Folding behavior of ribosomal protein S6 studied by modified Gō-like model

L. Wu, J. Zhang, J. Wang, W. F. Li, and W. Wang*

National Laboratory of Solid State Microstructure and Department of Physics, Nanjing University, 210093, China

(Received 15 October 2006; published 28 March 2007)

Recent experimental and theoretical studies suggest that, although topology is the determinant factor in protein folding, especially for small single-domain proteins, energetic factors also play an important role in the folding process. The ribosomal protein S6 has been subjected to intensive studies. A radical change of the transition state in its circular permutants has been observed, which is believed to be caused by a biased distribution of contact energies. Since the simplistic topology-only Gō-like model is not able to reproduce such an observation, we modify the model by introducing variable contact energies between residues based on their physicochemical properties. The modified Gō-like model can successfully reproduce the Φ-value distributions, folding nucleus, and folding pathways of both the wild-type and circular permutants of S6. Furthermore, by comparing the results of the modified and the simplistic models, we find that the hydrophobic effect constructs the major force that balances the loop entropies. This may indicate that nature maintains the folding cooperativity of this protein by carefully arranging the location of hydrophobic residues in the sequence. Our study reveals a strategy or mechanism used by nature to get out of the dilemma when the native structure, possibly required by biological function, conflicts with folding cooperativity. Finally, the possible relationship between such a design of nature and amyloidosis is also discussed.

DOI: 10.1103/PhysRevE.75.031914 PACS number(s): 87.14.Ee, 87.15.Cc

I. INTRODUCTION

It is widely accepted that the shape of the energy funnel is largely determined by the native topology of proteins. This is supported by extensive experimental observations and simulation results. For example, the structures of the transition states (TSs) are similar for proteins with similar topologies and are insensitive to site mutations. The folding rates exhibit a strong dependence on a topological parameter, namely, the contact order. The folding pathway and the structures of the TS and intermediate states can be predicted by simplistic Gō-models that incorporate only structural information about the native states. However, with the advancement of experimental techniques and simulation methodologies, it has been discovered that, in addition to the topological factor, the energy also plays an important role [1,2], especially in some proteins such as the ribosomal protein S6.

The speciality of S6 was noticed in circular permutation experiments [2,3]. Circular permutation works by breaking the chain at a certain position and connecting the original terminals by a chemical bond or peptide linker. This keeps the interactions between residues intact but changes the topology and loop entropy. The Φ value [4], which normalizes the perturbation of the site mutation in the overall barrier by the same perturbation in the stability of the native state, is used to quantify the nativeness of each residue in the structures of the transition states. It is found in such experiments that the distribution of Φ values of S6 changes from diffuse to polarized after circular permutation. In contrast, for example, the distribution remains diffuse (or polarized) for CI2 (or SH3) after circular permutation [5–7]. In detail, the Φ values of the wild-type S6 are uniformly distributed within 0.1–0.3, whereas those of the circular permutants $P^{13\text{-}14}$, $P^{68\text{-}69}$, and $P^{54\text{-}55}$ are characterized by a bimodal distribution, centering on 0 and 1, respectively.

To understand the interesting behavior of S6 mentioned above, extensive experiments and theoretical studies have been done, and it is concluded that such behavior is the consequence of competition between the biased contact energy and loop entropy in S6 [8–13]. This can be understood as follows. First, the ribosomal protein S6 has a biased loop entropy that may be requisite for its function. However, this puts the protein in danger of low folding cooperativity and thus of high possibility of partial unfolding and aggregation. To alleviate this danger, a biased contact energy is introduced into the native structure to compensate for the biased loop entropy, resulting in a diffuse TS. For the circular permutants, however, this delicate balance breaks where strong contact energy and short sequence separation both occur, rendering the TS atypically polarized. Previous work revealed such a biased contact energy in the native structure of S6, but how is it coded in?

In recent decades, the simplistic Gō model [14] has shown great success qualitatively or semiquantitively in characterizing the folding behaviors of most small proteins [15–17]. This model treats all the native interactions equally, ignoring the fact that the interaction strengths between residues are actually variable. In this way, it suppresses the energetic frustrations to the lowest level and emphasizes the importance of the topological factors in folding. The successes of such a simplistic model suggest that the energy landscapes of most proteins are mainly determined by their native topologies. However, despite many successes, this model fails to describe the folding behaviors of protein S6. This is not surprising due to the biased energy distribution in its native structure, as suggested in many previous works.

Theoretically, the folding of S6 and this biased contact energy have been studied by several groups. For example,

*Electronic address: wangwei@nju.edu.cn

Shakhnovich and co-workers investigated the folding process of S6 using an all-atom Gō model in conjunction with restraints from experimental Φ values. With those restraints, the biased contact energy is automatically introduced into their model which enables their model to reproduce most experimental observations. Following a similar idea but stepping further, Clementi and colleagues developed a systematic strategy to incorporate experimental data into a coarse-grained model. Their model can also describe the folding behaviors of S6 and indeed exhibits the atypical distribution of contact energies.

Although there are already several successful models, we feel that the situation is not satisfactory because, for the models mentioned above, the parameters of contact interactions have to be calibrated by introducing experimental data. This limitation reduces their ability to model the folding behaviors of new proteins. In our work, by analyzing the various interactions in S6, we modify the simplistic Gō model and obtain a set of parameters that reproduce the biased distribution of contact energies in the native structure. Using this modified Gō model, we investigate the folding behaviors of S6 and find that the predicted folding pathway, theoretical Φ values, and folding nucleus all agree well with the experimental observations. This success demonstrates that our model is workable for the protein S6 and can also be used to investigate other atypical proteins with energetic factors that play critical roles in the folding process of these proteins.

