BBL lie between 0.74 and 0.99, of which the distribution is a bit narrower than Naf-BBL. In comparison, the unfolding curves for proteins CI2 and SH3 distribute tightly, and midpoint temperatures have a small range from $T_m=0.98$ to 1.03 and from $T_m=1.07$ to 1.14, respectively [see Figs. 2(c) and 2(d)].

To explore the detailed mechanism of the weak thermodynamic coupling of Naf-BBL, the global transition region of the unfolding curves of Naf-BBL from temperatures of 0.6–1.0 is magnified in Fig. 2(e), and the unfolding curves are classified into two groups: the black curves corresponding to the residues that only form local contacts and the gray dashed curves corresponding to the residues that form nonlocal contacts. In Fig. 2(e), the black curves show more gradual transitions compared with the gray dashed curves, indicating smaller free energy barriers that are needed to be overcome by local native contacts during formation. However, the gray dashed curves have a much broader distribution of the midpoint temperatures, which indicates weak thermodynamic coupling of nonlocal contacts. The gradual transitions of local contacts can be explained by the small loop entropy between the two residues that form a native contact, because the loop entropy is a major source of the free energy barrier of a two-state folding. On the other hand, because of the lack of loop entropy, the formation of the local native contacts is more dependent on the interaction energy. Therefore, the local native contacts exhibit a relatively stronger thermodynamic coupling due to the homogeneous interaction energies in the Gō model. From Fig. 2(e), it can be concluded that the weak thermodynamic coupling of folding of Naf-BBL in the Gō model is mainly originated from the nonlocal native contacts. Moreover, although the local native contacts have relatively stronger thermodynamic coupling, they have gradual transitions of formation and will lower the free energy barrier of the folding of Naf-BBL. This is in accordance with the previous opinion that the fraction of local native contacts can be used to identify the downhill folders.[16,32]

The small difference between the folding behavior of Naf-BBL and QNND-BBL in the Gō model should result from the additional native contacts in QNND-BBL [see Fig. 2(f)]. However, as shown in Fig. 2(f), although most of the additional native contacts are formed by the N-terminal tail of QNND-BBL, the statistical result shows that most of these native contacts only occasionally appear in some of the

NMR structures. Thus, we believe that the QNND peptide does not form stable native contacts, but the following alanine (130A) has large probability to form native contacts with 156G, 157R, and 158L. Through structure analysis, the 130A is likely to form a backbone hydrogen bond with 158L. This hydrogen bond may be responsible for the more compact structure around it in the QNND-BBL. However, whether this difference in native structures of Naf-BBL and QNND-BBL in protein data bank (PDB) files is actually responsible for the different folding behaviors is still questionable, because the nuclear overhauser effect (NOE) maps obtained from NMR experiments for these two variants are similar, indicating an almost identical interaction network of these two variants.[12] Here, the different degrees of folding cooperativity of Naf-BBL and QNND-BBL in our simulation are clearly originated from the different native structures in their PDB files. The sensitivity of folding behavior to the change in tertiary native contacts indicates the importance of the tertiary native structure in determining the folding cooperativity.

Finally, to quantify the degree of thermodynamic coupling, the MTCI values are calculated from the theoretical unfolding curves for Naf-BBL, QNND-BBL, CI2, and SH3 and are 0.93, 1.05, 2.58, and 2.96, respectively. The MTCI value correctly reflects the relative magnitude of folding cooperativity of the four proteins. Even the slight increase in the free energy barrier of QNND-BBL can be discerned by the MTCI value. The Naf-BBL has the lowest MTCI value due to the poor thermodynamic coupling between unfolding curves. The MTCI value only for local native contacts or nonlocal native contacts of Naf-BBL are further calculated, and the values are 1.36 and 0.89, respectively, indicating stronger coupling between local native contacts. Therefore, although the gradual transition of local native contacts is an important cause of the barrierless folding of Naf-BBL, the weak thermodynamic coupling of Naf-BBL is yet mainly attributed to nonlocal native contacts, which are determined by the tertiary structure. If the specific topological feature of Naf-BBL plays a generic role in other weak cooperative folding, revealing the folding mechanism of downhill folding of Naf-BBL can help us to understand the relationship between native topology and the weak cooperative folding of proteins.

