PROTEINS: Structure, Function, and Bioinformatics 57:153–171 (2004)

Transition States for Folding of Circular-Permuted Proteins

Jie Chen, Jun Wang, and Wei Wang*
National Laboratory of Solid State Microstructure and Department of Physics, Nanjing University, Nanjing, China

ABSTRACT **To explore the role of entropy and chain connectivity in protein folding, a particularly interesting scheme, namely, the circular permutation, has been used. Recently, experimental observations showed that there are large differences in the folding mechanisms between the wild-type proteins and their circular permutants. These differences are strongly related to the change in the intrachain connectivity. Some results obtained by molecular dynamics simulations also showed a good agreement with the experimental findings. Here, we use a topology-based free-energy functional method to study the role of the chain connectivity in folding by comparing features of transition states of the wild-type proteins with those of their circular permutants. We concentrate our study on 3 small globular proteins, namely, the α-spectrin SH3 domain (SH3), the chymotrypsin inhibitor 2 (CI2), and the ribosomal protein S6, and obtain exciting results that are consistent with the available experimental and simulation results. A heterogeneity of the interaction energies between contacts for protein CI2 and for protein S6 is also introduced, which characterizes the strong interactions between contacts with long loops, as speculated from experiments for protein S6. The comparison between the folding nucleus of the wild-type proteins and those of their circular permutants indicates that chain connectivity affects remarkably the shapes of the energy profiles and thus the folding mechanism. Further comparisons between our theoretical calculated ϕ^{th} values and the experimental observed ϕ^{exp} values for the 3 proteins and their permutants show that our results are in good agreement with experimental ones and that correlations between them are high. These indicate that the free-energy functional method really provides a way to analyze the folding behavior of the circular-permuted proteins and therefore the folding mechanism of the wild-type proteins. Proteins 2004;57:153–171.** © 2004 Wiley-Liss, Inc.

Key words: topology; circular permutation; folding nucleus; free-energy functional; contact map

INTRODUCTION

Protein folding is a fundamental, important, and unsolved problem in molecular biology. Folding resembles a diffusive process on a rugged, funnel-like energy landscape.[1–8] The bottleneck affecting the folding dynamics, namely, the saddles on the free-energy landscape, is defined as the transition state.[1,2,4,7] Transition state acts as a watershed between the denatured state and the folded state, and behaves as the main barrier for the folding processes from the denatured state to the folded state.[4,5,7] Thus, studying the transition state is an essential task in understanding the protein folding mechanism, and it has become a hot topic in recent years.

Near to the native state, the transition state may be acted on by the energetic factors (encoded in amino acid sequence) and the entropic ingredients (originating from the denatured conformations).[1–10] Due to the rare appearance of the transition state in the equilibrium population, the information of the transition state is generally collected from mutational studies.[11] With a knowledge of the denatured state (considerably disordered) and of the folded state (highly ordered), the comparison between the nativeness of residues of a wild-type protein and its permutants provides us some insight into the features of the transition-state ensemble. The point mutations that vary the identities of a few amino acids perturb the energetics of proteins but keep the topology of the native structure unchanged.[12,13] On the contrary, the circular permutations, implemented by connecting two ends of a protein, change the connectivity of a protein chain.[13] For the circular permutants, various microscopic states are totally rearranged to produce a new free-energy landscape. Experimenters have discovered that the transition states show distinguishable differences between the proteins before and after the circular permutations.[12–18] Contact with the same pair of residues may play different roles in different transition states. Theoretical simulations have also illustrated results similar to those of experiments.[19,20] Recently, Clementi et al.[19] concluded that the changes in native-state topology induce different folding mechanisms for the wild-type proteins and their circular permutants. They also argued that the entropy plays an important role in the folding of the circular permutants. However, how do the entropy and the connectivity of the proteins affect the transition states of the circular-permuted proteins? Is it possible to study the effects of the entropy with an analytical scheme? A theoretical discussion would be

Grant sponsor: Foundation of NNSF; Grant numbers: 90103031, 10204013, 10074030, 10021001. Grant sponsor: Nonlinear Project; Grant number: 973.

*Correspondence to: Wei Wang, Physics Department, Nanjing University, Nanjing 210093, China. E-mail: wangwei@nju.edu.cn

Received 31 July 2003; Accepted 11 March 2004

Published online 11 June 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20175

instructive and necessary for understanding the experimental results of the circular permutations.

Here, based on a free-energy functional method developed by Plotkin et al.,[7,8,21–23] we present an approach on the folding behavior of the circular-permuted proteins. The free-energy profiles of 3 circular-permuted proteins, namely, the α-spectrin SH3 domain (SH3),[24,25] the chymotrypsin inhibitor 2 (CI2),[13] and the ribosomal protein S6,[14] are obtained by minimizing the functional under the Gō-like interactions.[26] The heterogeneity of the transition state ensemble is described using a distribution of various native contacts. Our results show that the architecture of the transition states from the free-energy functional theory is consistent with the findings by both experiments and simulations. The comparison between the folding nucleus of the wild-type proteins and those of their circular-permuted proteins indicates that the connectivity affects remarkably the shape of the energy landscape, and thus the folding mechanism. The free-energy functional method provides a way to analyze the folding behavior of the circular-permuted proteins.

METHODS

Free-Energy Functional

The energetic features of natural proteins are believed to be well optimized by longtime evolutionary selection,[1–6,9,26,27] and it has been argued that a simple and reasonable model of interactions between amino acids is the Gō-like model, which favors the native contacts energetically.[26] For the entropic component of the free-energy functional, the polymeric nature of the protein chain suggests that the loop between two residues is directly related to the probability for forming a contact between them. Consequently, relative to coarse graining, a protein system can be described by a collection of its native contacts with energies $\{\epsilon_i\}$ and corresponding loop lengths $\{l_i\}$.[7,8,22,23] Here, the contact energy ϵ_i (< 0) ensures the stability of the native structure and includes the solvent-averaged effect of contact i. $l_i >$ is the sequential length pinched off by contact i. Two residues, α and β, are defined as contacting each other when the spatial distance between their C_α atoms is smaller than 8.0 Å, and at the same time their sequential distance along the chain is larger than 3 residues (e.g., $\alpha > \beta + 3$). Note that if a small cutoff distance between the C_α atoms, say 7.0 Å, is used, there will be a smaller number of contacts in the contact map. However, from our calculation, the variance of contacts has less effect on the features of transition state [i.e., the main features of the islands on the contact maps (see the Results and Discussion section)]. A quantity Q, defined as the overall fraction of the native contacts, is used to describe the similarity of a given conformation to the native conformation. It is known that various contacts contribute to the global similarity differently. For an ensemble of Γ conformations with various values of nativeness Q relating to the folding pathways, there is a subset C of conformations with a certain value of Q. Then, a probability of the ith contact in this subset C is defined as

$$Q_i(Q) = \frac{\sum_j^\Gamma q_j^i \delta_j^C}{\sum_j^\Gamma \delta_j^C},$$

where j represents the jth conformation in the ensemble Γ. $q_j^i = 1$ if the ith contact is observed in the jth conformation, and $q_j^i = 0$ otherwise. $\delta_j^C = 1$ if the jth conformation is in the subset C (i.e., with its nativeness as Q), and $\delta_j^C = 0$ otherwise. Q_i indicates the degree of nativeness for the occurrence of the ith contact in the ensemble of states with global similarity Q. Thus, a distribution $Q_i(Q)$ can be used to describe the heterogeneities of various contacts in the ensemble with a fixed value of Q. Note that these degrees of nativeness satisfy the relation

$$\sum_i^M Q_i = MQ, \tag{1}$$

where M is the number of the native contacts. Eq. (1) is a constraint for the distributions of Q_i. Using the degrees of nativeness of the contacts, the folding processes can be described using a statistical picture. The degree of nativeness of the ith contact at equilibrium is defined as $Q_i^\dagger(Q)$, which is derived from the minimization of the free-energy functional. Compared with the usual characterization of the transition states of folding, the value of $Q_i^\dagger(Q^\ddagger)$ for the transition-state ensemble (with $Q = Q^\ddagger$) is related to the φ value.[23] Therefore, the study on the distribution of these degrees of nativeness of contacts provides us a way to figure out the structure of the transition state ensemble and further enables us to compare the results of different mutations. In this work, we use a variational approach to search for the optimal distribution, $\{Q_i^\dagger\}$, which minimizes the free-energy functional. With this scheme, the effect of the circular permutations can be characterized systematically.

The free-energy functional serves as the basic starting point for our studies. We employ the free-energy functional theory developed by Plotkin et al.[7,8,21–23] as follows. With the optimal features of natural proteins, native interactions are considered to play a more important role than non-native interactions in determining the folding mechanism of minimally frustrated proteins, such as the Gō model. Thus, the non-native interactions are treated as an averaged background field that contributes to the system as an energy dispersion, while the native interactions are retained. The energy of the system is obtained by averaging the Hamiltonian[7,8,21]

$$\mathcal{H} = \sum_{\alpha<\beta} [\varepsilon_{\alpha\beta}^N \Delta_{\alpha\beta} \Delta_{\alpha\beta}^N + \varepsilon_{\alpha\beta} \Delta_{\alpha\beta} (1 - \Delta_{\alpha\beta}^N)] \tag{2}$$

over a Gaussian distribution of the non-native interactions with variance b^2,

$$P(E|E_N, \{\Delta_{\alpha\beta}\Delta_{\alpha\beta}^N\}) = \langle \delta[E - \mathcal{H}\{\Delta_{\alpha\beta}\}] \delta[E_N - \mathcal{H}\{\Delta_{\alpha\beta}^N\}] \rangle_{\text{n-nat}}. \tag{3}$$

Here, the summation runs over all residue indices. The function $\Delta_{\alpha\beta} = 1$ or 0, respectively, if residues α and β do or do not contact with each other, and $\Delta_{\alpha\beta}^N = 1$ or 0, respec-

tively, if these contacts do or do not exist in the native configuration. $\varepsilon^N_{\alpha\beta}$ and $\varepsilon_{\alpha\beta}$ are the energies of the native and non-native contacts, respectively. Eq. (2) describes the energetics of the model protein system. After the averaging in Eq. (2), the native interaction $\varepsilon^N_{\alpha\beta}\Delta_{\alpha\beta}\Delta^N_{\alpha\beta}$ is written in a single index for simplification; that is, the native interaction is written as $\Sigma_i\, \varepsilon_i \mathscr{L}_i$, where $\mathscr{L}_i \in \{\mathscr{L}_{\alpha\beta} = 1, \alpha = 1, N,$ and $\beta = 1, N\}$ with $\mathscr{L}_{\alpha\beta} = \Delta_{\alpha\beta}\Delta^N_{\alpha\beta}$, and $\mathscr{L}_i \equiv 1$ if the ith contact is a native contact formed between the two residues α and β. At the same time, the non-native interactions $\varepsilon_{\alpha\beta}\Delta_{\alpha\beta}(1 - \Delta^N_{\alpha\beta})$ are averaged and are found to be $Mb^2(1 - Q)$. Generally, the energies of the native contacts $\{\epsilon_i\}$ vary from contact to contact. The heterogeneity of these energies may relate to the type and structural features of the residues as shown by the statistical potentials[28] and also to different features of secondary and tertiary interactions. For the sake of simplification, all the native contacts could be assumed to be energetically equivalent for exploring the role of entropy and chain connectivity in protein folding. However, it is worth noting that it may improve the modeling to some extent to employ some certain distributions of contact energies. A recent work by Oliveberg and colleagues[14] showed that a biased dispersion of the contact energies has a dramatic effect on the transition state of small protein S6 and its circular permutants.

For the entropic component of the free-energy functional, a simplification related to the coordinate Q is introduced. The entropy functional is divided into two parts: one related to the partition of the contacts in structural ensemble with coordinate Q (which corresponds to the multiplicity of the contact pattern), and the other depicting the entropic decrease due to closure of the contacts from the coil (which considers the effect of polymeric connectivity). Here we take similar notations to those used by Plotkin et al., namely, $\mathscr{S}_{\rm route}$ for the entropy due to the multiplicity of the contact pattern, and $\mathscr{S}_{\rm bond}$ for the entropy loss for the formation of different contacts. Based on a mean-field consideration of the correlations between different native contacts, the entropy functional $\mathscr{S}_{route}$ is given as[21]

$$\mathscr{S}_{route}(\{Q_i\}|Q) = -\lambda(Q)\sum_{i=1}^{M}[Q_i \ln Q_i + (1 - Q_i)\ln(1 - Q_i)], \tag{4}$$

where the prefactor $\lambda(Q) \equiv 1 - Q^\alpha$ originates from the reduction of the number of combinatory states by chain topology, since the residues connected by a chain have less mixing entropy than when they are free. The index $\alpha > 0$ decreases as system size N increases, since for a large system, the more part of the chain is buried, the stronger the constraint by surrounding contacts.[22,23] In our study, we set $\alpha = 1$. Due to the volume variation during folding, the entropy loss for forming a contact depends on the detailed structural features, that is, not only on the loop length related to the contact (which describes the polymeric connectivity) but also on the distribution $\{Q_i(Q)\}$ of formation probabilities of other contacts (which define the interactive environment). Neglecting the specific influence of other contacts, the entropy loss s_i to form a contact i may take the form $s_i = s_i\,[l_i, \{Q_i(Q)\}]$. The additivity of the entropic term s_i is also a kind of simplification, which is suggested from some qualitative analysis.[29] Thus, the entropy loss of the model protein system during a change from the state $\{Q_i(Q_0)\}$ to the state $\{Q_i(Q_f)\}$ would have the form

$$\mathscr{S}_{bond}(\{Q_i(Q_f)\}|\{Q_i(Q_0)\}) = \sum_i \int_{Q_0}^{Q_f} DQ_i(Q)s_i(l_i, \{Q_i(Q)\}). \tag{5}$$

The integrations are performed along an M-dimensional path specified by $\{Q_i(Q)\}$. The dependence of the term s_i on the state variables l_i and $\{Q_i\}$ is also consistent with the requirement that the contact-closure entropy would be a thermodynamic potential determined by the initial and final states. As further progress, the entropy s_i in Eq. (5) is approximated as the Flory form

$$s_i(l_i, \{Q_j\}) = \ln(a/l_{\rm eff})^{3/2} \tag{6}$$

with an effective loop length $l_{eff}(l_i,Q)$. The fluctuation of $\{Q_i\}$ in s_i is neglected as a mean-field approximation. As a result, the entropy loss to form the state $\{Q_i\}$ from the coil state $(Q = 0)$ is

$$\mathscr{S}_{bond} = \mathscr{S}_{MF}(Q, \bar{l}) - \frac{3}{2}M\langle \delta Q \delta \ln l\rangle, \tag{7}$$

where the $\langle \ldots \rangle$ means an ensemble average in subset C of the ensemble Γ.