II. MODELS AND METHODS

A. Protein S6 and its circular permutants

The protein S6 is a ribosomal protein consisting of 97 residues (Protein Date Bank 1RIS). It has two helices and four β strands which are symmetrically distributed along the sequence [Fig. 1(a)]. The four parallel β strands are termed strand 1 (s1) to strand 4 (s4) following the sequence order. We define three interfaces between these β strands; the interface 1 is between strand 1 and strand 3, the interface 2 is between strand 2 and strand 3, and the interface 3 is between strand 1 and strand 4. Interface 3 in particular contains a large number of long-range contacts. The folding behavior of the circular permutant $P^{13\text{-}14}$ is also investigated; this is created by cutting the peptide bond between residues N13 and L14, and connecting the two original termini by a small loop.

B. Contact energies

In the simplistic C_α Gō model, all native contacts are assigned the same interaction strength. Here a native contact is defined when the distance between any heavy atom pair from two residues is below a cutoff 5 Å. However, intuitively, the interaction strength between residues should be distance dependent and residue dependent, and these interactions should mainly comprise van der Waals (vdW) interactions, hydrogen-bond interactions, and hydrophobic interactions. Here, we investigate the relative strengths of these interactions, and we take their magnitudes as coefficients in our model.

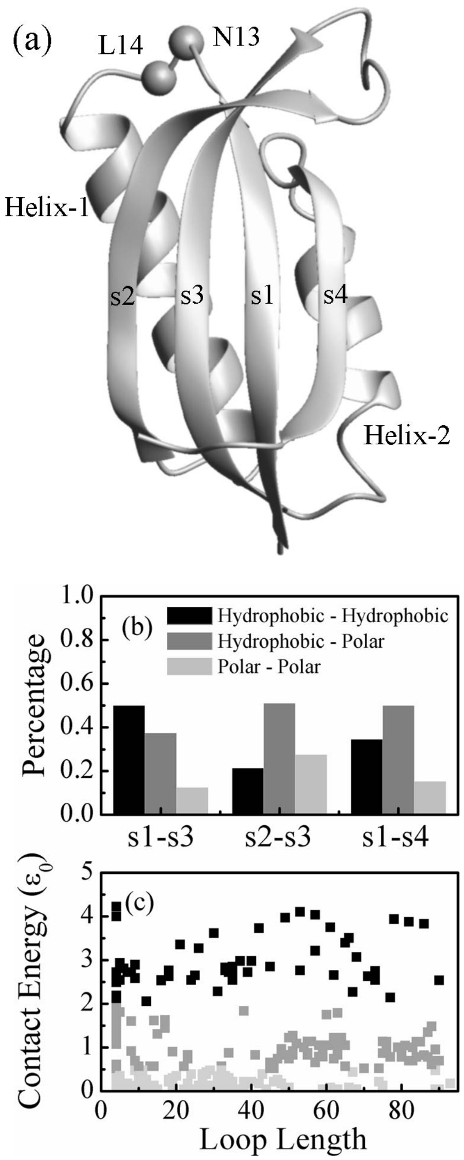

FIG. 1. (a) Native structure of the ribosomal protein S6. The figure is generated using the software MOLMOL [32]. (b) Distribution of hydrophobic and polar contacts in the three interfaces between β strands in S6. In total, eight kinds of hydrophobic residues are recognized from the Miyazawa-Jernigan matrix; they are Ile, Leu, Val, Cys, Phe, Tyr, Trp, and Met. (c) Strengths of contact energies (in units of ε_0) versus loop length, which is defined as the number of residues separating the contacted residues. The contacts are colored according to the strength of contact interactions. Those with strength below $0.5\varepsilon_0$ are colored light gray; those ranging from $0.5\varepsilon_0$ to $2\varepsilon_0$, gray; and those larger than $2\varepsilon_0$, black.

Physically, the vdW interaction contributing to the strength of contact interaction between two residues should be proportional to the number of heavy atom pairs within a cutoff distance. The coefficient of the vdW interaction then can be written as follows:

$$\varepsilon_{vdw}(i,j) = \varepsilon_1 n(i,j)/N, \tag{1}$$

where $n(i,j)$ is the number of heavy atom pairs between residues i and j. N is the averaged value of $n(i,j)$ on all native contacts; it normalizes the mean pairwise vdW interaction energy to ε_1.

The hydrogen-bond interaction plays a critical role in protein folding; however, it is not considered explicitly in the simplistic Gō model either. In this work, it is introduced into our model as follows:

$$\varepsilon_{\mathrm{H}\ bond} = \varepsilon_2 \frac{1}{2}\left[1 + \cos\left(\frac{\theta}{35}\pi\right)\right], \qquad (2)$$

where θ is the angle between the proton-donor bond and the line connecting donor and acceptor atoms. Following the idea of the Gō model, only the native hydrogen bonds are considered. A native hydrogen bond is defined when the proton-acceptor distance is less than 2.4 Å and the angle $\theta < 35°$ [18]. The expression smooths the values of $\varepsilon_{\mathrm{H}\ bond}$ between 0 (when $\theta=35°$) and ε_2 (when $\theta=0°$).