B. Topology factors that determine folding cooperativity

The general relationship between topological features and folding cooperativity is studied in previous theoretical and experimental works, in particular, the important roles of local native contacts in protein folding are emphasized.[16,33,34] However, the local native contacts have only limited effects on the folding behavior for most of the proteins, and what the exact relationship between the tertiary structures of proteins and the folding cooperativity is remains unclear. In order to discuss how the tertiary structures of proteins influence the folding cooperativity, the concept of kinetic coupling is introduced, because of its close relation with the covalent and noncovalent interaction networks which are determined by the tertiary structures.[35] The kinetic coupling refers to the cooperative kinetic behaviors of structural elements within a protein. For example, when residues in a protein are kinetically coupled, the perturbation in the position of a residue directly affects the kinetic behavior of other coupled residues. In certain case, the low kinetic coupling does not necessarily lead to low thermodynamic coupling, because the thermodynamic coupling can also result from very similar thermodynamic stabilities of local structures.[36] However, this situation is not common for proteins having the complex native structures and interaction networks, so the two kinds of couplings are usually correlated with each other.

Here, the native topology of protein Naf-BBL is analyzed, aiming to identify the most important topological factors that determine the weak kinetic coupling, along with the weak thermodynamic coupling of protein Naf-BBL. A data set of 86 proteins whose folding types are already known is used for analyzing. This data set includes all of the proteins that have been used in a previous study,[37] and the Naf-BBL is added. The folding type is roughly classified into two-state and multistate foldings, and the Naf-BBL is classified as multistate folder because of its weak folding cooperativity. The downhill folding will be separately discussed later because both downhill and multistate foldings have lower folding cooperativity than two-state folding, and thus cannot be discriminated only by the degree of folding cooperativity. The different distributions of two-state proteins and multistate proteins based on our proposed topological factors are analyzed to verify whether they also greatly influence the folding cooperativity of other proteins.

1. Fraction of scattered native contacts

Since the native contact network determines the folding behavior of a protein in the Gō-like models, it should also contain clues for the weak cooperative folding of Naf-BBL. In order to analyze the specificity of the native contact network of protein Naf-BBL, the native contact map is used. In Fig. 3(a), the contact map of Naf-BBL and CI2 are plotted to make comparison, and a more scattered distribution of the native contacts of Naf-BBL is observed. However, most of the native contacts in the protein CI2 are clustered into three groups, corresponding to one paralleled β-sheet and two antiparalleled β-sheets. These contact groups are basically constituted by sequential native contacts, i.e., the contact formed between residue i and residues $j-2$, $j-1$, j, $j+1$, and $j+2$ [see Fig. 3(b)]. Such a pattern represents a close interface between two peptides, and these sequential native contacts should have strong kinetic coupling because of the restriction from the covalent bonds along the peptide. The increase in folding cooperativity through introducing the nonadditivity of interactions in simulation is also found to be more sensitive to the structure with slightly more grouped nonlocal native contacts.[10] Therefore, the native contacts within a whole contact group between two regular secondary structures should have strong coupling between each other. However, the scattered native contacts in protein Naf-BBL are mostly formed between random coil structures and are relatively independent of each other.

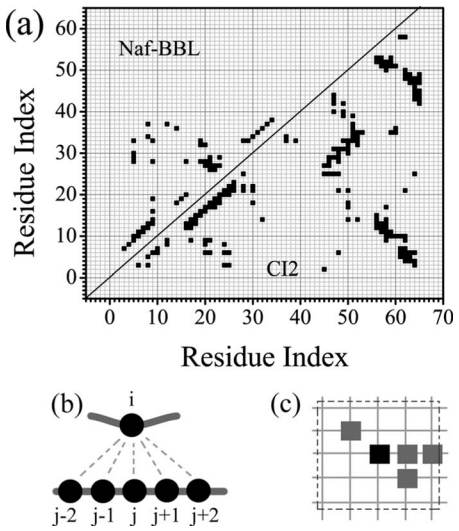

FIG. 3. (a) Native contact map for a typical native structure of Naf-BBL (upper left corner) and CI2 (lower right corner). The native contacts between residues that are separated by less than three residues are excluded. (b) The schematic figure for sequential native contacts. (c) The schematic figure for the criteria of judging a scattered native contact. If the number of native contacts within the range of two amino acids excluding the selected native contact is less than 4, the selected native contact will be identified as a scattered native contact.

