PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Identifying folding nucleus based on residue contact networks of proteins

Jie Li,[1,2] Jun Wang,[1] and Wei Wang[1]*

[1] National Laboratory of Solid State Microstructure, Department of Physics, Nanjing University, Nanjing 210093, China
[2] The State Key Laboratory of Pharmaceutical Biotechnology, School of Life Science, Nanjing University, Nanjing 210093, China

ABSTRACT

In the native structure of a protein, all the residues are tightly parked together in a specific order following its folding and every residue contacts with some spatially neighbor residues. A residue contact network can be constructed by defining the residues as nodes and the native contacts as edges. During the folding of small single-domain proteins, there is a set of contacts (or bonds), defined as the folding nucleus (FN), which is formed around the transition state, i.e., a rate-limiting barrier located at about the middle between the unfolded states and the native state on the free energy landscape. Such a FN plays an essential role in the folding dynamics and the residues, which form the related contacts called as folding nucleus residues (FNRs). In this work, the FNRs in proteins are identified by using quantities which characterize the topology of residue contact networks of proteins. By comparing the specificities of residues with the network quantities K^R, L^R, and D^R, up to 90% FNRs of six typical proteins found experimentally are identified. It is found that the FNRs behave the full-closeness centrals rather than degree or closeness centers in the residue contact network, implying that they are important to the folding cooperativity of proteins. Our study shows that the FNRs can be identified solely from the native structures of proteins based on the analysis of residue contact network without any knowledge of the transition state ensemble.

Proteins 2008; 71:1899–1907.

Key words: folding nucleus; folding nucleus residues; transition state; residue contact networks of proteins.

INTRODUCTION

It is widely accepted that folding of small single-domain proteins follows the nucleation-condensation mechanism and shows a two-state behavior. The transition state (TS), i.e., a rate-limiting barrier, is located at about the middle between the unfolding states and the native state on the free energy landscape.[1–3] It is found that around the TS there are key contacts (or bonds) which are defined as folding nuclei (FNs), and the related residues of these contacts are called as folding nucleus residues (FNRs). The FNs and the related FNRs play an essential role in the folding dynamics. There are two kinds of the FNs, namely the critical FNs and the postcritical FNs. The critical FNs drive fast folding when they are formed in the TS, and the conformations where the critical FNs are formed have equal folding probability (or 1/2) to be the unfolded or the native. The postcritical FNs drive inevitable and fast folding when they are formed after the TS. That is, once these key contacts between the FNRs are well formed, a framework of the native conformation of the protein is well established and then the other non-FN residues are pulled to form native contacts by the FNRs. The folding is downhill to the native state along the free energy funnel rapidly. Thus, the features of the postcritical FN should be characterized for studying the folding behavior of the two-state proteins since the subsequent folding of the proteins proceeds after such a nucleus is formed (see a detailed discussion in Ref. 1). Precisely, this nucleus or the minimal set of contacts which seeds the folding can be found from conformations around the TS because the postcritical FN is close to the barrier state.[1] Therefore, to characterize the FN, a transition state ensemble (TSE) needs to be collected and the related FNRs should be identified. To identify the FNRs and characterize the features of the FN is very important for studying the folding and evolution of proteins, as well as for designing sequences of proteins with specific structures.

Many experimental and theoretical efforts have been made to identify the FNRs and to characterize the feature of the FN.[4–16] In experiments, the FNRs are identified generally by the so-called ϕ-value. The ϕ-value is defined as the ratio of the changes in stability between the TS and the native state when a single site mutation on residue is made in the protein,