Hydrophobic interaction is of vital importance in the protein folding process. Unfortunately, due to its many-body nature it is difficult to map this interaction accurately in a minimalist model using additive interactions. However, we believe that a semiquantitative characterization of this interaction can greatly improve the model and is enough to describe the folding of S6, as will be shown in this work. Motivated by this assumption, we carefully checked the distribution of the hydrophobic and hydrophilic contacts in the native state of S6. Figure 1(b) shows such distributions within the three interfaces between four β strands. We can see that the hydrophobic contacts within the interfaces 1, 2, and 3 are 50%, 20%, and 35%, respectively. The extremely high percentage of hydrophobic contacts in interface 1 is worth investigating. Presumably, such a biased distribution of hydrophobic contacts in S6 will contribute greatly to its special folding mechanism. Furthermore, taking into consideration the different magnitudes of hydrophobicity of various residues and thus their different contributions to the hydrophobic interaction, we classify the residues Ile, Leu, Val as "strong hydrophobic residues" and the contacts between them as "strong hydrophobic contacts," according to the experimentally determined magnitude of hydrophobicity of each residue [19]. Careful study of the composition of hydrophobic contacts reveals that the percentage of strong hydrophobic contacts within interfaces 1, 2, and 3 is 25%, 20%, and 67%, respectively. These features will be considered when parametrizing the model. Besides the hydrophobic contacts within the three interfaces between the β strands, there are some long-range strong hydrophobic contacts discretely distributed between helices and strands, such as the contacts within the experimentally determined hydrophobic core (V6-L30 and I8-I26), and the gatekeeper contacts (E22 and V85) [12]. These contacts are also considered in our model. Thus, the coefficient corresponding to the hydrophobic interaction can be described as follows:

$$\varepsilon_{hydrophobic}(i,j) = h_k, \qquad (3)$$

where $k=1,2,3,4$. The parameters h_1–h_3 correspond to the native contacts within the three interfaces, respectively. h_4 corresponds to the native hydrophobic contacts discretely distributed between helices and strands. To reflect the many-body features of the hydrophobic interaction, we make a partition of the total hydrophobic energy of each interface to all the native contacts within the interface. Thus the values of the parameters h_1, h_2, and h_3 should be much lower than that of h_4. Although the folding cooperativity may be decreased by such a partition, we believe that the correctness of the predicted folding pathway should basically not be changed since it is determined by the energy-entropy interplay along the energy landscape. That this is really the case is indicated by comparing our results with experiments (see the following sections). Furthermore, h_1 and h_3 should be stronger than h_2 because interface 1 comprises the largest percentage of the hydrophobic contacts and interface 3 comprises the largest percentage of strong hydrophobic contacts. Thus, we should semiquantitatively weigh the hydrophobic intensities for the three interfaces.

Actually, such a weight among three kinds of interactions is relevant to many experimental results showing that the hydrogen bond is about five times stronger than vdW interactions, and the hydrophobic interaction energy is comparable with the hydrogen bond energy. Accordingly, the ratio of ε_2 to ε_1 is fixed to 5:1, and the interaction energy contributed by each hydrophobic contact is set around ε_2. By testing a set of h_k values without changing their qualitative relationship, we obtain a cooperative folding process. ε_1 and ε_2 are chosen as $0.4\varepsilon_0$ and $2\varepsilon_0$; the hydrophobic parameters h_1–h_4 are set as $0.75\varepsilon_0$, 0, $0.65\varepsilon_0$, and $2.5\varepsilon_0$, respectively. Although the hydrophobic parameters are somewhat arbitrary due to the many-body features and thus the complex nature of the hydrophobic effect, the parameters are chosen to reflect the relative hydrophobic strengths between the three β interfaces, the β-strand–helix, and the helix-helix interfaces. As will be seen later, such an assumption can greatly improve the performance of our model.

The distribution of overall contact energy versus loop length (i.e., the distance in sequence between two contacted residues) in our model is shown in Fig. 1(c). The plot can be divided into three areas by two loop lengths, 40 and 70. The areas I, II, and III mainly contain contacts within the interfaces 2, 1, and 3, respectively. The average contact energy in area I is $0.99\varepsilon_0$, while those in areas II and III are $1.26\varepsilon_0$ and $1.23\varepsilon_0$, respectively. This distribution is in accordance with the conclusion that interactions between residues with larger separation in sequence are stronger [2]. The bias of interaction energies comes from the hydrophobic factor, while neither vdW nor hydrogen bond interactions show apparent biased distributions.

It should be noted that the parameters in our model can hardly be uniquely determined, especially that for the hydrophobic interactions, due to their many-body nature. However, we believe that the major factors that affect the folding pathway are captured by correctly parametrizing the relative magnitudes of hydrophobic interactions within the three interfaces. It is also found that further refinement of these hydrophobic parameters can only improve the results slightly, possibly due to the limitation of the coarse-grained model itself. Therefore, although the possibility of other suitable parameters may exist, our conclusions drawn later are still robust.