Based on the above consideration, a quantity S_{NC} is defined to evaluate the degree of scattering of the native contacts,

$$S_{NC} = N_s/N_{tot}. \tag{7}$$

Here, the N_s represents the number of the nonlocal scattered native contacts and N_{tot} represents the total number of the long range native contacts. Therefore, the S_{NC} value measures the fraction of the scattered nonlocal native contacts, which is defined between the residues from coil regions and are relatively isolated on the contact map. The degree of isolation of a native contact is measured by the number of native contacts surrounding this contact on the contact map within the range of two residues, and a threshold number of 4 is used to discriminate the isolated and the grouped native contacts [see Fig. 3(c)].

The distribution of the 86 proteins in the data set according to their S_{NC} values is shown in Fig. 4(a). The S_{NC} values of multistate proteins are higher on average than that of two-state proteins which mostly have S_{NC} values smaller than 0.4. The relatively smaller S_{NC} values of two-state proteins support the conclusion that the grouping of native contacts provides strong kinetic coupling and maintains the folding cooperativity. Moreover, there is considerable overlap between multistate and two-state proteins in the region of S_{NC} value from 0.2 to 0.4, indicating that the folding intermediates not only result from the degree of scattering of native contacts. The downhill protein Naf-BBL has an extraordinary large S_{NC} value of 0.76.

2. Loop entropy difference in native contacts

Loop entropy is an important topological factor that affects the folding behavior.[38] However, the effect of loop entropy is difficult to be directly measured, because the thermal

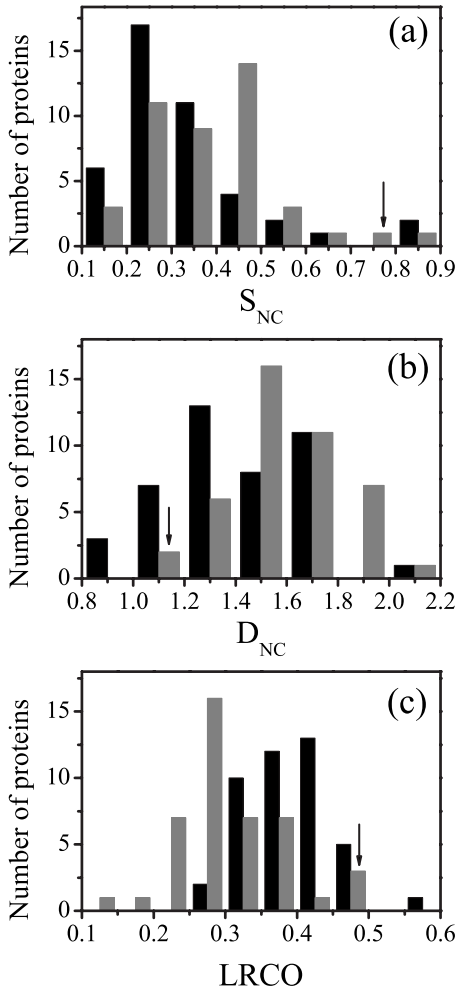

FIG. 4. Statistical distributions for two-state proteins (black bar) and multistate proteins (gray bar) on different topological features: The corresponding positions of protein Naf-BBL are marked by a down arrow. (a) Histogram of S_{NC} values which represent the fraction of scattered native contacts. (b) Histogram of D_{NC} values which represent the difference in contact loop entropy. (c) Histogram of LRCO values which measures the localness of native contacts.

behaviors of native contacts are determined by the combination of many factors. Therefore, although the energetic frustration is minimized in the Gō model, the chain connectivity will provide kinetic coupling and influence the formation probability of the individual native contacts. Here, benefiting from the weak kinetic coupling of Naf-BBL, the difference in thermal behaviors of native contacts should be mostly dominated by the difference in loop entropy of native contacts. Based on this consideration, the midpoint temperatures of each unfolding curves of Naf-BBL versus the logarithm of contact loop length are plotted in Fig. 5. Theoretically, the loop entropy varies linearly with the logarithm of the number of residues N in the loop between the two contacted residues.[39] In Fig. 5(a), the midpoint temperatures show linear correlation with loop entropy, and the correlation coefficient is 0.7. Such a correlation indicates that the difference in loop entropy is another important factor leading to poor thermal coupling between nonlocal contacts in Naf-BBL. The midpoint temperatures for unfolding curves that correspond

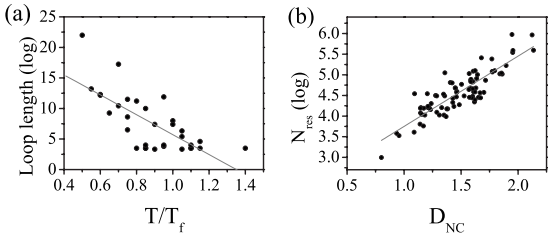

FIG. 5. (a) Logarithmized average of loop length of native contacts formed by each residue vs the midpoint temperatures of unfolding curves: The correlation coefficient of the linear fit is $R=0.71$. (b) Logarithmized number of residues of proteins vs the D_{NC} values. The correlation coefficient of the linear fit is $R=0.75$.

to local contacts are higher than that for nonlocal contacts, indicating that small loop entropy will enhance the contact thermal stability.