Together with Eqs. (4) and (7) as the total entropy $\mathscr{S} = \mathscr{S}_{route} + \mathscr{S}_{bond}$, the free-energy functional has the form[21]

$$\mathscr{F} = \sum_{i=1}^{M}[\varepsilon_i \mathscr{L}_i - T\mathscr{S}(\{\mathscr{L}_i\}|Q)] - \frac{Mb^2}{2T}(1 - Q). \tag{8}$$

For each value of Q, there are many values of the free-energy $\mathscr{F}$ due to many distributions of Q_i. Among them, the most probable distribution is noted as $\{Q_i^\dagger\}$. It is easy to find out that the most probable distribution $\{Q_i^\dagger\}$ has the form of a Fermi distribution by minimizing the above free-energy functional Eq. (8) for a fixed Q as

$$Q_i^\dagger(Q, \{\varepsilon_i\}, \{l_i\}) = \left\{1 + \exp\left[\frac{1}{\lambda T}\left(\mu + \varepsilon_i - Ts_i(l_i, Q) - T\lambda'\langle s^0_{route}\rangle + \frac{b^2}{2T}\right)\right]\right\}^{-1}, \tag{9}$$

where the Lagrange multiplier is $\mu \sim -(1/M)\,\delta\mathscr{F}/\delta Q$. Using Eq. (7), we can calculate the entropy loss for contact i with native loop length l_i, which represents the native topology of the protein. Then the entropy loss of the contact is substituted into Eq. (9), together with the contact energy. Thus, the distribution of contact probabilities can be evaluated easily with simple calculations. For different values of Q, the distribution $\{Q_i^\dagger\}$ varies; consequently, the

free-energy $\mathcal{F}[Q_i^\dagger(Q)]$ along the folding progress could be outlined.

By varying the reaction coordinate Q in the free-energy functional, the values of Q that maximize the free energy $\mathcal{F}[Q_i^\dagger(Q)]$ of the wild-type protein and their circular permutants during the folding process are obtained. These values of Q approximately correspond to the position of the barrier between the native state and the denatured state (i.e., the location of the transition states).[7,8,21,30,31] We note them as $Q^\ddagger$ in the following. In our study, the values of $Q^\ddagger$ are found to be around 0.65 for both the wild-type protein SH3, CI2, and their permutants, while for protein S6 and its permutants, the value of $Q^\ddagger$ is about 0.45. With the same energy for each contact, the energy term in the free-energy functional is determined explicitly from the reaction coordinate Q. However, the entropy loss changes significantly due to the topological changes, as mentioned above.

In short, for a given value of Q, we get the most probable distribution $\{Q_i^\dagger\}$ by minimizing the free-energy functional [see Eq. (8)], and then all the most probable distributions [see Eq. (9)] are obtained for various values of Q. Among the most probable distributions, we get the transition state by finding the value of Q that makes the free-energy functional Eq. (8) be maximal. This value of Q describes the transition state and is defined as $Q^\ddagger$.

Generally, the contacts that have been well formed i.e., with large values of $Q_i^\dagger(Q^\ddagger)$, are important parts of the transition state and are believed to be essential for the stability of the folding nucleus, and thus to be helpful for further formation of the native structure. Consequently, these contacts are collected to describe the structural features of the transition state, which are largely populated in the folding nucleus. As a result, some characteristics of the transition-state ensemble are deduced with the minimizing procedures on the functional. Since various proteins and their circular permutants have distinct distributions $\{Q_i(Q)\}$ due to different folding mechanisms, to calculate the distribution of $\{Q_i^\dagger(Q^\ddagger)\}$ using Eq. (9) is our main task in this work.

Examples of the free-energy profiles for protein CI2 are demonstrated in Figure 1 for 3 temperatures, namely, $T_1 < T_f$, $T_2 \sim T_f$, and $T_3 > T_f$, with T_f being the folding transition temperature. For other proteins and their circular permutants, the shapes of the free-energy profiles are similar (data not shown). Here, T_f is identified as 0.58. It is lower than that found by simulations $T_f \approx 1$[19] due to the difference between the energy functions used here and in the simulations. In this work, the non-native interactions are treated as an averaged background field to derive the free-energy functional, and the interaction energies of native contacts are all set as -1 except when they are mentioned specifically with different values. In addition, the value of b^2 is taken to be 0.15.

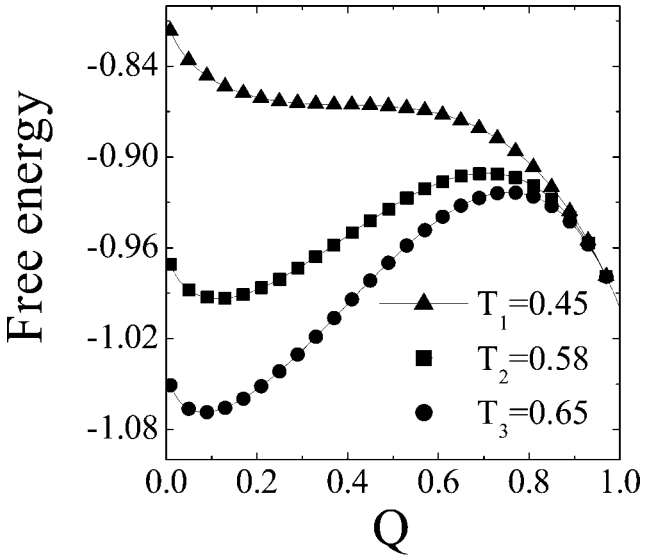

Fig. 1. Free-energy profiles for protein CI2 at 3 temperatures: $T_1 = 0.45 < T_f$, $T_2 = 0.58 \sim T_f$, and $T_3 = 0.65 > T_f$. The free energies are scaled by their values at the native state. The values of contact energies are set as $\epsilon_i = -1.0$.

Circular-Permuted Proteins

A circular permutation works by joining the N- and the C- termini of a protein with a peptide linker or chemical linkage, and at the same time, disrupting the peptide bond of the original protein at a certain given position.[12,13,20] In general, the circular permutation works well when the N- and C- termini of the protein are spatially close to each other, which makes the linker or linkage easy to manipulate. Here, 3 small proteins meet this constraint, namely, protein SH3 [Protein Data Bank (PDB) entry: 1bk2],[24] protein CI2 (PDB entry: 1ypc),[13] and protein S6 (PDB entry: 1ris),[14] and were chosen for our studies. A series of circular permutations for protein SH3 are performed by joining residues 1 and 57, and disrupting the bond between residues 14 and 15, 38 and 39, or 42 and 43, respectively. In the following, the corresponding permutants are referred to as *perm* 14–15, *perm* 38–39, and *perm* 42–43,[19] respectively (as shown in Fig. 2). Similarly, we perform the circular permutation for protein CI2 by joining residues 3 and 63 and disrupting the bond between residues 40 and 41, and refer to this permutant as *perm* 40–41 (as shown in Fig. 3). For protein S6, we join the two terminal residues 1 and 97, and disrupt the peptide bond between residues 13 and 14.[14] We obtain *perm* 13–14 (see Fig. 4).

We show the distributions of loop lengths versus the contact index for the 3 wild-type proteins and their permutants in Figures 5, 6, and 7, respectively. From these figures, we can see that due to the cutting and connecting, the loop lengths of the contacts near the cutting sites become longer and the loop lengths of the contacts near the connecting sites become shorter when compared with those in the cases of the wild-types. For *perm* 14–15 of protein SH3, the changes of the loop lengths occur only for contacts $i < 70$, that is, around the two termini [see Fig. 5(b)]. However, for both permutants of *perm* 38–39 and *perm* 42–43, the loop lengths change dramatically for all the contacts, and most of the contacts with long loops become short [see Fig. 5(c and d)]. For protein CI2, we can

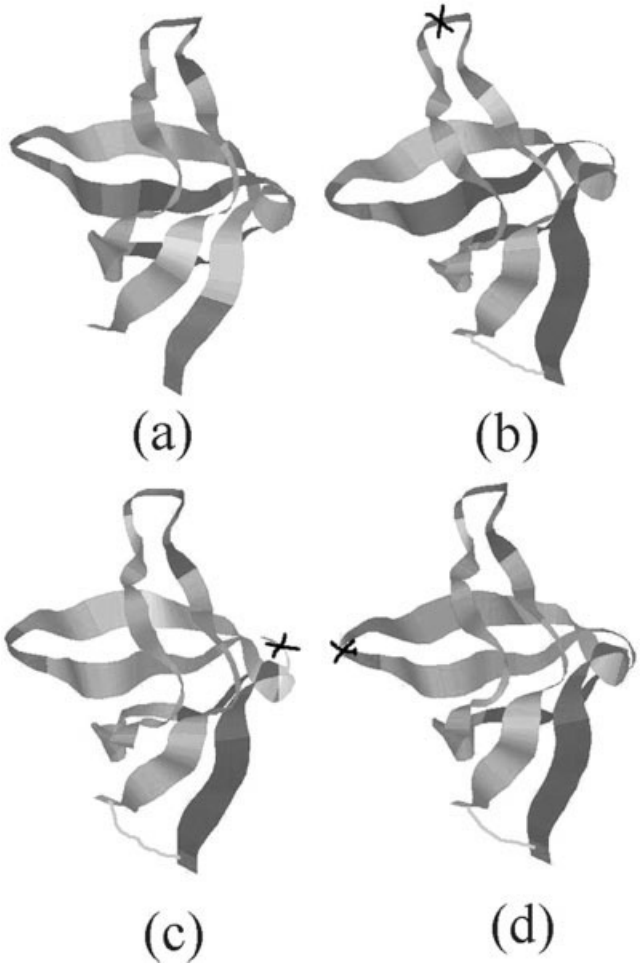

Fig. 2. The degree of nativeness of residues, P_α, for the wild-type protein SH3 and its different permutants in the transition state defined by $Q^\ddagger = 0.65$. The values of all contact energies are set as $\epsilon_i = -1.0$. The grayscale is the same as in Figure 8. (**a**) wild-type; (**b**) *perm* 14–15; (**c**) *perm* 38–39; (**d**) *perm* 42–43.

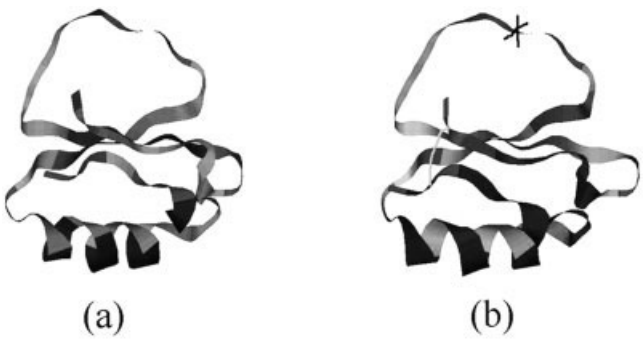

Fig. 3. The degree of nativeness of residues, P_α, for the wild-type protein CI2 and its permutant in the transition state defined by $Q^\ddagger = 0.65$. The values of all contact energies are set as $\epsilon_i = -1.0$. The grayscale is the same as in Figure 9. (**a**) wild-type; (**b**) *perm* 40–41.

also see that the loop lengths of contacts from $i = 1$–40 around the termini of the wild-type protein CI2 become short completely, while the loop lengths of contacts from $i = 60$–90 become longer (see Fig. 6). For *perm* 13–14 of

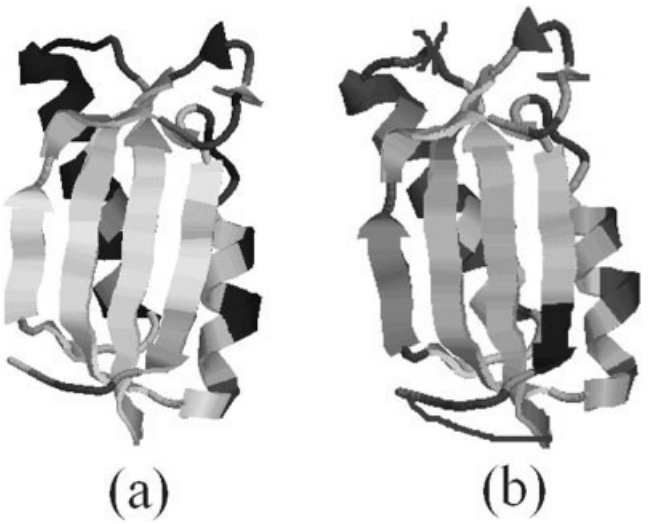

Fig. 4. The degree of nativeness of residues, P_α, for protein S6 and its permutant in the transition state defined by $Q^\ddagger = 0.45$. The values of contact energies are set as $\epsilon_i = -1.0$ for the loop lengths $l_i \leq 46$ and $\epsilon_i = -4.0$ for $l_i > 46$. The grayscale is the same as in Figure 10. (**a**) wild-type; (**b**) *perm* 13–14.