Grant sponsor: NSF of China; Grant numbers: 90403120, 10474041; Grant sponsor: National Basic Research Program; Grant numbers: 2006CB910302, 2007CB814800; Grant sponsor: FANEDD.
*Correspondence to: Wei Wang, National Laboratory of Solid State Microstructure, Department of Physics, Nanjing University, Nanjing 210093, China. E-mail: wangwei@nju.edu.cn
Received 22 June 2007; Revised 8 October 2007; Accepted 25 October 2007
Published online 3 January 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21891

i.e., $\phi = \Delta\Delta G_{\ddagger-U}/\Delta\Delta G_{N-U}$. Note that $\Delta\Delta G_{\ddagger-U} = (G_{\ddagger}-G_U)_m-(G_{\ddagger}-G_U)_w$ (or $\Delta\Delta G_{N-U} = (G_N-G_U)_m-(G_N-G_U)_w$) is the change of free-energy differences between the TS (or the native state) and the unfolded states due to the mutation. Here the $G_{\ddagger}$, G_U, and G_N are the free energy at the TS, the unfolded states, and the native state, and the subscripts "w" and "m" represent the wild-type and the mutated case of the protein, respectively.[4] Generally, the ϕ-value measures the contribution of the mutated residue to the stability of the TS. Residues with high ϕ-values usually take important role in formation and stability of the TS and are relevant to the FNRs. While in theory, by monitoring the performances of various contacts in the TS following the folding trajectories, residues with important contacts are identified as the FNRs.[5,6] Thus, no matter in experiment or theory, the knowledge of the TSE and the native state is needed for the identification. However, it is not easy to obtain a well-defined TSE since intensive simulations on folding should be made, resulting in that the identification of nuclei is difficult.[17,18]

In the native state of a protein, balance of various contacts between residues stabilizes the native conformation. If residues are regarded as nodes and contacts as edges between the residues, the native structure can be transformed into a residue contact network[19–23] although the linear chain of the protein introduces some constraints for local residues. As a result, some quantities, which characterize the features of various complex networks proposed in different fields, can be used to study the properties of the residue contact networks. Thus, the interactions between the residues in the native state and the topology of the native state can be well described, and the folding behavior can be characterized since the topology of the native state is the dominant factors for the folding of proteins.[24–31] Furthermore, from the quantification of the properties of the residue contact network, some important residues and contacts should present quite special characters, which may be related to the FN. Therefore, the FNRs of a protein could be worked out from its native conformation only by characterizing the properties of the residue contact networks.

In this article, after transforming the native structures of proteins into residue contact networks, three quantities associated with the relative solvent accessibility (RSA) are used to identify the nuclei. It is shown that up to 90% FNRs found experimentally could be identified correctly for six proteins, and that the FNRs are those residues which are essential to the folding dynamics of a protein.

MATERIALS AND METHODS

Construction of the residue contact network

As to the native structure of a protein, residues can be regarded as nodes and contacts as edges. Here, a contact is defined when any two nonhydrogen atoms from residues i and j are within a distance $r_c = 4.5$ Å. The contacts then can be transformed to an adjacent matrix $A_{ij} = H(r_c - r_{ij})$ for $|i - j| > 3$ and zero otherwise. Here $H(x)$ is the Heavyside step function with $H(x) = 1$ for $x > 0$ and $H(x) = 0$ for $x \leq 0$. An example of the transformation for protein src-SH3 is shown in Figure 1.

Quantities characterizing the topology of network

Similar to the quantities used to characterize the complex network, we defined three quantities to study the topology of the residue contact network. They are the degree K_i, the path length L_i,[32] and the specialty D_i for node-i.

The degree of node-i, K_i, is defined as the number of neighbors of node-i,

$$K_i = \sum_{j=1}^{N} A_{ij}, \tag{1}$$

where N is the total number of the nodes in the network. The path length L_i for node-i is defined as the shortest path length from node-i to another node-j on average,

$$L_i = \frac{1}{N_i}\sum_{j\neq i}^{N} l_{ij}, \tag{2}$$

where l_{ij} represents the shortest path length from node-i to node-j, and N_i is the total number of nodes from node-i reaching to other nodes through edges. Clearly, the average path length L for the whole network is

$$L = \frac{1}{N}\sum_{i=1}^{N} L_i. \tag{3}$$

Obviously, a high value of K_i means that the node-i has a large number of edges and is a degree central in the network,[23] and a low value of L_i means that the node-i has short transmitting path from the node to all the related nodes and is a closeness central.[23] Since K_i characterizes the local topology of the network while L_i the global nature, the degree centrals are the local centrals while the closeness centrals are the global centrals in the network. These centrals are important to the topology of protein structures and usually related to the key residues in proteins.[23]