Here, another quantity D_{NC} is defined to evaluate the degree in difference in loop entropy of native contacts, which is expressed as

$$D_{NC} = \sqrt{\sum_i \sum_{j>i} (\log(l_m) - \log(l_n))^2 / N}, \tag{8}$$

where l_m and l_n refer to the numbers of residues between the two residues forming the mth and the nth native contacts. N represents the total number of pairs of native contacts. The D_{NC} value is the root mean square difference in the logarithmic loop length of native contacts. If most of the native contacts within a protein have similar loop lengths, the D_{NC} value will be small. Figure 4(b) shows the distribution of D_{NC} values for the 86 proteins. Generally, the two-state and multistate proteins have different distributions and peak values. Most multistate proteins have D_{NC} values >1.4, indicating that the difference in loop entropy is actually an important source of low cooperativity. However, a number of two-state proteins also have D_{NC} values above 1.4, leading to substantial superposition between the proteins with two-state and multistate folding types. As discussed above, the native contacts within an interface between regular secondary structures should have strong kinetic coupling. Such a coupling effectively reduces the difference in formation probabilities of contacts. Therefore, the formation of native contacts between antiparalleled β-sheets has cooperative behavior, even though they have considerable difference in loop length. This effect of contact groupings cannot be accounted by the D_{NC} value.

In addition, the D_{NC} value increases almost linearly with the logarithm of the number of residues[see Fig. 5(b)], and the correlation coefficient reaches 0.75. The size of protein is a simple but important topological characteristic. Large proteins usually have complex structures and interaction networks, which not only lead to low folding cooperativity[37] but also increase folding time.[40] The difference in loop entropy of native contacts is just one aspect of the complexity of protein structure.

3. Long range relative contact order

The relative contact order is also an important topological factor that was previously suggested to correlate with the folding rate[41,42] and was also used to discriminate two-state and multistate proteins.[37] However, the reason why the relative contact order correlates with folding types is not fully understood. Since our analysis is focusing on the relationship between tertiary structure of proteins and their folding cooperativity, only the long range native contacts separated by four or more residues are taken into account, and the long-range relative contact order is named LRCO. The LRCO value is defined similar with the LRCO value,

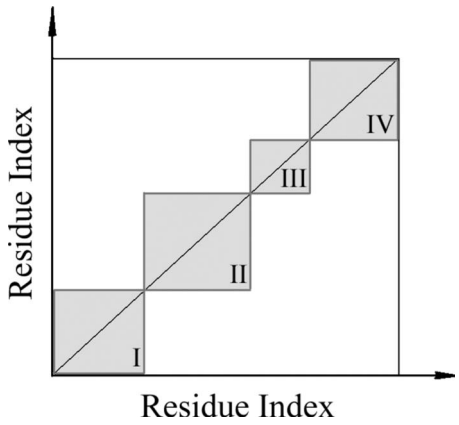

FIG. 6. A schematic contact map for the special feature of native conformation that the structural units are sequentially arranged along the peptide chain. The native contacts of proteins with such a kind of native conformation are mostly located in the gray regions (I, II, III, IV, etc.) near the diagonal, and each region represents a relatively independent structural unit.

$$\mathrm{LRCO} = \sum_k^{\text{long range}} l_k / nL, \tag{9}$$

where n refers to the number of long range native contacts, and L refers to the total number of residues. l_k is the number of residues between the two residues forming the kth long range native contact. The LRCO parameter reflects the relative localness of the native contacts. Low LRCO value usually leads to a special feature of native conformation that contains several relatively independent structural units. These structural units are sequentially arranged along the peptide chain, as shown in Fig. 6. The formation of native contacts within a single structure unit is usually cooperative and gives rise to a free energy barrier between two thermodynamic stable states. However, the coupling between different structure units is weak. This image of weak cooperative folding is typical for multistate proteins and makes them differ from the downhill folding.