C. Gō model

The energy function of our Gō model has a similar form as that in previous work and is

$$E(C,C_0) = \sum_{bonds} K_r(r-r_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{dihedral} K_\phi^{(n)}\{1+\cos[n(\phi-\phi_0)]\} + \sum_{i<j-3}\left\{\varepsilon(i,j)\left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10}\right] + \varepsilon_0(i,j)\left(\frac{\sigma_0}{r_{ij}}\right)^{12}\right\}. \quad (4)$$

Here, $E(C,C_0)$ stands for the total potential energy of conformation C with C_0 as its native conformation. r is the bond length between two neighbor residues in conformation C. θ and ϕ represent the bond angle formed by three subsequent residues and the dihedral angle formed by four subsequent residues in conformation C, respectively. r_0, θ_0, and ϕ_0 refer to the values of r, θ, and ϕ in the native conformation C_0. The last term is the contact energy between two residues i and $j>i+3$. Note that if residues i and j form a native contact, the 12-10 Lennard-Jones potential is used; otherwise the second repulsive potential is used. r_{ij} and σ_{ij} represent the distances between two residues i and j in conformations C and C_0, respectively. The parameters in the above equation are chosen as $K_r=100\varepsilon_0$, $K_\theta=20\varepsilon_0$, $K_\phi=\varepsilon_0$ for $n=1$ or $0.5\varepsilon_0$ for $n=3$, $\sigma_0=4$ Å. $\varepsilon(i,j)$ is the sum of ε_{vdW}, $\varepsilon_{\text{H}-bond}$, and $\varepsilon_{hydrophobic}$ if they exist between residues i and j.

Off-lattice molecular dynamics (MD) simulations are carried out using the Langevin equation:

$$m\ddot{r} = -\nabla_r E(C,C_0) - \gamma\dot{r} + \Gamma, \quad (5)$$

where γ is the friction coefficient and Γ is a random force that is Gaussian distributed and depends on temperature.

D. Data analysis

The simulation is performed by running several very long-time MD simulations at different temperatures, and the running is long enough to guarantee equilibrium. These trajectories are combined using the weighted histogram analysis method (WHAM) method [20] to calculate the thermodynamic properties such as the free energy and specific heat. The logic of the WHAM is to calculate the partition functions by solving the following equations self-consistently:

$$W(Q,E) = \frac{\sum_{k=1}^{R} N_k(Q,E)}{\sum_{j=1}^{R} n_j \exp(f_j - \beta_j E)}, \quad (6)$$

$$P_\beta(Q,E) = \frac{\sum_{k=1}^{R} N_k(Q,E)\exp(-\beta E)}{\sum_{j=1}^{R} n_j \exp(f_j - \beta E)}, \quad (7)$$

$$\exp(-f_j) = \sum_{Q,E} W(Q,E)\exp(-\beta_j E). \quad (8)$$

Here Q is the fraction of native contacts and is used as a reaction coordinate. It has been accepted in many works that, within the framework of Gō-like models, Q is a good enough reaction coordinate to characterize the folding process. β_j, n_j, f_j, and $N_k(Q,E)$ are the corresponding inverse temperature, number of conformations collected, free energy, and histogram on a (Q,E) grid of each MD run, respectively. $W(Q,E)$ is the density of states, and $P_\beta(Q,E)$ is the probability for conformations with their Q and E in the range of $Q \sim Q + \Delta Q$ and $E \sim E + \Delta E$ at temperature β.

The free energy projected on the reaction coordinate Q, or the so-called potential of mean force, is calculated by

$$F_\beta(Q) = -k_B T \ln \sum_E P_\beta(Q,E). \quad (9)$$

To investigate the folding pathway or the folding order of different regions of the native structure, the contacts are clustered according to their distances in space in the native state. For example, the contacts between strands 1 and 3 are treated as one cluster and referred to as Q_{13} hereafter. Such clusters of contacts are generally denoted by Q_{part}. The folding pathway can be described by following the change of $\langle Q_{part}\rangle$ as a function of Q, where the average $\langle Q_{part}\rangle$ is over the ensemble of conformations with the same Q. In a formula, it is

$$\langle Q_{part}(Q)\rangle = \frac{\sum_E Q_{part}(Q) w(Q_{part},Q,E)\exp(E/k_B T)}{\sum_E w(Q_{part},Q,E)\exp(E/k_B T)}, \quad (10)$$

where $w(Q_{part},Q,E)$ is the density of states, which is obtained in the same way as $W(Q,E)$, but with an additional dimension of Q_{part}.

The free energy can also be projected on two reaction coordinates to illustrate the detailed geometry of the free energy landscape. It can be done as follows:

$$F_\beta(Q_{part},Q) = -k_B T \sum_E \ln P_\beta(Q_{part},Q,E). \quad (11)$$

Φ-value analysis is a powerful technique to characterize the structure of transition states. Each residue has a Φ value that measures the involvement of this residue in the transition state according to the following equation:

$$\Phi \equiv \frac{-RT\ln(k_{mut}/k_{wt})}{\Delta\Delta G^0}, \tag{12}$$

where k_{wt} and k_{mut} are the folding rates of the wild-type protein and its site mutant, respectively. $\Delta\Delta G^0$ is the change of stability after the site mutation.

In theoretical calculations, Φ values are often calculated with free energy perturbation theory:

$$\Phi = \frac{\Delta\Delta G^{\ddagger}}{\Delta\Delta G^0} = \frac{\Delta G^{TS} - \Delta G^{U}}{\Delta G^{F} - \Delta G^{U}} = \frac{\ln\langle e^{\Delta E/RT}\rangle_{TS} - \ln\langle e^{\Delta E/RT}\rangle_{U}}{\ln\langle e^{\Delta E/RT}\rangle_{F} - \ln\langle e^{\Delta E/RT}\rangle_{U}}. \tag{13}$$

However, this theory is based on the assumption that the perturbation induced by a mutation is small and will not cause significant distortion of the free energy landscape, which may not be the case for S6. It has been demonstrated that S6 has a plastic transition state that can be greatly affected by mutations at certain sites [21]. In this work, considering that the Φ value also measures the fraction of native structure formed in the transition state ensemble around the mutation site i, it can be calculated by [22,23]

$$\Phi(i) = \frac{N_{TS} - N_U}{N_F - N_U}, \tag{14}$$

where N_{TS}, N_F, and N_U are the average numbers of contacts involving residue i in the transition state ensemble, native ensemble, and unfolded ensemble, respectively. A Φ value close to 0 means that few contacts involving residue i are formed in the transition states ensemble, whereas a value close to 1 means that most of them have been formed.