Furthermore, in the error analysis of network,[32,33] the specialty of a node can be obtained by comparing changes of some quantities, say L, for the network when the node is removed with respect to the original case. Thus the specialty of node-i can be described by the difference of L between these two cases using

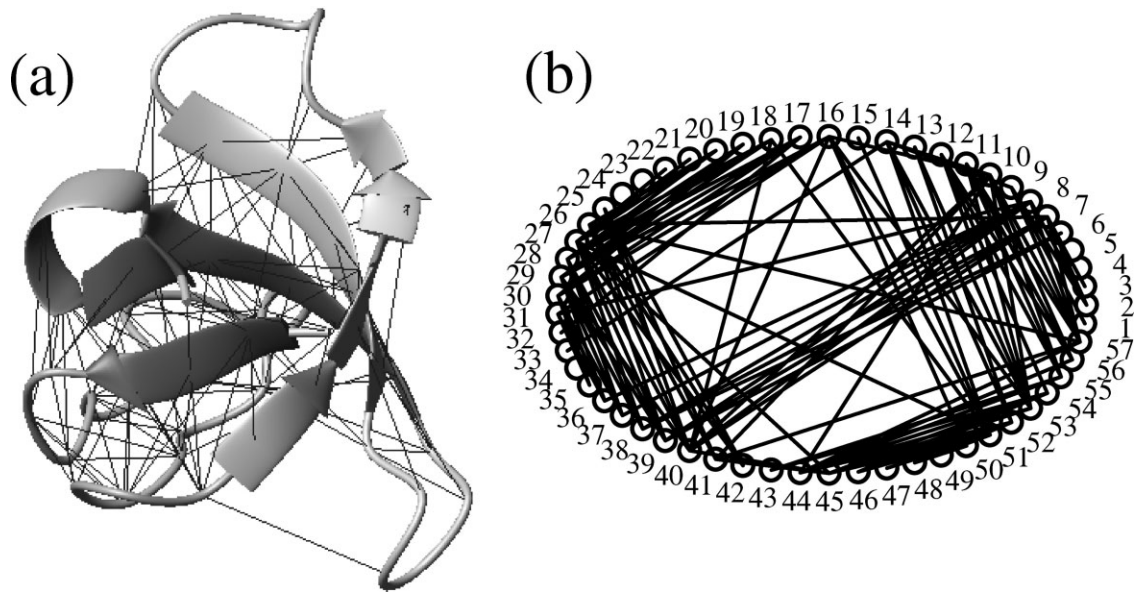

Figure 1

*The native structure of protein src-SH3 (pdb code: 1nlo) (**a**), and the related residue contact network (**b**). Dark area in (a) are the hydrophobic core and the lines are the noncovalent contacts.*

$$D_i = L_i^{\mathrm{RM}} - L^{\mathrm{WD}}. \quad (4)$$

Here L^{WD} is the average path length for the network of the wild-type of the protein and L_i^{RM} is the average path length for the error network when node-i is removed. Note that as described earlier, L_i scales the closeness from node-i to other nodes while D_i scales the effect of node-i taking the closeness between all the pairs of nodes in the network differently. This means that D_i includes the effect of node-i acting as a bridge in the shortest path between two other nodes. Thus, D_i characterizes the contribution of node-i to the closeness of more pairs of nodes in network than that of L_i, and nodes with high values of D_i can be termed as the full-closeness centrals with respect to the closeness centrals.

Quantities for identification on the FNRs

As mentioned earlier, the FNRs take important roles to the folding of proteins and identifying them is very useful to understand the mechanism of protein folding. Generally, there are two key factors, i.e., the geometrical one and the energetic one which determine the folding process of a protein together. The formation of the FN, so the key contacts for folding, is affected by both factors. This means that, to accomplish a good identification on the FNRs, these two factors should be considered. In our identification, the geometrical factor is considered by the three quantities related to the contact network which characterize the topology of the native structures of proteins. Thus, to consider the energetic factor, the RSA of a residue in the native structure of a protein,[34] which represents the effect of the hydrophobic interactions, a main driving force for folding, is included in our identification. Intuitively, residues buried in the interior of a protein molecule would have small values of the RSA while residues exposed to solvent would have large values of the RSA. Additionally, both geometry and stability of proteins imply that the FNRs often prefer to appear in the interior of the protein molecule.[34] Therefore, we can further formulate quantities K_i, L_i, and D_i by taking into account the factor RSA as follows