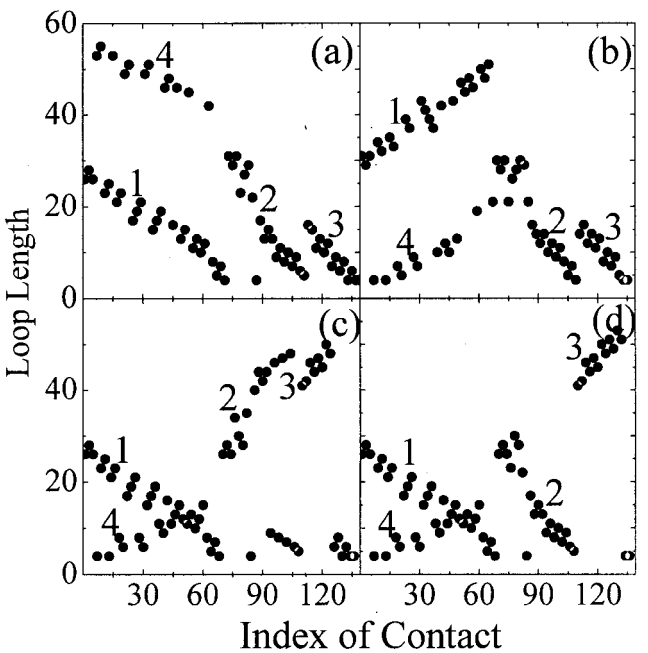

Fig. 5. Distribution of contact loop lengths for protein SH3 and its permutants. Indices 1, 2, 3, and 4 indicate the 4 islands, respectively (see definition of island in text). (**a**) wild-type; (**b**) *perm* 14–15; (**c**) *perm* 38–39; (d) *perm* 42–43.

protein S6, we can see that all the contacts with loop lengths longer than 46, which are formed between β_1 and β_2, β_1 and β_4, respectively, change into shorter ones. Actually, the change of loop length of a contact indicates the change of the role of the contact. For example, a nonlocal contact may become a local one if its loop length become small dramatically, or vice versa.

Previously, there were some experimental and simulational studies on how the circular permutation influences

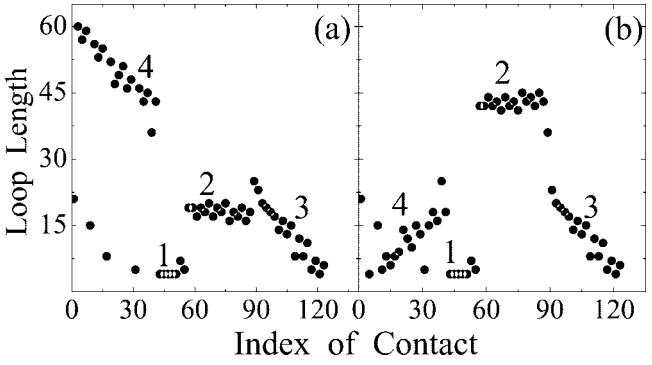

Fig. 6. Distribution of contact loop lengths for protein CI2 and its permutant. Indices 1, 2, 3, and 4 indicate the 4 islands, respectively (see definition of island in text). (**a**) wild-type; (**b**) *perm* 40–41.

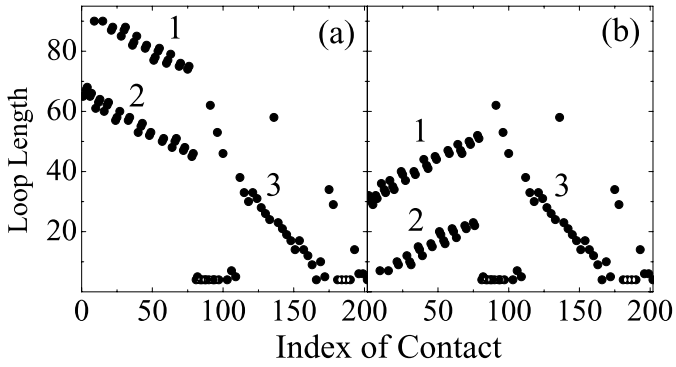

Fig. 7. Distribution of contact loop lengths for protein S6 and its permutant. Indices 1, 2, and 3 indicate the 3 islands, respectively (see definition of island in text). (**a**) wild-type; (**b**) *perm* 13–14.

the native states and folding rates of proteins.[13–20] Although conflict still exists as to whether native states are affected by the circular permutations, it is found that almost all the results derived for proteins SH3 and CI2 from experiments and simulations indicate no changes in their native conformations.

RESULTS AND DISCUSSION

On the level of mean-field theory, the folding pathways can be characterized with the local order parameter Q_i. In a space with coordinate $\{Q_i\}$, different pathways indicate different folding processes. By comparing various pathways related to different permutations, the effects of the chain architectures can be illustrated. The information from these comparisons serves as the basic starting point of our analysis.

Degree of Nativeness of Residues

To obtain an intuitive picture of the folding processes, we investigate the degrees of the nativeness of the residues, namely, the degrees of the residues approaching to the native state, for 3 proteins defined as

$$P_\alpha = \sum_\beta \mathcal{L}_{\alpha\beta} Q_{\alpha\beta} / \sum_\beta \mathcal{L}_{\alpha\beta}, \qquad (10)$$

where α and β are the indices of the residues, and $\mathcal{L}_{\alpha\beta}$ defines the native contact map as declared in the Methods section. The degrees of the nativeness of the residues in the transition state ensemble are calculated using Eq. (10), as shown in Figures 2, 3, and 4, with different grayscales representing different degrees of the nativeness of residues in the native state. For the circular permutants, it is found that some residues change their roles in the transition state, resulting in a rearrangement of the grayscaled spots. The results demonstrated for protein SH3 in Figure 2 show that the transition states of 3 permutants are considerably different. Figure 3 illustrates the transition state of protein CI2 and its permutant. It is clearly seen that the changes of the transition state of the permutant with respect to transition state of its wild-type are not so obvious as those in the case of protein SH3. These are consistent with the experimental observations.[13,24] Especially in the case of *perm* 38–39 for protein SH3, the sites with dark gray shift from the distal β-hairpin to the C-terminal strand. Thus, the folding nuclei of the transition states are changed thoroughly according to the large change in the chain connectivity. For protein S6, Figure 4 shows that the globally diffuse nucleus in the transition state for the wild-type protein S6 changes to a polarized one for the circular permutant of protein S6.[14] All these detailed features are discussed in the following sections. In addition, since P_α is actually related to the ϕ value of residue α when the system is in the transition state (i.e., $Q = Q^\ddagger$), a quantitative comparison between the ϕ values of experimental observations and our theoretical calculations can be made and will be presented in the Comparison With Experiments section.

Degree of Nativeness of Contacts

More quantitatively, the degrees of the nativeness of the contacts are shown in Figures 8, 9, and 10, respectively. Note that the temperature used in our calculations is $T = 0.58$, which corresponds to the folding transition temperature T_f. The heterogeneities of the nativeness of the contacts in the transition states of 3 proteins and their permuted proteins are clearly illustrated.

For the wild-type SH3 domain, the well-formed contacts, namely, those having large values of Q_i, are all near the diagonal [see Fig.8(a)]. This observation is in agreement with the experimental results.[12,24] Clearly, there are 3 main islands representing the interactions in the reactive loop (*RT loop*), between loop 2 and β_4, and between β_3 and β_4, respectively. In the following, these islands are referred to as *island*-1, *island*-2, and *island*-3, and are marked as 1, 2, and 3, respectively, in Figure 8(a). Obviously, the contacts in the island far from the diagonal, namely, *island*-4, are poorly formed. *Island*-4 (marked as 4) corresponds to the contacts around the N- and C-terminal regions, which have been discussed and, we concluded, were not as ordered in the transition state as they are in the native state.[24,25,32] The above observations imply that the contacts with short sequential distances (i.e., the local interactions) are more important than those with long loops in the SH3 domain for the folding kinet-

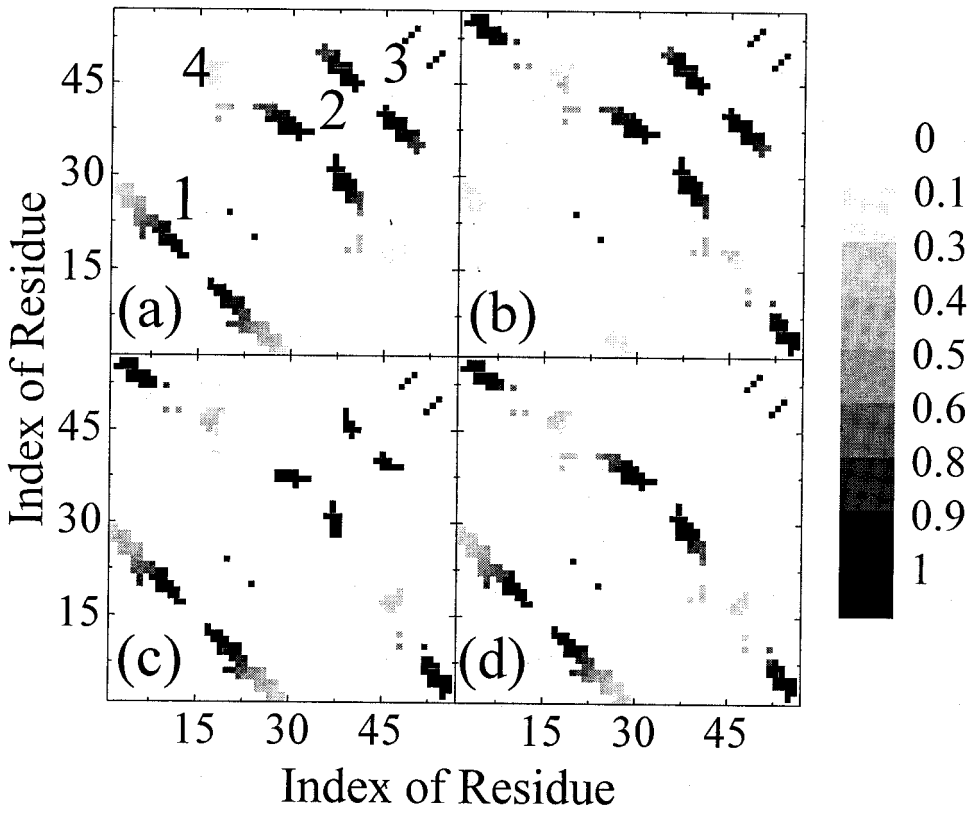

Fig. 8. The degrees of nativeness of contacts for protein SH3 and its permutants in the transition state shown by the contact map. The grayscale indicates the values of the degrees. The values of contact energies are set as $\epsilon_i = -1.0$. (**a**) wild-type; (**b**) *perm* 14–15; (**c**) *perm* 38–39; (d) *perm* 42–43.

ics.[33] Note that in our case, the local interactions actually are short-range interactions only for contacts $l_i > 4$.[19]

However, different permutants show different features of the contacts when compared with those in the wild-type case. The poorly formed *island*-4 becomes well formed due to the connection of the N- and the C- termini for all 3 permutants [see Fig. 8(b–d)]. Meanwhile, the effect of cutting at different sites varies distinctly. For *perm* 14–15, the contacts of *island*-1 become weak because of the cutting in the *RT loop*, while the contacts around the other two islands show no changes [see Fig. 8(b)]. In the case of *perm* 42–43, *island*-3 is totally destroyed, and *island*-1 and *island*-2 exhibit no changes [see Fig. 8(d)]. In comparison with *perm* 14–15, *perm* 42–43 shows no changes in *island*-1 because of the conservation of the *RT loop*. At the same time, in both *perm* 14–15 and *perm* 42–43, the loop lengths of the contacts in *island*-2 exhibit no changes either (also see Fig. 5). These effects are also observed in experiments,[19,24,25] for example, the lesser nativeness of *island*-1 for *perm* 14–15 due to the loss of conserved hydrogen bonds in the *RT loop*. The most interesting results are for *perm* 38–39. In the transition state, both *island*-2 and *island*-3 in *perm* 38–39 have much lower values of $Q_i \sim 0.3$ than those in the wild-type protein [see Fig. 8(c)]. This is due to the cutting near both islands.