III. RESULTS

A. Free energy and specific heat profile

First, we compare the thermodynamic properties between the modified model and the simplistic model. Both models show the typical two-state folding behavior, consistent with the experimental observations (Fig. 2). However, the modified model has a higher cooperativity comparing with the simplistic model, illustrated by the slightly narrower distribution of unfolded states and the higher barrier for folding [Figs. 2(a) and 2(b)]. This feature can also be seen in the specific heat profile [Fig. 2(c)], which shows that the peak corresponding to the modified model is much higher and narrower. The higher cooperativity of the present model indicates that it is superior to the simplistic one because of the general assumption that the biased contact energy in S6 is used by nature to balance the biased entropy to increase the folding cooperativity. Clearly, the improvement of the cooperativity of our model is consistent with such an issue.

B. Folding pathway

Experimentally, the folding pathway of S6 has been studied by making a continuous Φ-value movie of the growth of the critical nucleus upon addition of denaturant. The folding process of S6 is as follows. At early time, the nucleation occurs around the V6-I8-I26-L30 cluster. Then, the large loop separating strands 2 and 3 closes in around F60, L48, and I52, docking with V9 after nucleation. After this, helix 2 forms secondary structure and docks to strand 1. At the final stage, the entropically disfavored strands 1 and 4 come together [24].

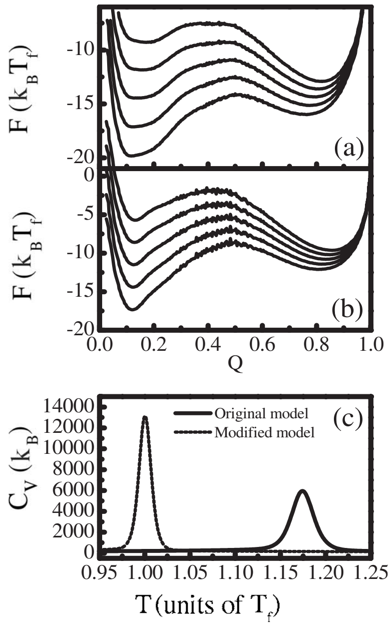

FIG. 2. Free energy curves at temperature from $1.02T_f$ (top) to $0.98T_f$ (bottom) in steps of $0.01T_f$ as a function of the reaction coordinate Q calculated using the WHAM [20], sampling at several temperatures near the folding temperature T_f of S6. (a) Free energy curves (in units of k_BT) calculated by the simplistic model. (b) Free energy curves calculated by the modified model. (c) Specific heat (in units of k_BT) calculated by both models. The solid line is for the modified model and the dotted line is for the simplistic model.

The folding pathways predicted by the simplistic and modified models are compared in Fig. 3(a). Several important structural clusters, including those located in the three interfaces and in the hydrophobic core, are used to characterize the folding process. The most important difference is that in the modified model the formation of the hydrophobic core between strand 1 and helix 1 is much accelerated, and at the same time the formation of the central hairpin (formed by strands 2 and 3) is slowed. The former feature is closer to the experimental observations, which show that this event occurs earlier in the overall folding process [24]. This improvement is due to the implementation of strong hydrophobic interactions in the core in the modified model. As for the central hairpin, its formation is very fast in the simplistic model due to the low entropy cost of this process. However, this advantage has been counteracted by the relatively low energy gain in the modified model, resulting in a low formation speed. It is also interesting to see that, after passing the transition state, the formation of interface 2 slows down and drops behind the formation of interface 3. This phenomenon is also consistent with experiments, which show that the Φ values of strand 3 have comparatively small factional values when the TS ensemble is shifted close to the native state.

To investigate the folding pathway in detail, we calculate the free energy landscape of S6 by projecting it onto two

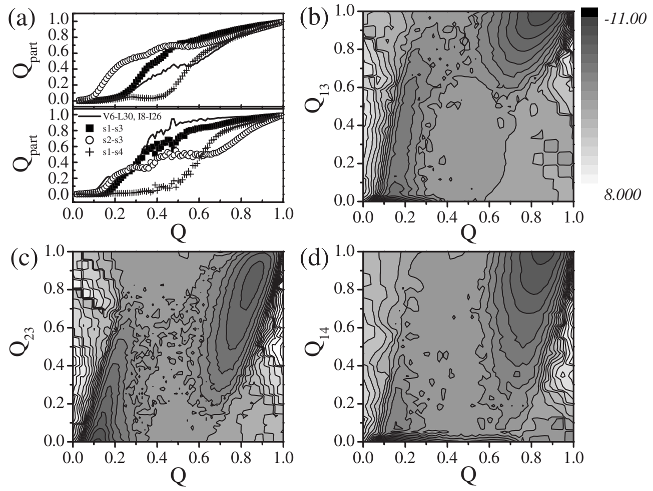

FIG. 3. (a) Probability of forming several important contact clusters as a function of Q, demonstrating the folding order of these clusters or the folding pathways. The upper panel is calculated from the simplistic model and the lower one is from the modified model. (b)–(d) Two-dimensional free energy contour plots for S6; darker gray corresponds to higher population. Q_{13}, Q_{23}, and Q_{14} denote the fraction of formed contacts within the interface between strands 1 and 3, between strands 2 and 3, and between strands 1 and 4, respectively.