$$K_i^{\mathrm{R}} = K_i/\mathrm{RSA}_i, \quad (5)$$

$$L_i^{\mathrm{R}} = L_i \times \mathrm{RSA}_i, \quad (6)$$

$$D_i^{\mathrm{R}} = D_i/\mathrm{RSA}_i. \quad (7)$$

Such kind of combinations bias to the residues in the interior of protein molecule, and improves the efficiency of our identification on the FNRs. Clearly, the FNRs would have large values of K_i, small values of L_i, and large values of D_i (see results). The values of the RSA in our study are evaluated by the software MOLMOL.[35] However, as argued in Ref. 23, the RSA and the network-based quantities K_i, L_i, and D_i in Eqs. (5)–(7) are independent.

It is worthy noting that the identification is realized by defining a reference line for the values of quantities K_c^R, D_c^R, or L_c^R (defined as X_c). When residue-i has a value of $K^R \gg K_c^R$, $D^R \gg D_c^R$, or $L^R \ll L_c^R$, this residue is picked out as a possibly FNR. Here X_c is obtained via a statistical analysis on the average distribution of the quantity X. In detail, assuming that the quantity X for various residues follows a Gaussian distribution, the most probable value X_p of the quantity X is taken as the central of the distribution, and the variation could be calculated with the regular procedure, $\sigma_X = <(X - X_p)^2>$ where the average runs over all residues of the concerned protein. Then X_c is picked as the $X_p + \sigma_X$ for K^R and D^R, or $X_p - \sigma_X$ for L^R to ensure that the identified residues are significantly different from average. Note that small deviation of X_c may introduce some false nucleus in our identification, but it would not affect the successful identification on correct nucleus.

Six proteins to be identified

Six proteins, i.e., src-SH3 (1nlo), CI2 (1ypc), Acp (1aps), S6 (1ris), U1A (1urn), and Tendamistat (2ait) are selected to test our procedure. All these proteins fold following the nucleation-condensation mechanism and their FNRs have been well identified experimentally and simulationally.[4–16]

All the possible FNRs for these six proteins are collected from previous experiments and simulations, and are classified into two groups, namely the assured nuclei residue (ANR) and the unsure nuclei residue (UNR). An ANR is defined when it is confirmed both experimentally and theoretically, while an UNR is declared only experimentally or theoretically. In our study, the ANRs are the exclusive benchmark, and the UNRs are included as a complementary reference. In addition, a residue which is neither the ANR nor the UNR but identified as a nucleus in our procedure, is termed as a possible nucleus residue (PNR). To evaluate our identification procedure, both the sensitivity (SE) and the specificity (SP) for the identification are defined. The SE of identification of the ANRs is defined as the number of identified ANRs divided by the total number of ANRs, and the SP as the total number of the ANRs divided by the number of our identified residues, respectively.[23]