For protein CI2 and its *perm* 40–41, the variation of the degrees of nativeness of contacts in *island*-1 [marked as 1 in Fig. 9(a)] in the transition state is not significant [see Fig. 9(a and b)]. However, there are significant changes of the degrees of nativeness of contacts in *island*-2 (marked as 2) corresponding to the interactions between β_1 and β_2. This is not consistent with previous experimental and simulation results,[13,19] which indicate that the transition state of protein CI2 is retained after the circular permutation. This inconsistency may result from the assumption that all contact energies are the same. Considering the heterogeneity of the contact energies, namely, stronger interactions are preferable to the contacts with longer loops, we recalculate the degrees of nativeness of contacts for both the wild-type protein CI2 and its *perm* 41–42 in their transition states by setting (e.g., $\epsilon_i = -1$ for the loop lengths $l_i < 29$, $\epsilon_i = -1.5$ for $29 \le l_i < 44$, and $\epsilon_i = -4$ for $l_i \ge 44$). As shown in Figure 9(c and d), we can see that *island*-1, *island*-2, and *island*-3 are basically the same; meanwhile, *island*-4 shows a significant change (i.e., it disappears basically due to the connection of the N- and C-termini). These are relevant to the experimental findings.[13]

The free-energy functional method can also be applied to the permutations of other proteins, such as protein S6, which has recently been experimentally studied. As concluded from the experimental results, the energies of all the native contacts in protein S6 are not in the same level, and the interactions between the contacts with long loops

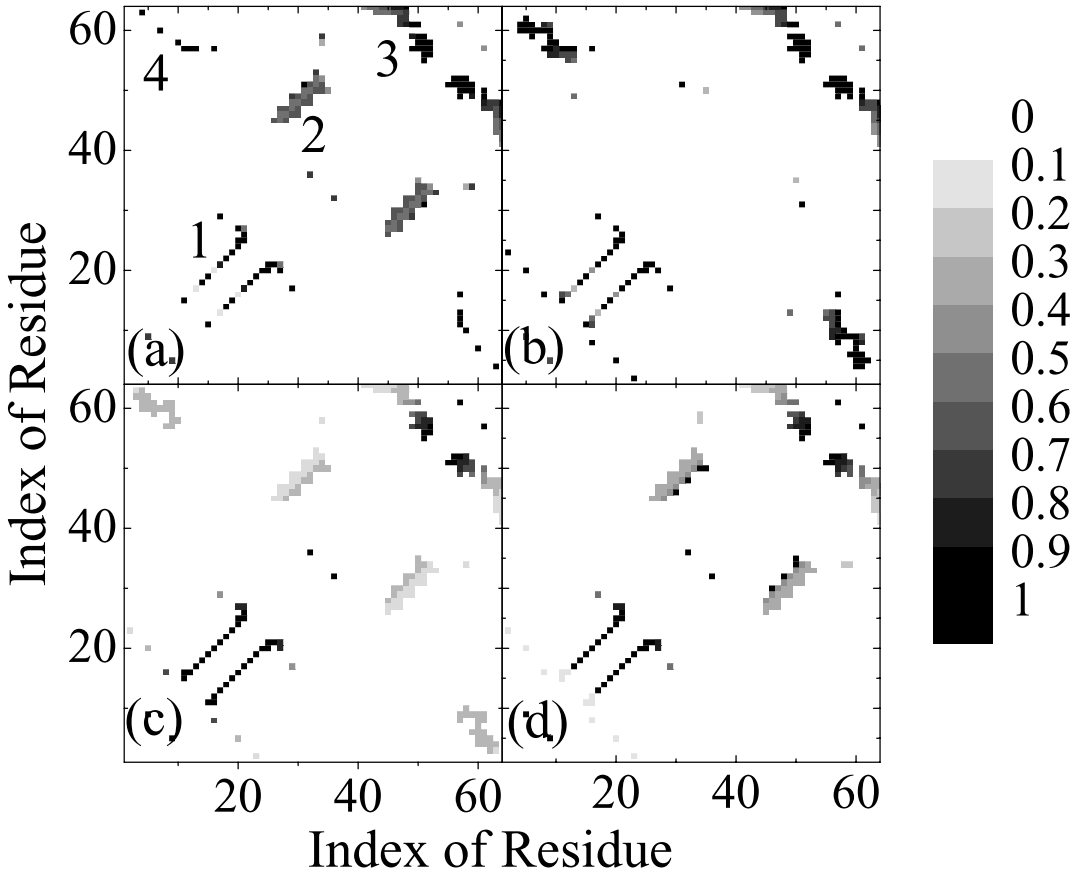

Fig. 9. The degrees of nativeness of contacts for protein CI2 and its permutant in the transition state shown on contact map. The grayscale indicates the values of the degrees. The values of contact energies are set as $\epsilon_i = -1.0$ in (**a**) and (**b**) and as $\epsilon_i = -1.0$ for loop lengths $l_i < 29$, $\epsilon_i = -1.5$ for $29 \leq l_i < 44$, and $\epsilon_i = -4.0$ for $l_i \geq 44$ in (*c*) and (**d**). (**a**) wild-type; (**b**) *perm* 40–41; (**c**) wild-type; (**d**) *perm* 40–41.

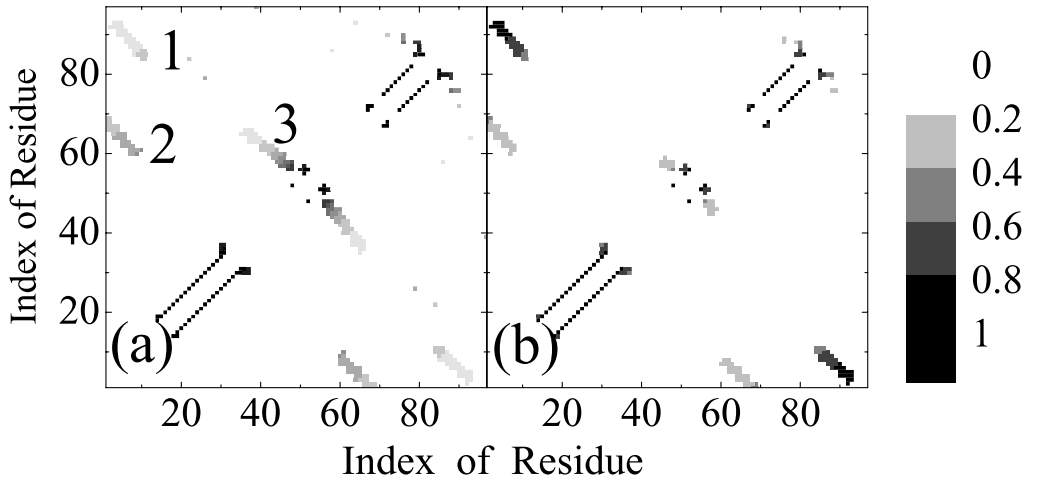

Fig. 10. The degrees of nativeness of contacts for protein S6 in the transition state shown on contact map. The values of contact energies are set as $\epsilon_i = -1.0$ for the loop lengths $l_i \leq 46$ and $\epsilon_i = -4.0$ for $l_i > 46$. The grayscale indicates the values of the degrees. (**a**) wild-type; (**b**) *perm* 13–14.

are speculated to be strong.[14] Thus, there is an apparent diversity of interaction intensities between the contacts. To properly model the protein S6, the assumption of equal values of the native contact energies should be modified, and a distribution of these energies related to the loop lengths should be introduced. This kind of distribution could maximize the effects of the permutation as shown in the experiments, and also provide a consistent comparison with experimental results. In detail, the contacts with long loops are strengthened by taking, for example, $\epsilon_i = -4$ for $l_i > 46$, and the contacts with short loops are weakened by, for example, $\epsilon_i = -1$ for $l_i \leq 46$. The physical reasons will

be discussed later in this article. Since the size of protein S6 is larger than those of proteins SH3 and CI2, the energetic effect is to some degree more important to guide the folding for protein S6. In addition, the distribution of interaction energies could also enhance the cooperativity of protein S6 and may reduce the possibility of appearance of intermediates during folding. These are actually relevant to the two-state properties of protein S6.

In Figure 10(a and b), *island*-1, *island*-2, and *island*-3 demonstrate the contacts that correspond to the interactions between β_1 and β_4, β_1 and β_3, and β_2 and β_3, respectively, in the transition states [also see Fig. 7(a and b)]. The contacts with short loops are favored by the entropic factor; thus, the local contacts in helical regions are always formed in the transition state. Due to the entropic deficiency for the longer loops, we can see that for the wild-type protein S6, any of these three islands with long loops in the transition state is less formed, that is, with $Q_i < 0.5$ [see Fig. 10(a)] when compared with their values in the native state $Q_i = 1$. Among these three islands, the loops of the contacts in *island*-1 and *island*-2 are longer than those in *island*-3, and the entropic contributions of these contacts are small. However, the interactions of these contacts are stronger than those in *island*-3. Thus, the probabilities to form these contacts are slightly increased due to the effect of strong interactions. As a result, the degrees of nativeness of the contacts in these three islands are more or less the same. This shows a globally diffuse nucleus for the wild-type of protein S6 as found in experiment.[14]

For *perm* 13–14 of protein S6, the contacts in *island*-1 are well formed, and the contacts in *island*-2 and *island*-3 are less formed with respect to those for the wild-type S6 [see Fig. 10(b)]. Since all the contacts in *island*-1 now become short due to the permutation and at the same time keep the feature of strong interactions, both the energetic and entropic effects promote the formation of contacts in *island*-1. Thus, the transition state is composed of contacts in *island*-1 with short loops (in the permutant) and strong interactions. The contacts with long loops (i.e., the contacts with short loops in the wild-type S6) basically play no role in the transition state due to unfavorable entropic and energetic effects. Consequently, the transition state of the permutant becomes polarized, as argued by Lindberg et al.[14] Thus, clearly, energetic heterogeneity plays an important role in changes of the contact patterns of the transition states.

From the 3 cases mentioned above, we find that due to the changes of the degrees of the nativeness for various contacts in the permuted proteins, the related energy profiles are changed dramatically with respect to those of their wild-type proteins. When a cutting is made in the high-density region of the folding nucleus, the contacts presented in the transition state of the wild-type proteins will experience a great change in their loop entropy due to the considerable variation of the loop lengths related to these contacts. As a result, the transition state would change thoroughly. *Perm* 38–39 of protein SH3 is an example of such case, since residue 39 is in the third β-strand and is believed experimentally to be important for folding (i.e., with a large value of ϕ).[32] In contrast, when the cuttings are in regions with fewer nonlocal contacts, such as the loops, there is less effect on the folding (e.g., for the case of *perm* 14–15 of protein SH3). These results are consistent with those shown previously by Serrano and coworkers.[12] They mentioned that changes of the transition state by circular permutation depend on the cutting positions. In addition, our results partially illustrate the important effects of the so-called contact order as discussed by Baker[34] (and references therein), since the cutting changes the contact order but keeps the interactions unchanged. A favorable quantitative comparison between our calculations of the ϕ-values and experimental results will be presented in the Comparison With Experiment section.

Interestingly, besides the consistency with experimental studies, our results also provide a way to compare with simulation researches.[19,20] Our results show a similar pattern of contacts with respect to the simulation findings. This enables us to make further comparison of the folding pathways between different permutants, and indicates that the theory of variational free-energy functional captures qualitatively the thermodynamics of folding. Comparing the contours of the degrees of the nativeness of the contacts, the remarkable changes of the transition-state architecture can also be visualized as derived from the nativeness of the residues (as shown in Figs. 2, 3, and 4). In other words, after tailoring the sequential connections, the transition state of the protein molecules would be totally different. The protein molecules would go to the native state from different pathways.

Characterization of Folding Processes With Q_i versus Q

To expand our view of the whole folding processes, we show the variation of $Q_i(Q)$ along the global coordinate Q, as shown in Figures 11, 12, and 13, respectively. The degrees of nativeness of contact i, Q_i, are represented with different grayscale levels, and the darker colors indicate the better formed status. Most of the multiple nucleation sites (i.e., some darker horizontal bands) appear in the folding processes toward the native structures for both the wild-type proteins and their permutants. This means that some folding nuclei of the wild-type proteins are still kept in various permutants, but different permutants keep different nuclei. In detail, the degree of nativeness of contacts, Q_i, behaves in a distinct manner for each value of Q. The above-mentioned variation of the transition states, as slices of these reorganized pathways, is an example.

For the SH3 domain and its related permutants, *island*-1, *island*-2, and *island*-3 in Figure 11(a) (i.e., for i = 50–70, i = 88– 110, and i = 120– 136, respectively), are formed earlier than the loop regions, regularly beginning to emerge when Q is greater than 0.1 (see Fig. 11). Comparing the figure for *perm* 14–15 of protein SH3 [Fig. 11(b)] with that of the wild-type protein [Fig. 11(a)], we find that more than half of the contacts are the same, while the contacts formed by the residues in the reactive loop

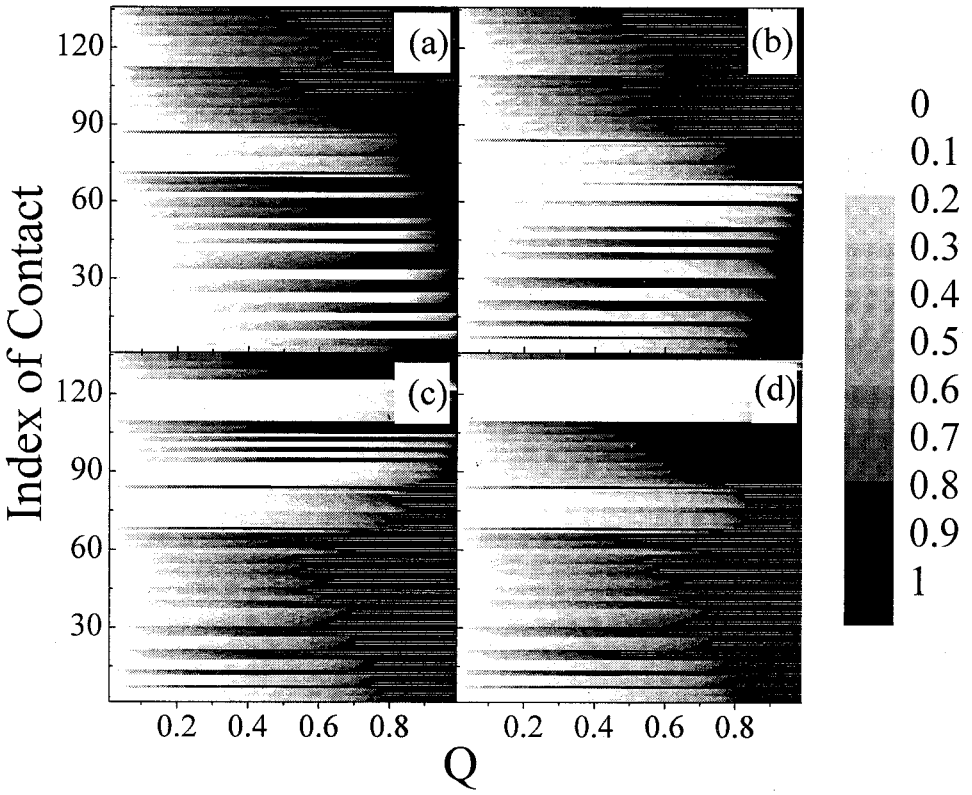

Fig. 11. The distribution of degrees of nativeness of contact for protein SH3 in the folding process shown by contact index *i* versus the *Q* value. The grayscale indicates the values of the degrees. The values of contact energies are the same as in Figure 8. (**a**) wild-type; (**b**) *perm* 14–15; (**c**) *perm* 38–39; (**d**) *perm* 42–43.