The LRCO values for the 86 proteins are calculated [see Fig. 4(c)], and it is noticeable that only nonlocal native contacts separated by four or more residues are taken into account here. Different from other two topological features, the histogram of the LRCO values shows the least superposition of distributions of two-state and multistate proteins [see Fig. 4(c)]. The proteins having LRCO values below 0.3 are mostly multistate proteins and above 0.4 are mostly two-state proteins. A relationship between the LRCO value and protein size is also found. The result shows that large proteins usually prefer low LRCO values (see Fig. 7). This is in accordance with previous studies[43] and again gives support to the

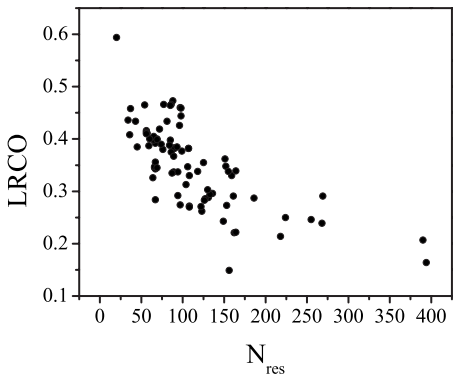

FIG. 7. Long range relative contact order vs the number of residues of proteins: Large proteins prefer low LRCO.

correlation between protein size and folding types. Moreover, the fraction of scattered native contacts also has inherent relationship with the LRCO value, because the scattered native contacts are mostly formed between local random coil structures and will also reduce the LRCO values. These two factors guarantee the reliability of discerning multistate proteins through low LRCO values. On the other hand, large LRCO value indicates small size or complex native topology with plentiful nonlocal native contacts. Small and single domain proteins have relatively small difference in loop entropy of native contacts, and proteins with large content of nonlocal contacts usually have couplings between different parts of the native structure. Therefore, proteins with large LRCO values are mostly two-state proteins.

C. Prediction of folding type

The topological features of Naf-BBL and other two-state and multistate proteins are thoroughly discussed. Although they have different distributions based on the three topological parameters, the two-state and multistate foldings of proteins cannot be discriminated through merely one topological factor. Despite that the LRCO value shows the best performance in predicting protein folding type, the accuracy is still unsatisfying. This indicates that the folding cooperativity of proteins is determined by multiple factors. Since there is inherent relationship between these topology factors, they are not independent and cannot be simply combined linearly to predict the folding type. Therefore, the 86 proteins in the data set are plotted in a three-dimensional space, of which the x, y, and z axes correspond to D_{NC}, LRCO, and S_{NC}, respectively (see Fig. 8). The two-state and multistate proteins are represented by filled and open circles, respectively. In Fig. 8, most of the two-state proteins show more compact distribution than multistate proteins which distribute out of the bulk of two-state proteins with higher S_{NC} and D_{NC}, as well as the lower LRCO values. The superposition of the two folding types in three-dimensional space seems smaller than that along a single axis. This provides the possibility to predict the folding cooperativity more accurately through multiple topological factors.

By carefully analyzing the distribution of the 86 proteins, three thresholds are defined to identify multistate proteins. They are 0.42 for S_{NC}, 1.71 for D_{NC}, and 0.31 for

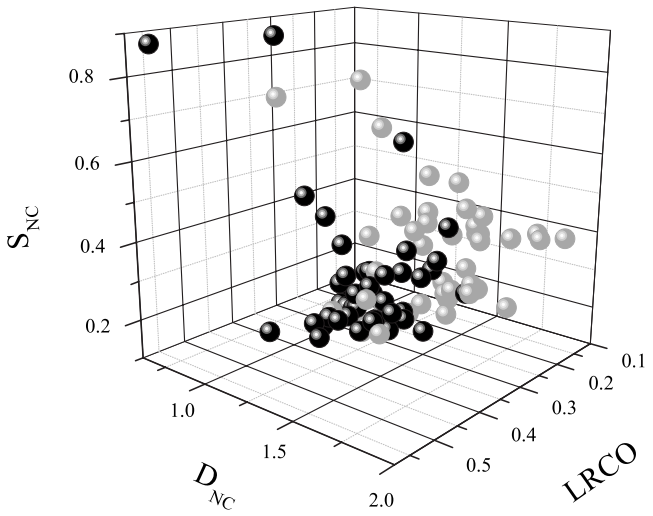

FIG. 8. Distribution of the 86 proteins in a three-dimensional space constructed by the S_{NC}, D_{NC}, and LRCO. Black spheres represent two-state proteins; light gray spheres represent multistate proteins.