RESULTS AND DISCUSSION

First, let us present the results for src-SH3 which has 57 residues and consists of five β-strands, a 3_{10}-helix, a RT loop, and a distal hairpin. Among them, β-strand-2 and β-strand-3 are orthogonally packed as the hydrophobic core [see the dark area in Fig. 1(a)]. The nuclei are formed via a hydrogen bonding network together with some residues in the hydrophobic core in β-strand-2, β-strand-3, and 3_{10}-helix.[7–9] This network involves residues Glu30, Ser47, and Thr50, while the hydrophobic core involves Phe10, Leu24, Ile34, Ala45, and Ile56. These residues concrete the packing of β-strand-2, β-strand-3, and 3_{10}-helix by local or nonlocal hydrogen bonds in the TSE, and mutations on them increase the free energies of the native state dramatically. These suggest that the eight residues are important to form and stabilize the native state and can be termed as the ANRs.[7–9] Four residues, i.e., Tyr16, Phe26, Leu32, and Val61, located inside or around the hydrophobic core, were argued to be the extensions of eight ANRs.[8] In the TSE, they hold the terminals of β-strand-2, β-strand-3, and 3_{10}-helix together, enabling the β-β-3_{10} pattern to be stable. Mutations on them also make the native state unstable.[8,9] These four residues are termed as the UNRs. To play the identification of the nuclei of SH3, its native structure is transformed into a residue network [see Fig. 1(b)], and then the related quantities K^R, L^R, and D^R (see Materials and Methods) are calculated and shown in Figure 2(a–f) in which the solid and cross circles are for the ANRs and UNRs, respectively. It is seen that most of the solid and cross circles are above the reference lines (see Materials and Methods) with high values of K^R and D^R, but low values of L^R (inversely axis), indicating well identification of the nuclei. Six residues out of the eight ANRs, and four UNRs, as well as seven PNRs in Figure 2(a,b); five out of the eight ANRs, four UNRs, as well as three PNRs in Figure 2(c,d); and eight ANRs, and four UNRs, as well as four PNRs in Figure 2(e,f) are identified. Thus, the values of SE of the identification of the ANRs using K^R, L^R, and D^R are 75.0, 62.5, and 100.0%, and the related values of SP are 35.3, 41.7, and 50.0%, respectively (Table I). Obviously, the values of SE and SP using D^R are higher than those using K^R and L^R, providing that the nuclei of src-SH3 can be well identified using D^R based on the native structure.

Similarly, nuclei in five other proteins are also identified (Table I). The averaged values of SE are 75.6, 77.7, and 90.5%, while the related SP are 40.7, 36.2, and 51.1%, respectively. These identifications using the combined quantities K^R, L^R, and D^R are generally better than those using the original quantities K_i, L_i, and D_i since the contribution of hydrophobicity in protein folding is included (see Materials and Methods). Especially, the values of SE are all increased about 10–20% (data are not presented) after the combinations. Note that a test solely using the RSA is worse than the identifications using K_i, L_i, and D_i (data not shown). Specifically, the ability in identification using D^R is stronger than those using K^R and L^R, indicating that D^R is a more effective quantity. This optimal identification using D^R can be understood by analyzing the properties of three quantities K_i, L_i, and D_i for the network because the RSA is uncorrelated with these network-based quantities. Usually, FNRs play important roles to the folding cooperativity of the proteins. That is, once the contacts between the FNRs are well