(*island*-1 in Fig. 8) (i.e., for $i = 50$–70) are slowly ordered. Nevertheless, the folding behaviors of *perm* 38–39 and *perm* 42–43 change significantly with respect to that of *perm* 14–15. Consequently, the sequence of folding events changes. From Figure11(c and d), we can see that the contacts between the terminal residues [*island*-4 in Fig. 8(a)] (i.e., for $i = 7$, 12, 21, 27–30) form first, before the formation of the distal β-hairpin (*island*-2).

For both the wild-type protein CI2 and its permutant, the observations are similar to the case of protein SH3. In the case of the wild-type, the contacts between the terminal residues (i.e., contacts $i = 2$–37) form later, around $Q \sim 0.8$, and the contacts in the reactive loop; that is, contacts $i = 55$–90, begin to form around $Q \sim 0.3$ [see Fig.12(a)]. On the contrary, for *perm* 40–41, the contacts $i = 2$–37 form early around $Q \sim 0.3$, and the contacts $i = 55$–90 form later after the transition state [see Fig. 12(b)]. The other contacts show basically no difference between the wild-type and its permutant of protein CI2. However, the folding pathways are basically the same before the transition state (i.e., $Q < 0.5$), when a distribution of the contact energies is introduced [see Fig.12(c and d)]. The main difference is for the contacts between the terminal residues (i.e., the contacts $i = 2$–37).

Similarly, the folding pathways of the wild-type protein S6 and its permutant are shown in Figure 13. From Figure13(a and b), we can see that the helical regions are obviously formed earlier than any other regions for both the wild-type and its permutant, whereas the contacts between the secondary structures often formed later. This observation is consistent with the well-known result that the secondary structural units usually form first, followed by the tertiary interactions, and thus the contacts between them. Comparing these two pathways shown in Figure 13(a and b) with each other, we find that the contacts $i < 80$ corresponding to *island*-1 and *island*-2 in Figure 10, namely, the β regions, are all less than half formed, with $Q_i < 0.5$ for the wild-type protein S6. (Note that the contacts $80 < i < 100$ are in helical region. They have loops $l_i \leq 5$ and are not included here.) However, for the circular permutant of protein S6, *island*-1 is well formed due to the energetic and entropic advantages, and at the same time, *island*-3 has only slight changes with respect to the case of the wild-type.

Distance Between the Wild-Type Proteins and Their Permutants

In order to characterize the difference between various folding pathways, a distance function can be defined as

$$\mathcal{D}(Q) = \frac{1}{M}\sum_{i=1}^{M}\left|\frac{Q_i^{\dagger perm}(Q)}{Q} - \frac{Q_i^{\dagger wild}(Q)}{Q}\right| = \frac{1}{MQ}\sum_{i=1}^{M}|Q_i^{\dagger perm}(Q) - Q_i^{\dagger wild}(Q)|, \quad (11)$$

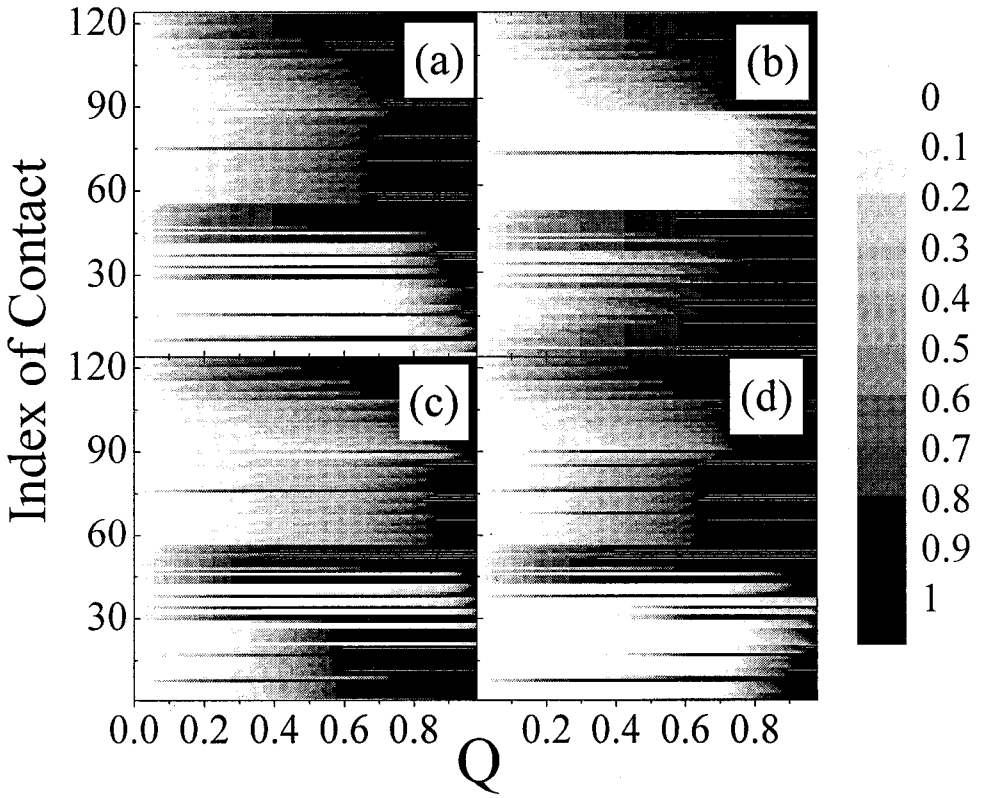

Fig. 12. The distribution of degrees of nativeness of contact for protein CI2 in the folding process shown by contact index *i* versus the *Q* value. The grayscale indicates the values of the degrees. The values of contact energies are the same as in Figure 9. (**a**) wild-type; (b) *perm* 40−41; (**c**) wild-type; (d) *perm* 40−41.

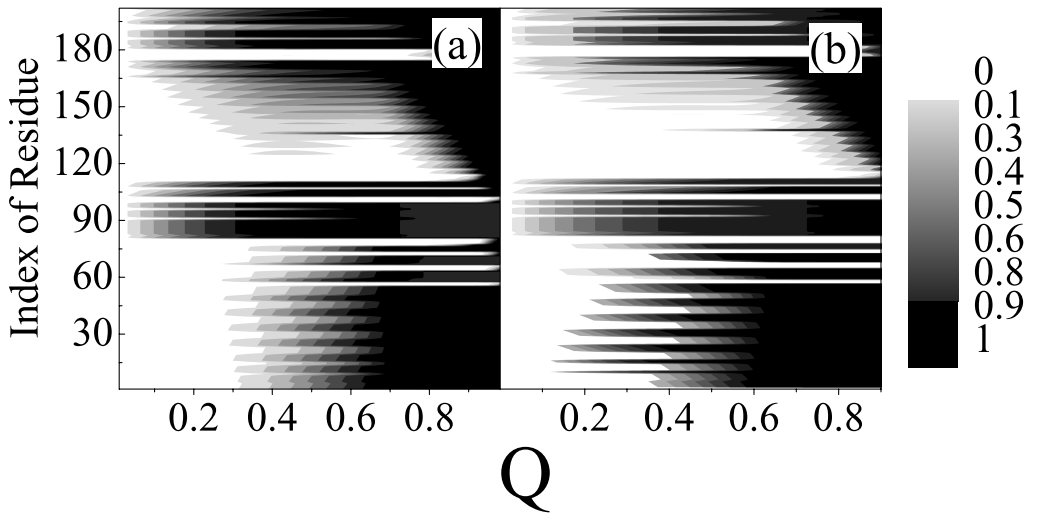

Fig. 13. The distribution of degrees of nativeness of contact for protein S6 in the folding process. The grayscale indicates the values of the degrees. The values of contact energies are the same as in Figure 10. (**a**) wild-type; (**b**) *perm* 13–14.

which describes the averaged difference of the degrees of nativeness of the contacts between a wild-type protein and its permutants. The distance approaches 0 when two distributions, $\{Q_i^{\dagger perm}(Q)\}$ and $\{Q_i^{\dagger wild}(Q)\}$, are identical, and the distance is 2 when two distributions are totally different, namely, $Q_i^{\dagger perm} = 1$ and $Q_i^{\dagger wild} = 0$ in two distributions, or vice versa. Each quantity $Q^{\dagger}(Q)$ is assumed to be a vector with M components, and at the same time, this vector is scaled by the value of M. Such a scale makes the comparison of the difference in degrees of nativeness of contacts between different pairs of $Q^{\dagger perm} - Q^{\dagger wild}$ be independent of the values of Q. Note that this

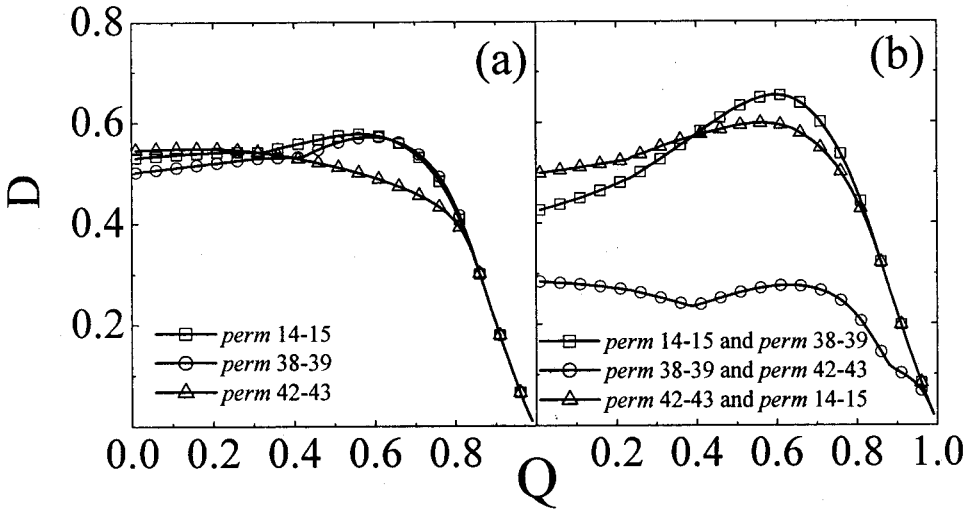

Fig. 14. Distance of degrees of nativeness of contacts $\mathcal{D}(Q)$ versus Q between the wild-type protein SH3 and its permutants (**a**), and between different permutants (**b**). The values of contact energies are the same as in Figure 8.

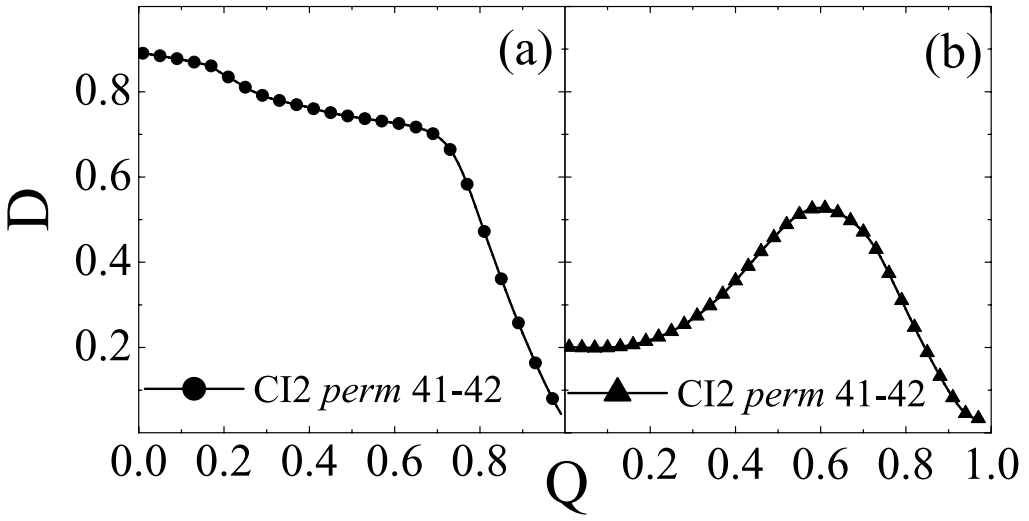

Fig. 15. Distance of degrees of nativeness of contacts $\mathcal{D}(Q)$ versus Q between the wild-type protein CI2 and its permutants. The values of contact energies are set as $\epsilon_i = -1.0$ for all contacts in (**a**), and as $\epsilon_i = -1.0$ for loop lengths $l_i < 29$, $\epsilon_i = -1.5$ for $29 < l_i < 44$, and $\epsilon_i = -4.0$ for $l_i \geq 44$ in (**b**).

scale is not the same as a normalization of $Q^{\dagger}(Q)$. The distances $\mathcal{D}$ of the contact degrees for the 3 proteins, SH3, CI2, and S6, are plotted in Figures 14, 15, and 16, respectively. As can be seen, when the value of $Q > 0.75$, the distances for proteins SH3 and CI2 go to 0 in a similar way. This can be attributed to the fact that after passing through the transition state, the final stages of folding would be similar for the wild-type protein and its permutants; that is, the effects of the permutation decrease as the protein approaches its native state. In addition, there are plateaus in $\mathcal{D}$ for both proteins SH3 and CI2 [see Figs.14(a) and 15(a)]. The existence of such plateau implies that the permutants have a large effect on the folding behavior before the transition state. However, the situation for protein S6 is quite different, and the $\mathcal{D}$ value decreases to 0 around $Q \approx 0.7$.