reaction coordinates. Figure 3(b) shows the projection on Q and Q_{13} where Q_{13} represents the fraction of formed contacts within the interface between strands 1 and 3, i.e., interface 1. The unfolded basin of attraction extends to a comparatively high Q_{13} value, indicating that a large fraction of native structure in this interface has been formed in the unfolded state. This is due to the existence of a large fraction of hydrophobic contacts in this region and the nucleation-condensation folding mechanism of S6. Further investigations show that the formation of interface 1 starts around the cluster Y4-V6-M67-V65, which is entropy disfavored compared with other contacts in this interface. This illustrates the role of the biased contact energy in mediating the folding process.

Figure 3(c) shows the free energy landscape by projecting it onto Q and Q_{23} where Q_{23} is the fraction of formed contacts within the interface between strands 2 and 3, i.e., the central hairpin. The free energy landscape shows a very broad transition region between the unfolded and native basins of attraction, indicating that the central hairpin is highly heterogeneous in the transition state ensemble. This feature is consistent with the experimental observation of the plastic nucleus when subjected to permutations. The more alternative folding pathways or heterogeneities exist in the transition state ensemble, the larger the plasticity in response to free energy perturbations [25]. The observed heterogeneity can be attributed to the competition between entropy and energy. The low conformational entropy in the central hairpin favors fast folding, while the low value of the contact energy implemented in the model is disadvantageous to fast folding.

The free energy landscape projected on Q and Q_{14} is shown in Fig. 3(d). The shape of the unfolded basin resembles a narrow belt with Q_{14} lower than 0.05 and Q extending beyond 0.6. Therefore the interface 3 is totally unfolded in the unfolded states and forms very late in the overall folding process, as shown in Fig. 3(a).

By comparing the folding pathways between the simplistic and modified models, it can be seen that the formation order of structural clusters predicted by the modified model is much closer to the experimental observations and this improvement is indeed due to the biased contact energy that balances the loop entropy. Physically, this strategy of balancing entropy by biased energy will introduce heterogeneity into the transition state ensemble and parallel pathways into the overall folding process, as has been observed both in experiments [26] and our simulations. The agreement of this characteristic illustrates again the crucial role of biased energy in folding. Moreover, a detailed analysis of the contact energy shows that among several energetic factors the dominant one is the hydrophobic energy. It is the hydrophobic energy that accelerates the nucleation around Y4-V6-M67-V65 and counteracts the favorable entropy of the central hairpin. This indicates that the bricks that nature prefers to use to build the cooperativity architecture of S6 are the hydrophobic residues.

C. Φ-value analysis

The theoretical Φ values for up to 20 sites for wild-type S6 and its circular permutant $P^{13\text{-}14}$ are calculated using both the simplistic and modified models, as shown in Figs. 4(a) and 4(b). The general shape of the distribution of Φ values calculated by the simplistic model resembles a hill, i.e., the residues located at the middle part of the sequence have higher Φ values than those at the two termini. Such a polarized distribution is due to the low entropy cost of the formation of corresponding structures for those middle residues, for example, the central hairpin. For the modified model, however, the low entropy cost is compensated by the low energy gain, resulting in a diffuse distribution, as shown by Fig. 4(b). The correlation coefficient between our simulations and experiments [24] is much improved compared with that between the prediction of the simplistic model and experiments, i.e., 0.54 versus 0.29. Specifically, the major improvements lie in the residues V6 and I8 within the β strand 1 and A35 and V37 within strand 2. The former two residues are involved in the formation of the nucleus, which has been accelerated in our simulations based on the modified model, thus manifesting the high Φ values. In contrast, the Φ values for the latter two residues are greatly decreased and also become closer to the experimental values since the formation of interface 2 is delayed in the modified model.

The Φ values for the circular permutant $P^{13\text{-}14}$ are also calculated by using the two models and the results are shown in Figs. 4(c) and 4(d). In both models, the contact energies between residues are kept the same as in the corresponding model of the wild type. The discrepancy of Φ values between the simplistic Gō model and experiments is very large [Fig. 4(c)], clearly demonstrating that this model fails to reproduce the correct transition state ensemble and folding pathways. For the modified model, however, the prediction is

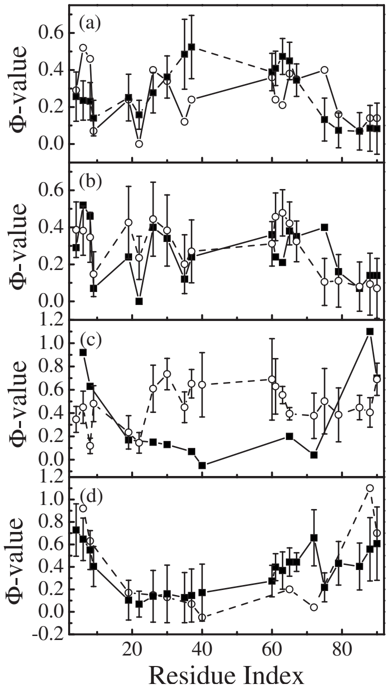

FIG. 4. Comparison of the Φ values calculated from the simplistic and modified models. Both are marked by open circles. The corresponding experimental values [24] are also plotted and marked by solid squares. (a) Φ values for wild-type S6 calculated from (a) simplistic model and (b) modified model. Φ values for $P^{13\text{-}14}$ calculated from (c) simplistic model and (d) modified model.