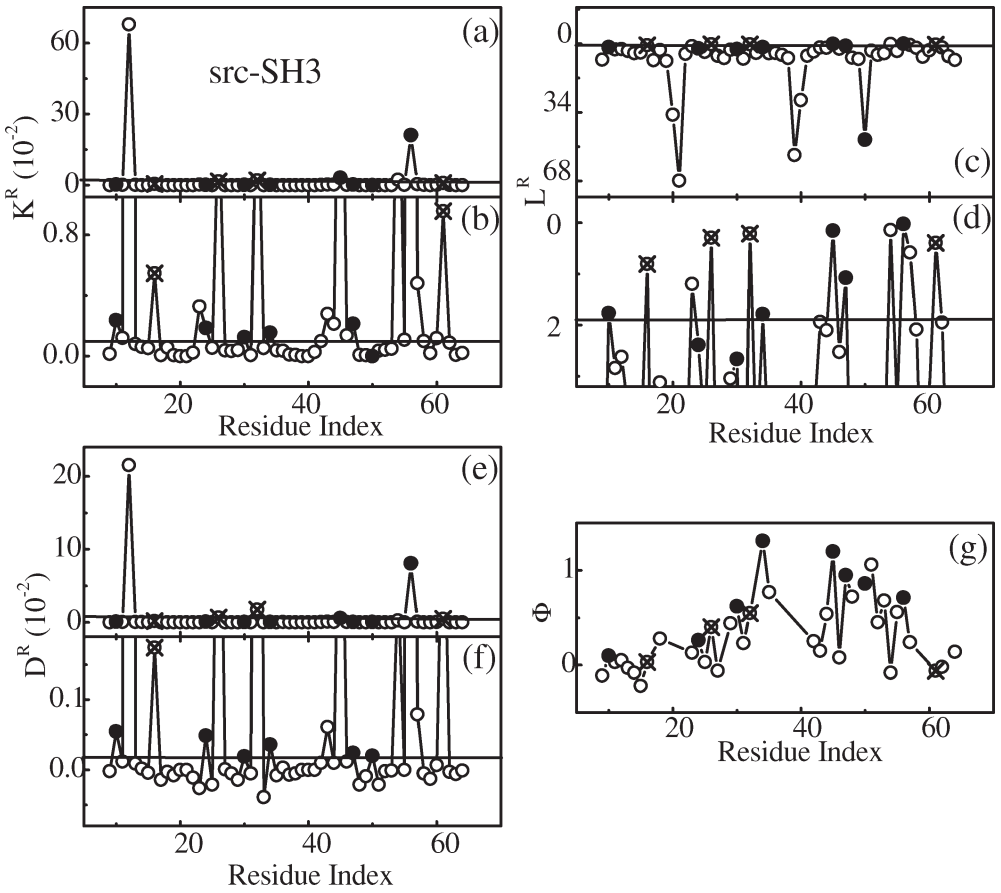

Figure 2

Identification for src-SH3 using K^R (***a,b***), L^R (***c,d***), D^R (***e,f***) *and experimental* ϕ*-values* (***g***). *(b), (d), and (f) show the detailed plotting of (a), (c), and (e) around the reference lines, respectively. The solid circles represent the ANRs and the cross circles present the UNRs. Some open circles are above the reference lines, showing identification of some possible nuclei. Note that the residue index is following the way in the related PDB file of the protein.*

formed, the other non-FN residues are pulled by the FNRs, and then the whole protein folds to its native state rapidly. Such a folding cooperativity is relevant to the special connections between nodes in the residue contact network, which relates to the average path length L (see Materials and Methods) of the whole residue network because the more less the L is, the more quickly the residues response to the pulling of by other residues. This can be further described by the quantity D_i according to Eq. (4). Differently, both K_i and L_i cannot characterize the coherent dynamics of the related contact network. The reason is as follows. K_i is a local quantity since it only related to the number of contacts of residue-i, and L_i is not sufficient to describe the global property of the network because it does not include the effect of all residue pairs in the network. Therefore, after combining these three quantities for the network with RSA, the identifications by D^R are more effective than those of K^R and L^R.

According to the above discussions on the identifications by using K, L, D, K^R, L^R, and D^R, it can be concluded that the quantity D^R is the best one. Thus, our following discussions on the nuclei identification for the six proteins are focused on the identifications by using D^R. Detailed identifications of these nuclei using D^R are shown in Figure 3. It is seen that the identifications perform very well. For example, all ANRs for protein SH3, CI2, UIA, and Tendamistat are identified, and only one ANR, i.e., residue Phe60, for S6 is not identified. Besides, four ANRs out of seven ANRs for Acp are identified. The identifications of the UNRs also perform well. Except residue Leu69 for U1A and residue Asp11 for Tendamistat, all other UNRs for the six proteins are identified. In addition, the related ϕ-values for these proteins are plotted in Figures 2(g) and 3(b,d,k), respectively. Note that the plots of ϕ-values for protein U1A and Tendamistate are absent. This is because that for protein U1A the ϕ-values in different denaturant concentrations[12] rather

Table I
Values of SE and SP for the identification using K^R, L^R, and D^R for six proteins. Detailed identification of the ANRs, the UNRs, and the PNRs using D^R are also listed