In detail, for protein SH3, 3 permuted proteins show small differences in the $\mathcal{D}$ values, and they all have plateaus when $Q < 0.50$ [see Fig. 14(a)]. It can be seen that for *perm* 42–43, the values of $\mathcal{D}$ decrease monotonically, while the $\mathcal{D}$ values for both *perm* 14–15 and *perm* 38–39 have a maximum around $Q = 0.60$ (i.e., around the transition state). This means that both *perm* 14–15 and *perm* 38–39 show a large difference around the transition state with respect to that of the wild-type protein, and the destruction of contacts around *island*-1 and *island*-2 due to the permutation has large effects on the transition state with $\mathcal{D} \approx 0.6$. Nevertheless, for *perm* 42–43, the destruction of the contacts around *island*-3 has fewer effects (i.e., with $\mathcal{D} \approx 0.5$) on the transition state when compared with those of *perm* 38–39. In addition, the contacts between the 2 termini are also important, because they result in a large

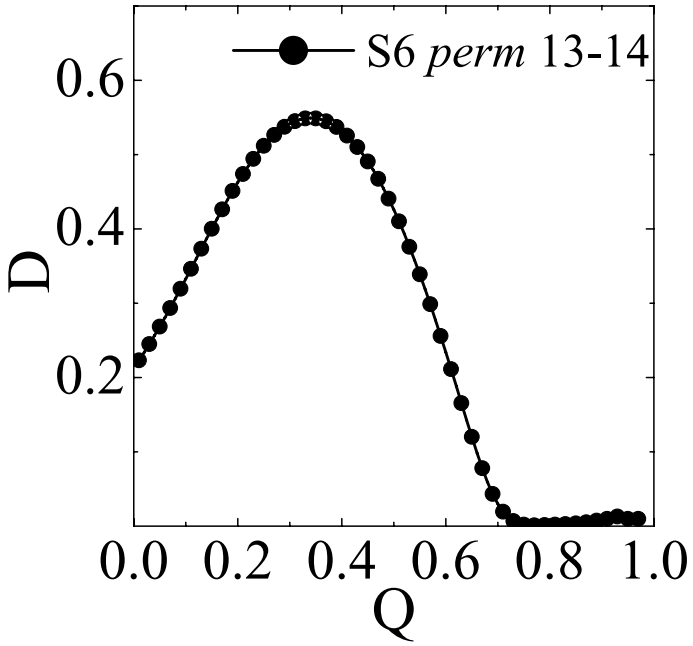

Fig. 16. Distance of degrees of nativeness of contacts $\mathcal{D}(Q)$ versus Q between the wild-type protein S6 and its permutant. The values of contact energies are the same as in Figure 10.

difference in $\mathcal{D}$ around the transition state. For *perm* 42–43, the interactions between β_3 and β_4 in the wild-type protein SH3 are replaced by those between the 2 termini. As a result, the value of distance $\mathcal{D}$ for *perm* 42–43 decreases slowly and shows no maximum around the transition state. Physically, we can interpret the results mentioned above as follows. For protein SH3, there are only sheet conformations (i.e., many interactions are between different strands), and the contacts have long loops (see Fig. 5). The permutations for protein SH3 generally increase the number of contacts with long loops, resulting in a slow formation of sheet conformations. This then enables different folding processes compared with those of the wild-type case, producing a plateau in the $\mathcal{D}$ value for $Q < 0.65$.

Distances between the 3 permutants of protein SH3 are also calculated. Figure 14(b) shows these distances between pairs of *perm* 14–15, *perm* 38–39, and *perm* 42–43. It can be seen that the distances between *perm* 14–15 and *perm* 38–39, and between *perm* 14–15 and *perm* 42–43 are large before the transition state and have maxima around the transition state. *Perm* 14–15 is similar to the wild-type protein SH3, since the $\mathcal{D}$ values of *perm* 38–39 with respect to the wild-type in Figure14(a) are basically the same as those between *perm* 14–15 and *perm* 38–39 in Figure 14(b). The similar distributions of loop lengths for both *perm* 14–15 and the wild-type (see Fig. 5) may be an explanation of this argument. It is also observed that the cutting of a peptide bond does not affect the 3-stranded β-sheet of *perm* 14–15 too much, indicating a strong interaction in that region. Thus, the 3-strand β-sheet is regarded to be important for the transition state and the whole folding mechanism of protein SH3. This result is relevant to previous unfolding simulations, which show that the strong interaction in 3-stranded antiparallel β-sheet of SH3 domain makes the disappearance of this region the last unfolding event.[35] Meanwhile, the small distance values between *perm* 38–39 and *perm* 42–43 mean that the folding pathways of these two permutants are comparable, which suggests that the folding mechanism of the permutants of protein SH3 might be kept by the retainment of the native contact maps and their distributions of contact loop lengths.

The permutant of protein CI2 shows a similar behavior, but the $\mathcal{D}$ values are larger than those of protein SH3 below the transition state, as shown in Figure 15(a). This is due to the large effect of the changes of the contact loop lengths for protein CI2 (see Fig. 6) below the transition state, resulting in large $\mathcal{D}$ values before the transition state with respect to the wild-type protein. The $\mathcal{D}$ values show two plateaus around $Q < 0.20$ and $0.40 < Q < 0.75$. These two plateaus correspond to two stages of the formation of helical conformations and other secondary conformations, respectively. For *perm* 40–41 of protein CI2, due to the connection of 2 termini, the contacts around the termini become local and form first and fast. Especially, these contacts near the helical conformations form fast during the folding process. Alternatively, cutting around the loop region makes the contacts near the cutting position become long-range and form slowly, or even after the transition state [see Fig.12(a and b) around the contacts $i = 55$–90]. Therefore, the effects of permutations on the unfolded state of the permutant of protein CI2 are large until the system reaches the transition state. This results, in the first plateau, in the curve of $\mathcal{D}$ versus Q, until the contacts around the termini are well formed for $Q \sim 0.2$, and then the second plateau, until other secondary conformations are formed gradually. Finally, the value

of $\mathcal{D}$ decreases as the contacts around the loop region are formed (see Fig. 12).

The situation is different when we consider the heterogeneity of the contact energies [Fig. 15(b)]. The $\mathcal{D}$ values change to become much lower than those in Figure 15(a), and there is a plateau for $Q < 0.2$ and a maximum around $Q \approx 0.6$ (i.e., around the transition state). This indicates that the permutation has less effect on the folding except around the transition state. Such behavior results from the contacts with long loops (around the 2 termini of the wild-type of protein CI2) having stronger interactions than those with short loops. These contacts with strong interactions between the 2 termini increase the stability of the wild-type protein CI2 and form early in the folding processes for both the wild-type and its *perm* 40–41. Besides, the position of permutation is located at the region of weak contact interactions. Therefore, the folding behaviors of both the wild-type and its *perm* 40–41 have small differences.

For protein S6, as shown in Figure 16, the $\mathcal{D}$ value versus Q is quite different with respect to proteins SH3 and CI2. There is no plateau while the maximum is located around $Q \sim 0.3$, which is below the transition state, while around the transition state with $Q \sim 0.45$, the $\mathcal{D}$ value becomes smaller (i.e., $\mathcal{D} < 0.4$). In addition, the distance is basically 0 when$Q > 0.7$, indicating that the native structure of the permutation is almost well formed. The large distance value around$Q \sim 0.2$–0.3 indicates that the folding behaviors of protein S6 and of its permutant are different in the loose structures (i.e., the unfolded states). In general, the permutations change the degrees of nativeness of contacts in the low$Q$ value region (i.e., change the structural features of the unfolded state). Thus, the loose structures, or the initial stage of folding, are important for the folding of protein S6, as also suggested in other literature.[36] This indicates that the folding behaviors of the wild-type and its permutant have large difference around the transition state and the unfolded state.

In a word, the 3 cases discussed above show that the permutations affect both the transition states and the unfolded states. From our results, the folding behaviors of the 3 proteins and their permutants could be well characterized by the free-energy functional.

COMPARISON WITH EXPERIMENTS

Now let us make a quantitative comparison of the results from our theoretical study with those from the experiments. The theoretical, calculated ϕ values compared with those from experiments[12–14] for proteins CI2, SH3, and S6 are plotted in Figures 17, 18, and 19, respectively. The theoretical calculated ϕ values, ϕ^{th}, are referred to as the nativeness of residues in the transition states, namely, $\phi^{th} = P_{\omega}$, with P_{α} defined as Eq. (10).[19] The experimental ϕ values (i.e., ϕ^{exp}) are taken from previous works.[12–14] Similar to the comparison made by Galzitskaya and Finkelstein[37] the negative ϕ^{exp} values are not included, the ϕ^{exp} values exceeding 1 are set as 1, and the largest ϕ^{exp} values are taken when there are several different mutations provided.

Figure 17(a and b) shows the results for the comparison of values of ϕ^{th} and ϕ^{exp} for protein CI2 and its circular permutant.[13] From Figure 17(a) for protein CI2, it can be seen clearly that both in the helical region and in the β-hairpin around the C-terminal, the ϕ^{th} values are large. This is in agreement with the experimental results as shown by the solid circles in Figure 17(a). The folding nucleus (i.e., residues 16, 49, and 57, as indicated by Otzen and Fersht[13]) have both high ϕ^{th} and ϕ^{exp} values, which correspond to the high formation probabilities of nativeness of these residues in the transition state. However, the ϕ^{th} values for the residues in the β-hairpin around the C-terminal are slightly greater than the ϕ^{exp} values. The correlation coefficient between the experimental ϕ^{exp} values and the theoretical ϕ^{th} values is $r = 0.88$ [see Fig.17(c)], which is better than previous results with $r = 0.56$[37] and $r = 0.58$.[38] Here, the contact energies between residues are taken as $\epsilon_i = -1$ for $l_i < 29$, $\epsilon_i = -3$ for $29 \le l_i < 47$, and $\epsilon_i = -0.2$ for $l_i \ge 47$, respectively. Note that the contact energies of the residues near the termini are weakened. Such a consideration relates to the dynamical flexibility of the residues at the termini. This set of the contact energies is slightly different from that used in the last section. However, the contact maps and the difference $\mathcal{D}$ are found to be similar to those in Figures 9(d) and 15(b). It is also worth noting that if all the contact energies are set as a unique value $\epsilon_i = -1$, or as two values (i.e., $\epsilon_i = -1$ for $l_i \le 46$ and $\epsilon_i = -3$ for $l_i > 46$), the correlation is not as good as in Figure 17(c). Roughly, for the latter case, it is found that $r = 0.6$ and for the former case, $r < 0.5$. Therefore, the values of contact energies play an important role in the detailed comparison of the ϕ values between experimental observations and theoretical calculations. The comparison of values of ϕ^{th} and ϕ^{exp} and their correlation for *perm* 40–41 of protein CI2 is plotted in Figure 17(b and d), respectively. Compared with previous theoretical works,[37–39] our ϕ^{th} values show better correlation with the experimental ones for the permutant.

Similarly, the comparisons of ϕ^{th} and ϕ^{exp} values for proteins SH3 and its circular permutants are shown in Figure 18(a–c), and the correlation coefficients between the experimental and theoretical results are plotted in Figure 18(d–f). [Here, the figures for *perm* 38–39 are not given due to the absence of related experimental values of ϕ^{exp}.] Considering that for the β-protein SH3 the conformation of loop regions is more flexible than the regions of β-sheet, the energies of contacts in the loop regions are set as $\epsilon_i = -0.5$, and energies of contacts in the formed β-regions as $\epsilon_i = -1.5$. From Figure 18(d–f), it can be seen clearly that for the wild-type, *perm* 14–15 and *perm* 42–43, the correlation coefficients are $r = 0.65$, $r = -0.83$, and $r = 0.43$, respectively. The best correlation is for the wild-type, and the worst is for *perm* 14–15. For both the wild-type and *perm* 42–43, the calculated ϕ^{th} values are smaller than the ϕ^{exp} values. However, the correlation between the ϕ^{th} values and the ϕ^{exp} values for *perm* 14–15 is negative. This was seen previously and is due to large changes in energies caused by the loss of 9 conserved hydrogen bonds.[12] Anyhow, the correlation coefficient