LRCO, respectively. If the native structure of a protein has S_{NC} or D_{NC} greater than 0.42 and 1.72 or the LRCO smaller than 0.31, it will be considered to have a strong possibility to fold with intermediates. By using this criterion, 81.4% of multistate proteins can be successfully identified, demonstrating the rationality of our considerations. There are 18.6% of both two-state and multistate proteins that are mistakenly identified due to the loose and simple criterion that multistate proteins could be identified by any of the three thresholds. Such a simple criterion still neglects certain inherent connections between these topological factors. For example, some two-state proteins have relatively large S_{NC} values, but the D_{NC} values are small and the LRCO values are large. This implies that, although the scattered distribution of native contacts of these proteins will lead to weak kinetic coupling, the small difference in loop entropy has little effect on thermodynamic coupling and thus their folding cooperativity is maintained. Meanwhile, the relatively high LRCO values with small D_{NC} values indicate a small single domain which makes it hard to form intermediates. According to these considerations, if the proteins with LRCO>0.4 and $D_{NC}<1.2$ are excluded from the category of multistate proteins, the false-positive multistate proteins (the two-state proteins that are incorrectly identified as multistate proteins) will decrease to 11.6%, with the fraction of true-positive ones (the multistate proteins that are correctly identified) only decreasing to 79.1%. The total percentage of proteins in the data set that are correctly classified as two-state or multistate proteins will increase to 83.7%.

The accuracy of the prediction of folding type based on topological considerations indicates the predominant effect of the tertiary structure of proteins in determining their folding cooperativity. The folding behaviors of proteins are determined by multiple factors, such as the native topology, the specific interaction energies, and the solvent environment. The information of the relationship between the tertiary structures of proteins and the folding cooperativity is partially masked by these factors. Therefore, the substantial correlation between the proposed three tertiary structural fea-

TABLE I. Kinetic parameters for the seven downhill candidates which are previously classified as two-state folders. The k_{mt} is the folding rate at the midpoint transition (see Table S1 in the supplementary material, Ref. 50).

Name	PDB	Length (N)	ln k_{mt} (s^{-1})	S_{NC}	D_{NC}	LRCO
Trp-cage protein	1L2Y	20	13.0	0.89	0.80	0.59
Villin headpeace subdomain	1VII	36	10.6	0.89	0.94	0.41
Peripheral subunit-binding domain	2PDD	41	9.8	0.53	1.17	0.43
Pin WW domain	1PIN	34	9.1	0.15	0.96	0.44
Engrailed homeodomain	1ENH	54	8.1	0.51	1.37	0.47
B domain of protein *A*	1BDD	60	5.8	0.40	1.29	0.40
λ-repressor	1LMB	80	5.2	0.65	1.46	0.33

tures and the folding types reflects the objective connections between tertiary structure and folding cooperativity. This demonstrates the rationality of our analysis and again gives support to the notion that the topology plays a significant role in determining the folding behavior of proteins.

Moreover, the issues of downhill folding behavior are widely concerned. In our analysis, the folding types are roughly classified into two-state and multistate folding. The determination of the folding types of the proteins within the data set is mostly according to the traditional method, which is to fit the kinetic date from experiments with the simple two-state or three-state chemical models. However, this traditional method cannot discern the downhill and two-state folding behaviors.[44] In order to examine if there is downhill protein classified as two-state folder, we apply the method of using relaxation rate to discriminate two-state and downhill foldings[44] and find seven proteins being downhill candidates (including global downhill and incipient downhill) (see Table I). Interestingly, all of the two-state proteins which are classified as multistate folders due to the large S_{NC} values can be found in these downhill folding candidates (1L2Y, 1VII, 2PDD, 1ENH, and 1LMB). In addition, Naf-BBL also has very large S_{NC} values. These small proteins or peptide have very short sequence length, so the reduction in conformational entropy during folding is too small to produce a substantial free energy barrier,[40] thus leading to fast folding rate. In other words, even if a very small protein folds in a perfect all-or-none manner, it still has marginal free energy barrier. In the case of the six downhill candidates with large S_{NC} values, we propose that the structures of these proteins have large downhill propensity, because the native contacts of these proteins should have weak kinetic coupling as discussed above. This point of view is supported by the existence of experimental and theoretical observations of the downhill folding behaviors of their mutants or structure homologies or themselves.[6,16,44–46]