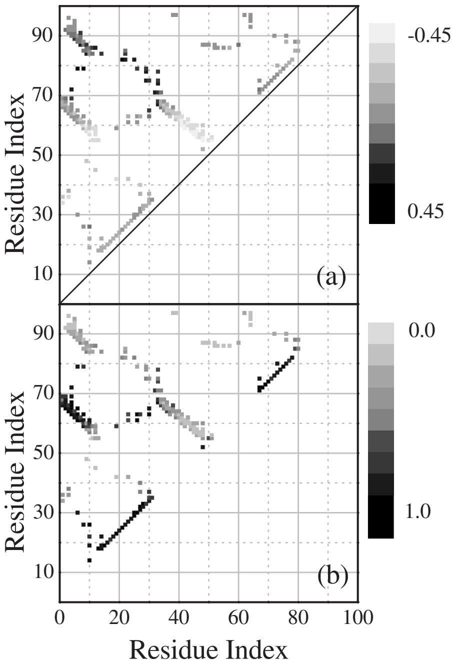

FIG. 5. (a) $f_{FF}-f_{UU}$ values of each contact for the wild-type S6 calculated by the modified model; (b) similar to (a) but only the f_{FF} values are shown.

much improved and the correlation coefficient between simulation and experiment increases from totally unrelated ($R=-0.26$) to generally conformed ($R=0.66$), as shown by Fig. 4(d). Furthermore, the distribution of Φ values is polarized and has a pattern similar to that of the experiments [2]. Specifically, the high Φ values ~ 0.7 [Fig. 4(c)] of residues 25–60, which largely correspond to the residues of the central hairpin, are decreased to below 0.2 in the simulations. This is due to the low interaction energies between the related contacts. Therefore, by comparing the Φ values calculated using two models and experiments, it can be seen again that for protein S6 the biased energy distribution plays significant roles in determining the transition state ensemble and folding pathways, and such a bias distribution has been characterized in our model.

We have also applied our model to other proteins that are intensively studied by the simplistic Gō model, such as chymotrypsin inhibitor 2 (CI2) and the SH3 domain. Our model shows only inconspicuous improvement on these proteins (the results are not shown here). Detailed analysis indicates that the percentage of hydrophobic contacts in these proteins is very low and the distribution shows no apparent bias. Thus, although our model is still applicable to these proteins, the simplistic model is good enough to describe their folding process.

D. Folding nucleus

To identify the folding nucleus, we follow the method used by Shakhnovich and co-workers [27]. A quantity $f_{FF}-f_{UU}$ is defined as the difference in frequencies of a contact appearing in two kinds of conformation, the FF and UU conformations. FF conformations are those located at the turning points of trajectories which start from the folded state and enter the transition state and then return to the folded state without reaching the unfolded state; correspondingly, UU is relevant to those trajectories starting from an unfolded state and returning to an unfolded state without descending to the native state. The contacts with high $f_{FF}-f_{UU}$ values are assumed as the key interactions that pull trajectories across the overall barriers and enter the native basin; thus they must belong to the folding nucleus.

Figure 5(a) shows the $f_{FF}-f_{UU}$ value of each native contact; the average value is −0.06 and the color is scaled from −0.45 to 0.45. Figure 5(b) shows the f_{FF} value only for each contact. The contacts with high values are those connecting the *N* termini of strand 2 and the *C* terminus of strand 3, those between helix 1 and helix 2, and those between strand 1 and helix 2. Considering that the nucleation site must be nativelike in the FF conformations, the contacts with high $f_{FF}-f_{UU}$ values but low f_{FF} values should be excluded [28]. By combining these two criteria, several nucleation sites can be identified. They are V6, I8, Y33, M67, and L79, with their corresponding contacts having $f_{FF}-f_{UU}$ values larger than 0.3 and f_{FF} larger than 0.6. These nucleation sites really coincide with those identified by experiments or by using evolutionary approaches [10]. If the $f_{FF}-f_{UU}$ threshold is lowered to 0.1, other nucleation sites, L30, Y63, and V65, can also be found. At the same time, some residues near the folding nucleus are also identified as nucleation sites, for example, N32 and D74. The success of our model in predicting most of the folding nucleus of S6 gives further support to the rationality of our model.

IV. CONCLUSION

In this work, by carefully analyzing the physicochemical properties of residues and the contact energies between them, we develop a C_α Gō-like model that balances the topological entropy by biased contact energies to study the special folding behaviors of ribosomal protein S6. The agreement of our model with experiments is much improved comparing with the topology-only simplistic Gō model. The improvement of our model illustrates that the contact energies are indeed biased in S6 and it is crucial to include both topological and energetic factors in the model, especially when studying such a type of protein. Furthermore, among several energetic factors, hydrophobic energy is the dominant one that counteracts with the loop entropy and mediates the folding pathway. It is the hydrophobic interactions that accelerates the entropy-disfavored nucleation and delays the formation of several entropy-favored structures. In this way the folding cooperativity is increased and a large plasticity is introduced into the folding nucleus.