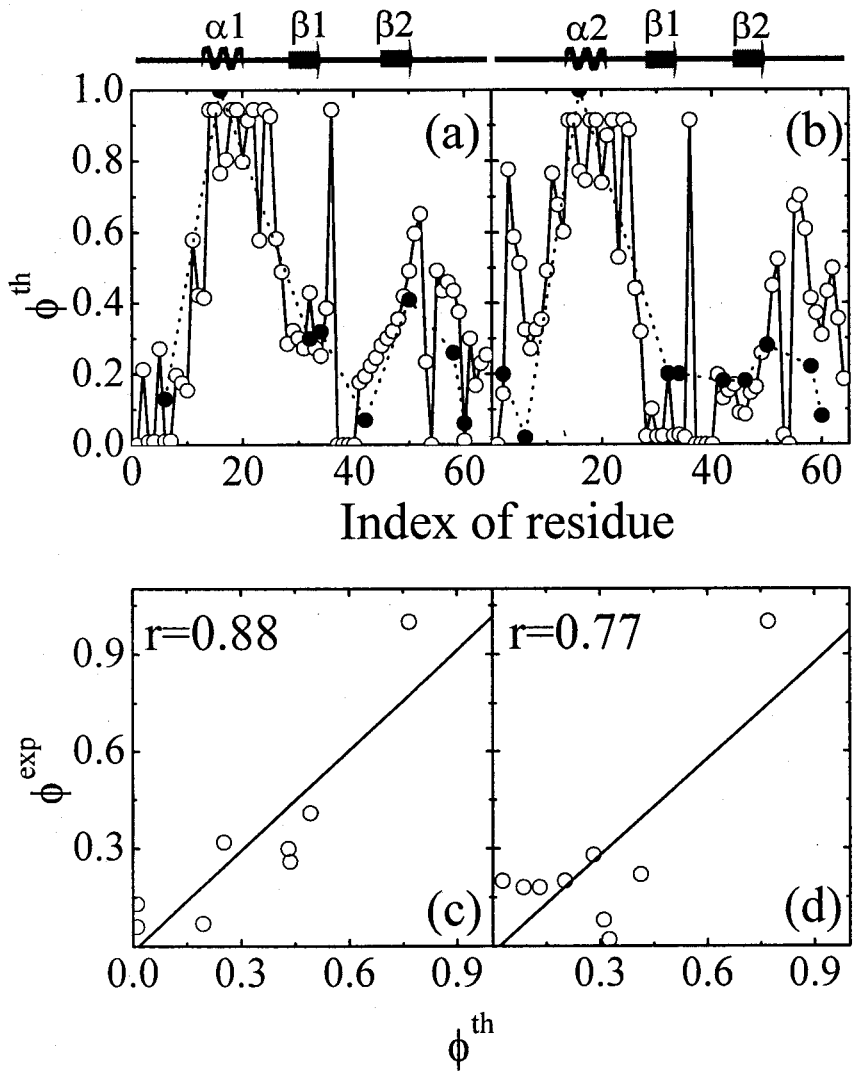

Fig. 17. Comparisons of the theoretical ϕ^{th} values and the experimental ϕ^{exp} values for protein CI2 (**a**) and *perm* 40–41 (–). The experimental ϕ^{exp} values are shown with filled circles and connected with dotted lines, while the theoretical ϕ^{th} values are shown with open circles and connected with solid lines. In (**c**) and (**d**), the correlation coefficients are shown as $r = 0.88$ for protein CI2, and $r = 0.77$ for CI2 *perm* 40–41.

obtained in this work for the wild-type is larger than that given in the previous theoretical works, namely, $r = 0.39$,[37] and $r = 0.5$.[38]

The comparisons between ϕ^{th} values and ϕ^{exp} values for protein S6 and *perm* 13–14 are shown in Figure 19(a and b), and the related correlations are plotted in Figure 19(c and d). In our calculations, the contact energies for both cases are set as $\epsilon_i = -0.1$ for $l_i \leq 40$, $\epsilon_i = -4$ for $40 < l_i \leq 60$, and $\epsilon_i = -3$ for $l_i > 60$, respectively. As with protein CI2, the energies of contacts formed around the terminal regions are weakened due to the flexibility of the residues around the two termini. This is consistent with the argument by Lindberg et al.[14] Actually, the interactions of contacts with long loops are strengthened for the loop length larger than 40 (i.e., $l_i > 40$), but not extending to the two termini. From Figure 19(a and b), it can be seen clearly that the circular permutation of protein S6 makes the diffused transition state of the wild-type become polarized,[14] and the theoretical and experimental points relate to each other very well. As shown in Figure 19(c and d), the correlation coefficients are $r = 0.69$ and $r = 0.80$, respectively.

From the aforementioned comparisons of experimental and theoretical ϕ values for the 3 proteins and their permutants, we have seen that the calculated ϕ^{th} values and the features of the transition states are consistent with previous, related experimental observations.[12–14] The correlation coefficients between theoretical and experimental data of ϕ values are better than those in previous works.[37–39] However, these correlations are not perfect. The absence of a general criterion for the modification of the contact energies might be a reason. Besides, the unfavorable results of *perm* 14–15 for protein SH3 indicate that the free-energy functional method should be improved further, and the non-native interaction may need to be considered. In short, the free-energy functional

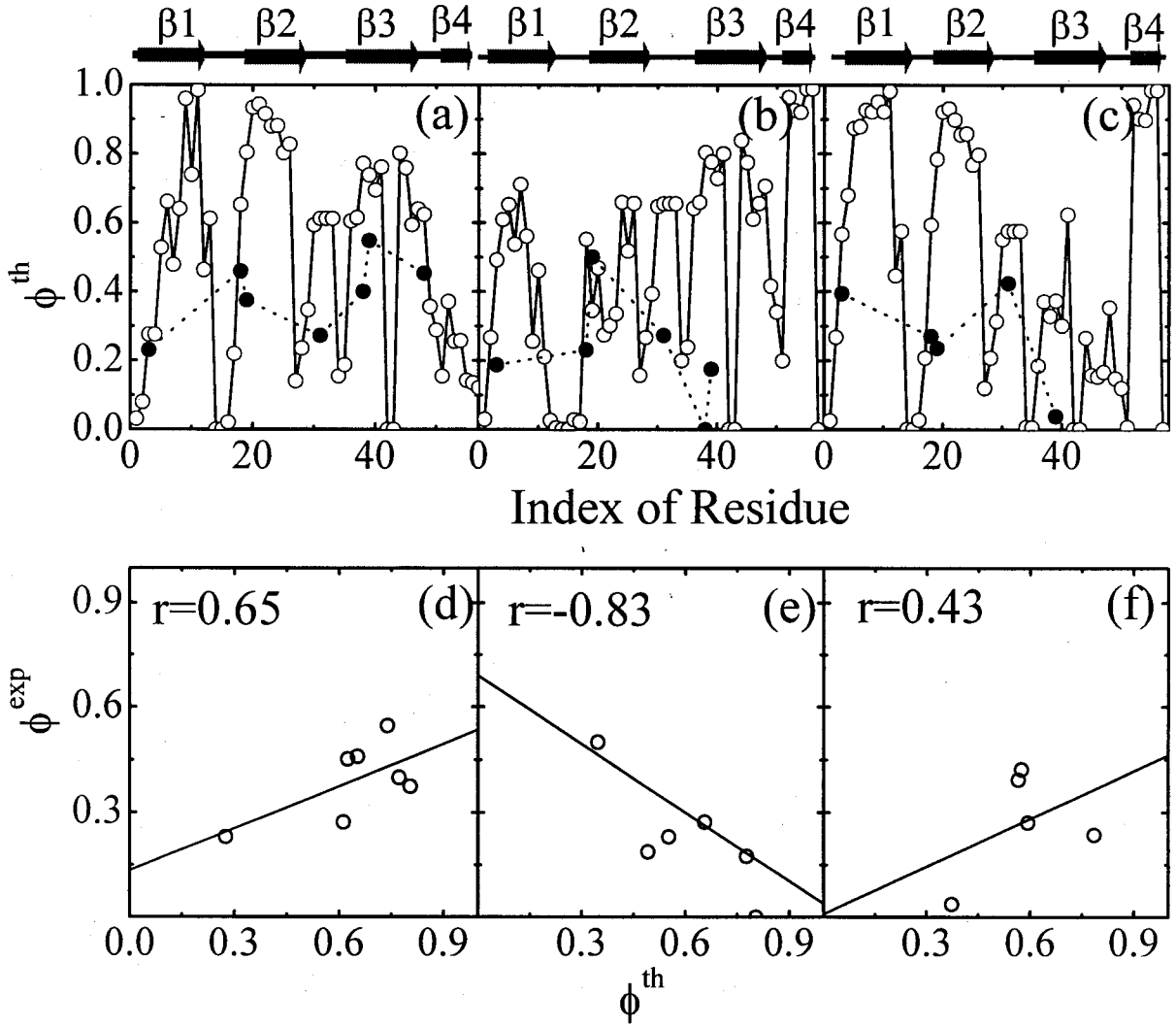

Fig. 18. Comparisons of the theoretical ϕ^{th} values and the experimental ϕ^{exp} values for protein SH3 (**a**), SH3 *perm* 14–15 (**b**), and SH3 *perm* 42–43 (**c**). The experimental ϕ^{exp} values are shown with filled circles and connected with dotted lines, while the theoretical ϕ^{th} values are shown with open circles and connected with solid lines. In (**d**), (**e**), and (**f**), the correlation coefficients are shown as $r = 0.64$ for protein SH3, $r = 0.83$ for SH3 *perm* 14–15, and $r = 0.43$ for SH3 *perm* 42–43.

method can be used to study the transition states of circular-permuted proteins.

REMARKS ON THE HETEROGENEITY OF CONTACT ENERGIES

In our calculations presented above, we have introduced the heterogeneity of the contact energies for protein CI2 and protein S6, respectively. Now, let us make a remark on its physical origins. It is known that different residues in a protein have different numbers of contacts with other residues. If a pair of residues that form a contact between each other both have large numbers of contacts with other residues, it may imply that these two residues show strong cooperativity, and the contact between them is much more stable.[9,16,40,41] Thus, the interaction of this contact is presumably strong and the value of contact energy is large (with a minus sign). The number of the contacts related to a pair of residues α and β of a certain contact is defined as

$$N^{C}(\alpha\beta) = \sum_{\alpha'}^{N} \mathcal{L}_{\alpha'\beta} + \sum_{\beta'}^{N} \mathcal{L}_{\alpha\beta'} - 1,$$

with $\mathcal{L}_{\alpha\beta}$ described in the Methods section. Note that many elements in $\mathcal{L}_{\alpha\beta}$ are zeros except the related contacts. The quantity $N^{C}(\alpha\beta)$ can be used roughly to evaluate the magnitude of the interaction of the contact formed between the residues α and β.

In Figure 20, the histograms of numbers of contacts over $N^{C}(\alpha\beta)$ versus the loop lengths of the contacts are plotted for the 3 proteins, where the grayscale represents the contact number with the same values of N^{C} and the loop length. The darker the area, the greater the number of contacts there. From Figure 20(a), we can see clearly that for protein SH3, most of the contacts are distributed in an area with loop lengths between 5 and 30, and the values of N^{C} between 9 and 18. Note that the number of contacts

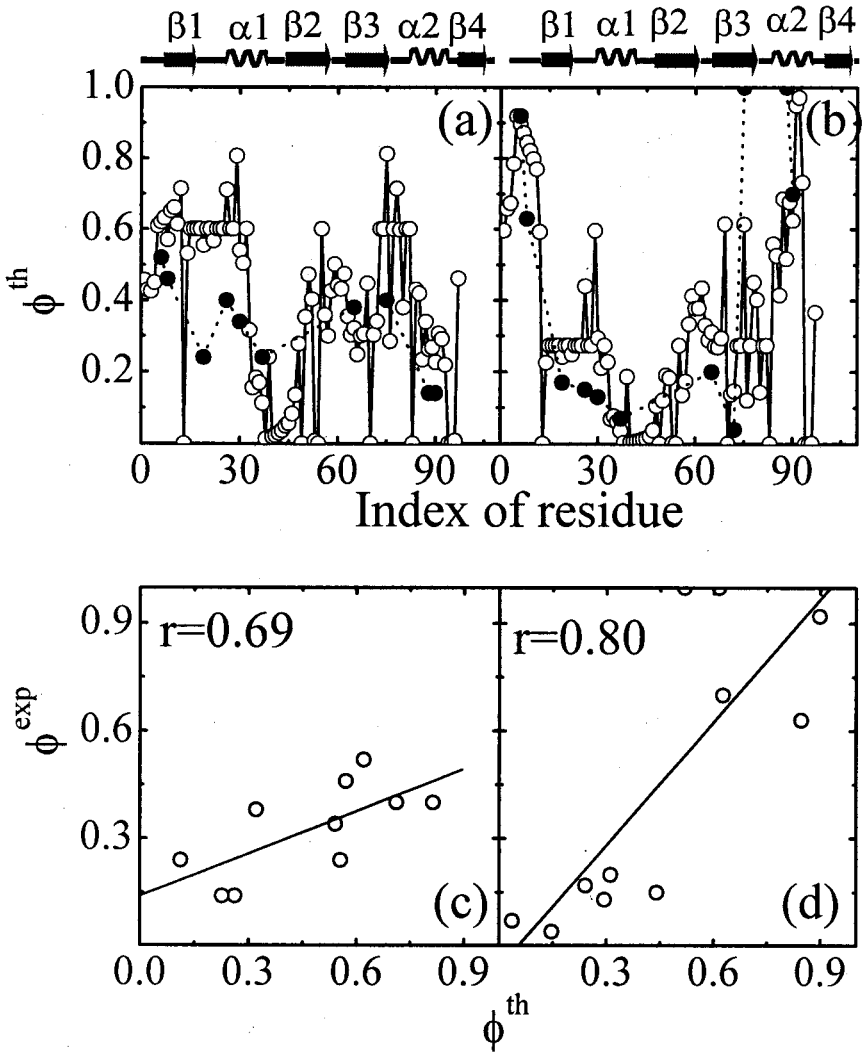

Fig. 19. Comparisons of the theoretical ϕ^{th} values and the experimental ϕ^{exp} values for protein S6 (**a**) and S6 *perm* 13–14 (**b**). The experimental ϕ^{th} values are shown with filled circles and connected with dotted lines, while the theoretical ϕth values are shown with open circles and connected with solid lines. In (**c**) and (**d**), the correlation coefficients are shown as $r = 0.69$ for protein S6, and $r = 0.80$ for S6 *perm* 13–14.

with loop lengths larger than 29 is quite small. Therefore, it is reasonable to use only one energy value for all the interactions of the contacts when the contacts with long loops are ignored. As a comparison, protein CI2 has a distribution of the contact numbers mainly around two regions: One region has many contacts with values of N^C from 5 to 20, and values of loop lengths from 5 to 25, and the other has values of N^C from 7 to 18, and values of loop lengths from 36 to 55 [see Fig. 20(b)]. It seems that a set of exact values of contact energies cannot be obtained based on the simplified consideration. However, physically, the contact energies can be regarded as weak when the loops are short. The interactions of contacts between residues can be assumed to have three values (e.g., as $\epsilon_i = -1$ for loop lengths $l_i < 29$, $\epsilon_i = -1.5$ for $29 \le l_i < 44$, and $\epsilon_i = -4$ for $l_i \ge 44$). Note that the interaction energies are not set as zeros for contacts with loop lengths from 29 to 44, even if no contacts exist for the native conformation. Actually they contribute zero to the total contact energy for the wild-type CI2. However, there may be contacts when a circular-permutation is made for the protein. For example, when a peptide bond between two sequential residues in this range is cut, some contacts may form between these two residues and their related neighbors, and some contacts with short or long loops will fall in this range. Similarly, for protein S6, Figure 20(c) shows roughly two regions (i.e., one region with N^C values from 8 to 16 for loop lengths smaller than 35, and other with N^C values from 12 to 18 for loop lengths larger than 46). Consequently, the interactions can be approximately considered as having two kinds of values (e.g., contact energies with $\epsilon_i = -1$ for the loop lengths $l_i \le 46$ and $\epsilon_i = -4$ for $l_i > 46$). Of course, the contact energies can be divided into more sections for different loop lengths.