IV. CONCLUSION

By following the methodology developed by Muñoz and co-workers, the theoretical MTCI values are calculated. The success of MTCI value in discriminating the different folding cooperativities between downhill and two-state folding behaviors demonstrates the rationality of using the degree of thermodynamic coupling to measure the folding cooperativity. On the other hand, although the actual effect of the extended tails of QNND-BBL on the folding behavior remains unclear, the simulation of the two variants of protein BBL shows a sensitive correlation between the folding behavior and native contact network. Through analyzing the folding of Naf-BBL in detail, we also found that the weak thermodynamic coupling of nonlocal contacts is another important factor, besides the proportion of local native contact, which influences the folding cooperativity of Naf-BBL. Therefore, we believe the network pattern of nonlocal native contact should also greatly influence the folding of other proteins, especially for the proteins with small proportion of local native contacts.

Through analyzing the topological features of Naf-BBL in detail, the scattered distribution and the difference in loop entropy of native contacts are found to be highly related to the weak thermodynamic coupling of tertiary structure formation in protein Naf-BBL. These two factors are also supposed to be the source of weak cooperative folding of other multistate proteins and are described by S_{NC} and D_{NC} values, respectively. The S_{NC} value is found to be related to the degree of kinetic coupling between native contacts, and the D_{NC} value qualitatively evaluates the difference in loop entropy of native contacts that directly influences the thermodynamic coupling. Furthermore, the long range relative contact order is also suggested to be correlated with folding cooperativity. The LRCO value provides a macroscopic view of the topological features of proteins. Low LRCO values indicate weak kinetic coupling between relatively independent structural units. Different from the S_{NC} parameter, LRCO is more suitable to measure the degree of kinetic coupling for larger proteins. In addition, the downhill candidates defined by the relaxation rate show similar structural features. That is, large S_{NC} values and low D_{NC} values due to their short length of sequence. Since the conformational entropy plays very limited role in the folding of these proteins, the free energy barrier becomes very small or disappear, and the specific interaction energy should be more important than large proteins in determining their folding behavior.

Our analysis reveals the topological origin of weak cooperative folding. First, the kinetic coupling is largely encoded in the interaction network which is determined by the native topology. Weak kinetic coupling usually leads to weak or fragile thermodynamic coupling. Second, the native topology directly influences the folding behavior through confor-

mational entropy. Large difference in loop entropy weakens the thermodynamic coupling. However, this difference can be compensated for by energy factors, for example, some two-state proteins have stronger interaction energy for native contacts between terminal structures.[47–49] In addition, the effects of local native contacts on folding cooperativity are also not negligible.[16,33,34] Therefore, despite that the tertiary structures of proteins have remarkable effects on the folding cooperativity, it can also be influenced by many other factors. Thus, to construct a comprehensive and generic image of protein folding requires continuous efforts.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant No. 10834002 and National Basic Research Program of China under Grant Nos. 2006CB910302 and 2007CB814800. We are grateful to the High Performance Computing Center of Nanjing University and Shanghai Supercomputer Center for the award of central processing unit hours to accomplish this work.