The native structures of proteins are designed to perform biological functions, which, in some cases, may conflict with the requirement of folding cooperativity. The latter is important since otherwise the protein may undergo partial unfolding and then harmful aggregation, especially taking consideration of the crowded environment *in vivo*. Therefore, trying to balance functional requirements and folding cooperativity must be a very common phenomenon. Our study reveals a possible strategy or mechanism by which nature gets out of this dilemma. This strategy is almost definitely requisite for S6 due to its β-rich structure, which is apt to aggregate or form amyloid fibrils [29]. Other examples that may face the same dilemma include the acyl-coenzyme A-binding proteins [30] and the human Pin1 WW domain [31]. It would be interesting to see whether a similar strategy is used in these proteins.

The success of our model also highlights the importance of building physical models to study protein folding. The most accurate all-atom simulations are limited to short time and spatial scales while knowledge-based or experimental information-based minimalist models lack physicochemical origins in their parameters. In contrast, our model has several advantages; for example, the role played by each physical factor, such as that of the hydrophobic interaction, can be easily discerned by "knockout" simulations or by comparing the results with those from simplistic models. The obtained knowledge can extend our understanding of the code of protein folding and our ability to design new proteins or drugs. Our work presents such an effort in this direction.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grants No. 90403120, No. 10504012, and No. 10474041) and the National Basic Research Program of China (Grant No. 2006CB910302). The simulation was done at Shanghai Supercomputer Center.

[1] S. S. Plotkin and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **97**, 6509 (2000).

[2] M. Lingberg, J. Tångrot, and M. Oliveberg, Nat. Struct. Biol. **9**, 818 (2002).

[3] M. O. Lindberg, J. Tångrot, D. E. Otzen, D. A. Dolgikh, A. V. Finkelstein, and M. Oliveberg, J. Mol. Biol. **314**, 891 (2001).

[4] A. R. Fersht, Curr. Opin. Struct. Biol. **5**, 79 (1994).

[5] D. E. Otzen and A. R. Fersht, Biochemistry **37**, 8139 (1998).

[6] V. P. Grantcharova, D. S. Riddle, and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **97**, 7084 (2000).

[7] V. P. Grantcharova and D. Baker, J. Mol. Biol. **306**, 555 (2001).

[8] T. R. Weikl and K. A. Dill, J. Mol. Biol. **332**, 953 (2003).

[9] T. Ternström, U. Mayor, M. Akke, and M. Oliveberg, Proc. Natl. Acad. Sci. U.S.A. **96**, 14854 (1999).

[10] I. A. Hubner, M. Oliveberg, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **101**, 8355 (2004).

[11] S. Matysiak and C. Clementi, J. Mol. Biol. **343**, 235 (2004).

[12] A. D. Stoycheva, C. L. Brooks III, and J. N. Onuchic, J. Mol. Biol. **340**, 571 (2004).

[13] J. Chen, J. Wang, and W. Wang, Proteins: Struct., Funct., Bioinf. **57**, 153 (2004).

[14] H. Taketomi, Y. Ueda, and N. Go, Int. J. Pept. Protein Res. **7**, 445 (1975).

[15] C. Clementi, H. Nymeyer, and J. N. Onuchic, J. Mol. Biol. **298**, 937 (2000).

[16] J. Zhang, M. Qin, and W. Wang, Proteins: Struct., Funct., Bioinf. **59**, 565 (2005).

[17] W. X. Xu, J. Wang, and W. Wang, Proteins: Struct., Funct., Bioinf. **61**, 777 (2005).

[18] K. D. Berndt, J. Beunink, W. Schröder, and K. Wührich, Biochemistry **32**, 4564 (1993).

[19] J. Kyte and R. Doolite, J. Mol. Biol. **157**, 105 (1982).

[20] R. H. Swendsen, Physica A **194**, 53 (1993).

[21] M. Silow and M. Oliveberg, Biochemistry **36**, 7633 (1997).

[22] M. Vendruscolo, E. Paci, C. Dobson, and M. Karplus, Nature (London) **409**, 641 (2001).

[23] F. Ding, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, Biophys. J. **83**, 3525 (2002).

[24] D. E. Otzen and M. Oliveberg, J. Mol. Biol. **317**, 613 (2002).

[25] V. Muñoz, Nat. Struct. Biol. **9**, 792 (2002).

[26] J. Karanicolas and C. L. Brooks III, J. Mol. Biol. **334**, 309 (2003).

[27] N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, J. Mol. Biol. **296**, 1183 (2000).

[28] J. M. Borreguero, N. V. Dokholyan, S. V. Buldyrev, E. I. Shakhnovich, and H. E. Stanley, J. Mol. Biol. **318**, 863 (2002).

[29] S. Matysiak and C. Clementi, J. Mol. Biol. **363**, 297 (2006).

[30] B. B. Kragelund, P. Osmarkl, T. B. Neergaard, J. Schiødt, K. Kristiansen, J. Knudsen, and F. M. Poulsen, Nat. Struct. Biol. **6**, 594 (1999).

[31] M. Jager, Y. Zhang, J. Bieschke, H. Nguyen, M. Dendle, M. E. Bowman, J. P. Noel, M. Gruebele, and J. W. Kelly, Proc. Natl. Acad. Sci. U.S.A. **103**, 10648 (2006).

[32] R. Koradi, M. Billeter, and K. Wüthrich, J. Mol. Graphics **14**, 51 (1996).