The above-mentioned argument can also be supported by the statistics on the numbers of hydrogen bonds (H-

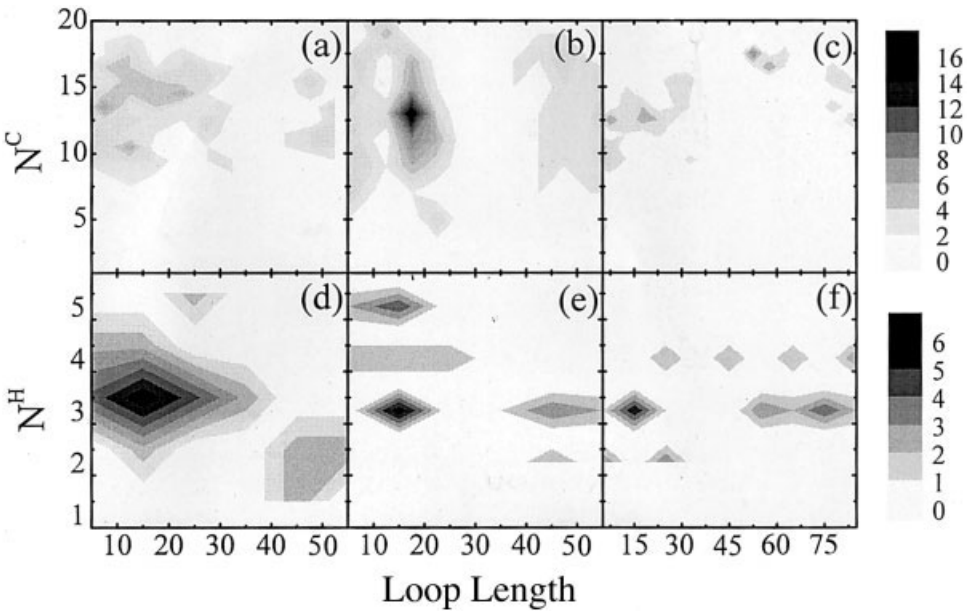

Fig. 20. The histograms of numbers of contacts over $N^C(\alpha\beta)$, versus the loop lengths of the contacts for protein SH3 (**a**), CI2 (**b**), and S6 (**c**). The grayscale represents the contact numbers with the same values of N^C and the same loop length. The histograms represent numbers of H-bonds over $N^H(\alpha\beta)$ versus the loop lengths of H-bond for protein SH3 (**d**), CI2 (**e**), and S6 (**f**). The grayscale represents H-bond numbers with the same values of N^H and the same loop length.

bonds) over their loop lengths. Here, for the sake of simplicity, only the H-bonds between the backbones with loop lengths $l_i > 5$ are considered. Figure 20(d–f) shows histograms of the numbers of the H-bonds over $N^H(\alpha\beta)$ versus the H-bond loop lengths for the 3 proteins. Here, $N^H(\alpha\beta)$ is the number of H-bonds between residues α and β. The grayscale represents the H-bond number with the same values of N^H and the loop length of the H-bond. Similar to Figure 20(a–c), the darker the area, the larger the number of H-bond there. These distributions relate well to those of the number of contacts. Thus, our discussion in fact describes the main physical reason for the heterogeneity of the contact energies.

Finally, it is worth noting that it is difficult to decide the exact values of the interaction energies of the contacts. Here we have followed the speculation for the interaction energies for protein S6.[14] The interactions of contacts with long loops should be strong. This is reasonable, since to form contacts with long loops there must be a large compensation for the entropy, thus a strong interaction. There is an exception for the contacts around the two termini that need to be weakened due to the flexibility. Our studies on the folding behaviors for proteins CI2 and S6 clearly show such features, especially when making detailed comparisons of the calculated ϕ values with those of experimental observations. However, for a quantified description, it deserves further, detailed study on the level of all atomic interactions.

SUMMARY

In summary, using the free-energy functional method, the folding processes and the transition states of 3 small proteins, namely, SH3, CI2, and S6, and their permutants, are studied. The heterogeneities of the folding pathways are characterized by the formation probability of the contacts $\{Q_i^\dagger(Q)\}$ derived from the variation of the free-energy functional. Except for the mean-field behavior of l_{eff} when $Q \to 0$ or $Q \to 1$,[23] we find that the free-energy functional method yields folding processes that are in good agreement with former experimental and simulational results.[12,13,19,25] The chain connectivity plays a critical role in the folding of these small, two-state proteins. For permutants of protein SH3, the folding nuclei become unstable or even completely changed, which depends on the cutting positions. For the permutant of protein CI2, the folding nuclei also change when the cutting position is in an unstructured region located at the transition state of the wild-type protein CI2. This may be evidence that besides chain connectivity, contact energies also play a role in the transition state of the folding process.[14] In our study, the results demonstrate the statistical nature of the folding reaction. As is widely accepted, the folding processes of a protein can be easily described as diffusion on the free-energy surface, and there are many parallel folding pathways leading to the native state on the free-energy landscape.[1–8] The number and populations of different microscopic pathways explored during folding may vary from protein to protein, which results in various folding kinetics for different proteins. The perturbations of the free-energy surface induced by circular permutations have significant effects on some of the folding pathways for some relative cutting sites. Thus, there is a redistribution of the populations of folding routes among existing pathways, which are explicitly shown as the shift of the folding

nucleus. Sometimes even new pathways that previously might cost much free energy to be visited are traversed as a result of the circular permutation. Thus, the folding transition states may be completely changed by the circular permutation.[14] Nevertheless, in spite of the large differences in transition states, these circular permutants are able to fold into the same native structure as that of the wild-type, indicating the large plasticity in folding[12,14] as long as the topological factor dominates the folding of these proteins.

Our studies indicate that the variation method based on the free-energy functional can capture well the key features of the folding processes of small, two-state proteins. It can be used for further studies on the folding transition states and prediction of folding mechanisms of various proteins. However, achievement of a more precise description for protein systems deserves an incorporation of some factors, such as how to treat the non-native interactions not simply as the averaged background, and how to go beyond the mean-field approximation to obtain the loop entropy.

ACKNOWLEDGMENTS

Our thanks to the two anonymous referees for their kind comments and suggestions.

REFERENCES

1. Wolynes PG, Onuchic JN, Thirumalai D. Navigating the folding routes. Science 1995;267:1619–1620.
2. Socci ND, Nymeyer H, Onuchic JN. Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 1997;48:545–600.
3. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.
4. Du R, Pande VS, Grosberg A Yu, Tanaka T, Shakhnovich EI. On the transition coordinate for protein folding. J Chem Phys 1998;108: 334–350.
5. Socci ND, Onuchic JN, Wolynes PG. Protein folding mechanisms and the multidimensional folding funnel. Proteins 1998;32:136–158.
6. Chan HS, Dill KA. Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. Proteins 1998;30:2–33.
7. Plotkin SS, Onuchic JN. Understanding protein folding with energy landscape theory: Part I. Basic concepts. Quart Rev Biophys 2002;35:111–167.
8. Plotkin SS, Onuchic JN. Understanding protein folding with energy landscape theory: Part II. Quantitative concepts. Quart Rev Biophys 2002;35:205–286.
9. Fan K, Wang J, Wang W. Modeling two-state cooperativity in protein folding. Phys Rev E 2001;64:041907.
10. Wang J, Fan K, and Wang W. Kinetic transition in model proteins with a denatured native spinodal. Phys Rev E 2002;65:041925.
11. Itzhaki LS, Otzen DE, Fersht AR. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. J Mol Biol 1995;24;254:260–288.
12. Viguera AR, Serrano L, Wilmanns M. Different folding transition states may result in the same native state. Nat Struct Biol 1996;3:874–880.
13. Otzen DE, Fersht AR. Folding of circular and permuted chymotrypsin inhibitor 2: retention of the folding nucleus. Biochemistry 1998;37:8139–8146.
14. Lindberg M, Tangrot J, Oliveberg M. Complete change of the protein folding transition state upon circular permutation. Nat Struct Biol 2002;9:818–822.
15. Hennecke J, Sebbel P, Glockshuber R. Random circular permutation of DsbA reveals segments that are essential for protein folding and stability. J Mol Biol 1999;286:1197–1215.
16. Lindberg MO, Tangrot J, Otzen DE, Dolgikh DA, Finkelstein AV, Oliveberg M. Folding of circular permutants with decreased contact order: general trend balanced by protein stability. J Mol Biol 2001;314:891–900.
17. Iwakura M, Nakamura T, Yamane C, Maki K. Systematic circular permutation of an entire protein reveals essential folding elements. Nat Struct Biol 2000;7:580–585.
18. Grantcharova VP, Baker D. Circularization changes the folding transition state of the src SH3 domain. J Mol Biol 2001;306:555–563.
19. Clementi C, Jennings PA, Onuchic JN. Prediction of folding mechanism for circular-permuted proteins. J Mol Biol 2001;311: 879–890.
20. Li L, Shakhnovich EI. Different circular permutations produced different folding nuclei in proteins: a computational study. J Mol Biol 2001;306:121–132.
21. Plotkin SS, Onuchic JN. Investigation of routes and funnels in protein folding by free energy functional methods. Proc Natl Acad Sci USA 2000;97:6509–6514.
22. Plotkin SS, Wang J, Wolynes PG. Statistical mechanics of a correlated energy landscape model for protein folding funnels. J Chem Phys USA 1997;106:2932–2948.
23. Plotkin SS, Onuchic JN. Structural and energetic heterogeneity in protein folding: I. Theory. J Chem Phys USA 2002;116:5263–5283.
24. Sadqi M, Casares S, Abril MA, Lopez-Mayorga O, Conejero-Lara F, Freire E. The native-state conformational ensemble of the SH3 domain from alpha-spectrin. Biochemistry 1999;38:8899–8906.
25. Grantcharova VP, Riddle DS, Baker D. Long-range order in the src SH3 folding transition state. Proc Natl Acad Sci USA 2000;97: 7084–7089.
26. Gō N. Theoretical studies of protein folding. Annu Rev Biophys Bioeng 1983;12:183–210.
27. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Identifying the protein folding nucleus using molecular dynamics. J Mol Biol 2000;296:1183–1188.
28. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256: 623–644.
29. Makarov DE, Keller CA, Plaxco KW, Metiu H. How the folding rate constant of simple-single domain proteins depends on number of native contacts. Proc Natl Acad Sci USA 2002;99:3535–3539.
30. Shoemaker BA, Wang J, Wolynes PG. Exploring structures in protein folding funnels with free energy functionals: the transition state ensemble. J Mol Biol 1999;287:675–694.
31. Shoemaker BA, Wolynes PG. Exploring structures in protein folding funnels with free energy functionals: the denatured ensemble. J Mol Biol 1999;287:657–674.
32. Riddle DS, Grantcharova VP, Santiago JV, Alm E, Ruczinski I, Baker D. Experiment and theory highlight role of native state topology in SH3 folding. Nat Struct Biol 1999;6:1016–1024.
33. Chan HS. Protein folding: matching speed and locality. Nature 1998;392:761–763.
34. Baker D. A surprising simplicity to protein folding. Nature 2000;405:39–42.
35. Gsponer J, Caflisch A. Role of native topology investigated by multiple unfolding simulations of four SH3 domains. J Mol Biol 2001;309:285–298.
36. Plotkin SS, Wolynes PG. Buffed energy landscapes: another solution to the kinetic paradoxes of protein folding. Proc Natl Acad Sci USA 2003;100:4417–4422.
37. Galzitskaya OV, Finkestein AV. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. Proc Natl Acad Sci USA 1999;96:11299–11304.
38. Alm E, Baker D. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. Proc Natl Acad Sci USA 1999;96:11305–11310.
39. Munoz V, Eaton WA. A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc Natl Acad Sci USA 1999;96:11311–11316.
40. Kaya H, Chan HS. Energetic components of cooperative protein Folding. Phys Rev Lett 2000;85:4823–4826.
41. Munoz V. Folding plasticity. Nat Struct Biol 2002;11:792–794.