[1] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, Science **254**, 1598 (1991).
[2] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).
[3] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[4] W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **96**, 5897 (1999).
[5] V. Muñoz, Int. J. Quantum Chem. **90**, 1522 (2002).
[6] M. M. Garci-Mira, M. Sadqi, N. Fisher, J. M. Sanchez-Ruiz, and V. Muñoz, Science **298**, 2191 (2002).
[7] V. Muñoz, Annu. Rev. Biophys. Biomol. Struct. **36**, 395 (2007).
[8] N. Ferguson, P. J. Schartau, T. D. Sharpe, S. Sato, and A. R. Fersht, J. Mol. Biol. **344**, 295 (2004).
[9] N. Ferguson, T. D. Sharpe, P. J. Schartau, S. Sato, M. D. Allen, C. M. Johnson, T. J. Rutherford, and A. R. Fersht, J. Mol. Biol. **353**, 427 (2005).
[10] S. S. Cho, P. Weinkam, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **105**, 118 (2008).
[11] A. N. Naganathan, R. Perez-Jimenez, J. M. Sanchez-Ruiz, and V. Muñoz, Biochemistry **44**, 7435 (2005).
[12] V. Muñoz, M. Sadqi, A. N. Naganathan, and D. de Sancho, HFSP J. **2**, 342 (2008).
[13] M. Sadqi, D. Fushman, and V. Muñoz, Nature (London) **445**, E17 (2007).
[14] A. N. Naganathan and V. Muñoz, Biochemistry **47**, 6752 (2008).
[15] P. Li, F. Y. Oliva, A. N. Naganathana, and V. Muñoz, Proc. Natl. Acad. Sci. U.S.A. **106**, 103 (2009).
[16] G. H. Zuo, J. Wang, and W. Wang, Proteins: Struct., Funct., Bioinf. **63**, 165 (2006).
[17] A. I. Jewett, V. S. Pande, and K. W. Plaxco, J. Mol. Biol. **326**, 247 (2003).
[18] T. R. Weikl, M. Palassini, and K. A. Dill, Science **13**, 822 (2004).
[19] A. Kolinski, W. Galazka, and J. Skolnick, Proteins: Struct., Funct., Bioinf. **26**, 271 (1996).
[20] H. S. Chan, Proteins **40**, 543 (2000).
[21] H. Kaya and H. S. Chan, Proteins **40**, 637 (2000).
[22] H. S. Chan, S. Shimizu, and H. Kaya, Methods Enzymol. **380**, 350 (2004).
[23] K. Fan, J. Wang, and W. Wang, Phys. Rev. E **64**, 041907 (2001).
[24] M. Sadqi, D. Fushman, and V. Muñoz, Nature (London) **442**, 317 (2006).
[25] C. Clementi, H. Nymeyer, and J. N. Onuchic, J. Mol. Biol. **298**, 937 (2000).
[26] C. Clementi, P. A. Jennings, and J. N. Onuchic, J. Mol. Biol. **311**, 879 (2001).
[27] C. Clementi, P. A. Jennings, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **97**, 5871 (2000).
[28] J. Karanicolas and C. L. Brooks III, J. Mol. Biol. **334**, 309 (2003).
[29] N. Koga and S. Takada, J. Mol. Biol. **313**, 171 (2001).
[30] H. Kaya and H. S. Chan, J. Mol. Biol. **326**, 911 (2003).
[31] R. H. Swendsen, Physica A **194**, 53 (1993).
[32] L. Prieto and A. Rey, J. Chem. Phys. **127**, 175101 (2007).
[33] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, J. Mol. Biol. **252**, 460 (1995).
[34] V. Muñoz and L. Serrano, Folding Des. **1**, R71 (1996).
[35] J. Li, J. Wang, and W. Wang, Proteins: Struct., Funct., Bioinf. **71**, 1899 (2008)m.
[36] C. M. Bradley and D. Barrick, J. Mol. Biol. **324**, 373 (2002).
[37] B. G. Ma, L. L. Chen, and H. Y. Zhang, J. Mol. Biol. **370**, 439 (2007).
[38] J. Chen, J. Wang, and W. Wang, Proteins: Struct., Funct., Bioinf. **57**, 153 (2004).
[39] A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **97**, 1525 (2000).
[40] A. N. Naganathan and V. Muñoz, J. Am. Chem. Soc. **127**, 480 (2005).
[41] K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[42] R. Bonneau, I. Ruczinski, J. Tsai, and D. Baker, Protein Sci. **11**, 1937 (2002).
[43] D. N. Ivankov, S. O. Garbuzynskiy, E. Alm, K. W. Plaxco, D. Baker, and A. V. Finkelstein, Protein Sci. **12**, 2057 (2003).
[44] A. N. Naganathan, U. Doshi, and V. Muñoz, J. Am. Chem. Soc. **129**, 5673 (2007).
[45] R. H. Zhou, Proc. Natl. Acad. Sci. U.S.A. **100**, 13280 (2003).
[46] F. Liu and M. Gruebele, J. Mol. Biol. **370**, 574 (2007).
[47] M. M. Krishna and S. W. Englander, Proc. Natl. Acad. Sci. U.S.A. **102**, 1053 (2005).
[48] M. Lingberg, J. Tångrot, and M. Oliveberg, Nat. Struct. Biol. **9**, 818 (2002).
[49] L. Wu, J. Zhang, J. Wang, W. F. Li, and W. Wang, Phys. Rev. E **75**, 031914 (2007).
[50] See EPAPS supplementary material at http://dx.doi.org/10.1063/1.3200952 E-JCPSA6-131-049931 for the kinetic data used for identifying downhill candidates, presented in table format.