Proteins (PDB ID)	K^R SE (%)	K^R SP (%)	L^R SE (%)	L^R SP (%)	D^R SE (%)	D^R SP (%)	ANs	UNs	PNs	ANs K^R	UNs L^R	PNs D^R
src-SH3 (1nlo)	75.0	35.3	62.5	41.7	100.0	50.0	10, 24, 30, 34, 45, 47, 50, 56	16, 26, 32, 61	12, 43, 54, 57	+, +, +, +, +, +, +, +	+, +, +, +	+, +, +, +
CI2 (1ypc)	100.0	42.8	100.0	33.3	100.0	50.0	35, 68, 76	39, 66	24	+, +, +	+, +	+
S6 (1ris)	75.0	26.1	100.0	32.0	100.0	42.1	6, 8, 26, 30, 65, 75, [60]	10, 63, 67, 79	22, 64	+, +, +, +, +, +, [+]	+, +, +, +	+, +
U1A (1urn)	85.7	40.0	71.4	42.8	85.7	50.0	14, 17, 26, 30, 34, 40, 43, 45	12, 55, 58, 84, [69]	11, 57, 59, 65, 68, 86, 87	+, +, +, +, +, +, +, +	+, +, +, +, [+]	+, +, +, +, +, +, +
Acp (1aps)	42.8	50.0	57.2	30.1	57.2	57.1	11, 25, 30, 94, [45, 47, 54]	13, 35	53	+, +, +, +, [+, +, +]	+, +	+
Tendamistat (2ait)	75.0	50.0	75.0	37.5	100.0	57.1	12, 22, 25, 33, 34, 35, 56, 70	16, 59, 71, [11]	21, 23, 37	+, +, +, +, +, +, +, +	+, +, +, [+]	+, +, +
Average	75.6	40.7	77.7	36.2	90.5	51.1						

Unidentified ANRs and UNRs are marked with square brackets. Signs of K^R, L^R, and D^R for the related ANRs, UNRs, PNRs are listed in the last three columns.

than in water are provided in experiment, and for protein Tendamistate there are no related experimental results. As to proteins src-SH3, CI2, S6, and Acp, the ANs are well identified by using ϕ-values since they usually have distinctively high ϕ-values. This is consistent with our identifications on ANs by using D^R, indicating that the nuclei areas identified by using the ϕ-values and D^R are generally the same [see Figures 2(g) and 3(b,d,k)]. However, there are limitations for the identifications by using ϕ-values since the ϕ-values for some residues in a protein are often absent. For example, most ϕ-values of UNs for protein Acp are absent [see Fig. 3(k)]. Thus, based on a study on all the residues in proteins, our identifications by using D^R can provide more candidates for the FNRs of proteins.

Now let us make a detailed discussion for protein CI2 which has 64 residues and is composed of an α-helix packed against six β-strands [Fig. 4(b)]. The nuclei confirmed by experiments or/and simulations are the α-β-β area including the α-helix, β-strand-3, and β-strand-4. Three ANRs, i.e., residues Ala35, Leu68, and Ile76 locate at the center of the nucleus area,[4,5,10,11] and two UNRs, i.e., residues Ile39 and Val66, locate at the nucleus margin.[5] The related interactions of three ANRs were found to be the key interactions to build up the α-β-β pattern in the TSE.[4,5] The two UNRs were not confirmed by experiments, but were argued to be important to the TSE since their native degrees in TSE are not only very close to those of the three ANRs, but also much higher than those of all other nonnucleus residues.[4,5] In the TSE, Ile39 located in the α-helix mainly connects with the β-strand-3, and contributes to the stabilization of the nuclei area between the α-helix and β-strand-3, while Val66 located in the β-strand-4 mainly connects with the N-terminal of the α-helix, and contributes to the stabilization of the nuclei area between the α-helix and β-strand-4. Using D^R, six residues, i.e., three ANRs, two UNRs, and one PNR [Fig. 3(a) and Table I], are identified, which is well consistent with those determined previously by others.

So far, our identification of the ANRs and UNRs is very efficient, although some PNRs are also found (Figs. 2 and 3 and Table I). These PNRs could be checked in detail one by one based on the native structures (Fig. 4). For example for src-SH3 [Fig. 4(a)], four PNRs, i.e., residues Ala12, Trp43, Gly54, and Pro57, are distributed in the neighbors of the ANRs (i.e., Ala12 near to Phe10, Trp43 near to Ala45, Gly54, and Pro57 near to Ile56). Since the orientations of their side chains are about the same as those of their neighboring nuclei, e.g., the side chain of Ala12 has basically the same orientation as Phe10 and faces to the nucleus area. Note that Ala12 has a special high value of D^R [Fig. 2(e)] since it is a member in the hydrophobic core with extremely low value of *RSA*. Thus, presumably, these four PNRs play the same important role in the TSE as the neighboring nucleus.

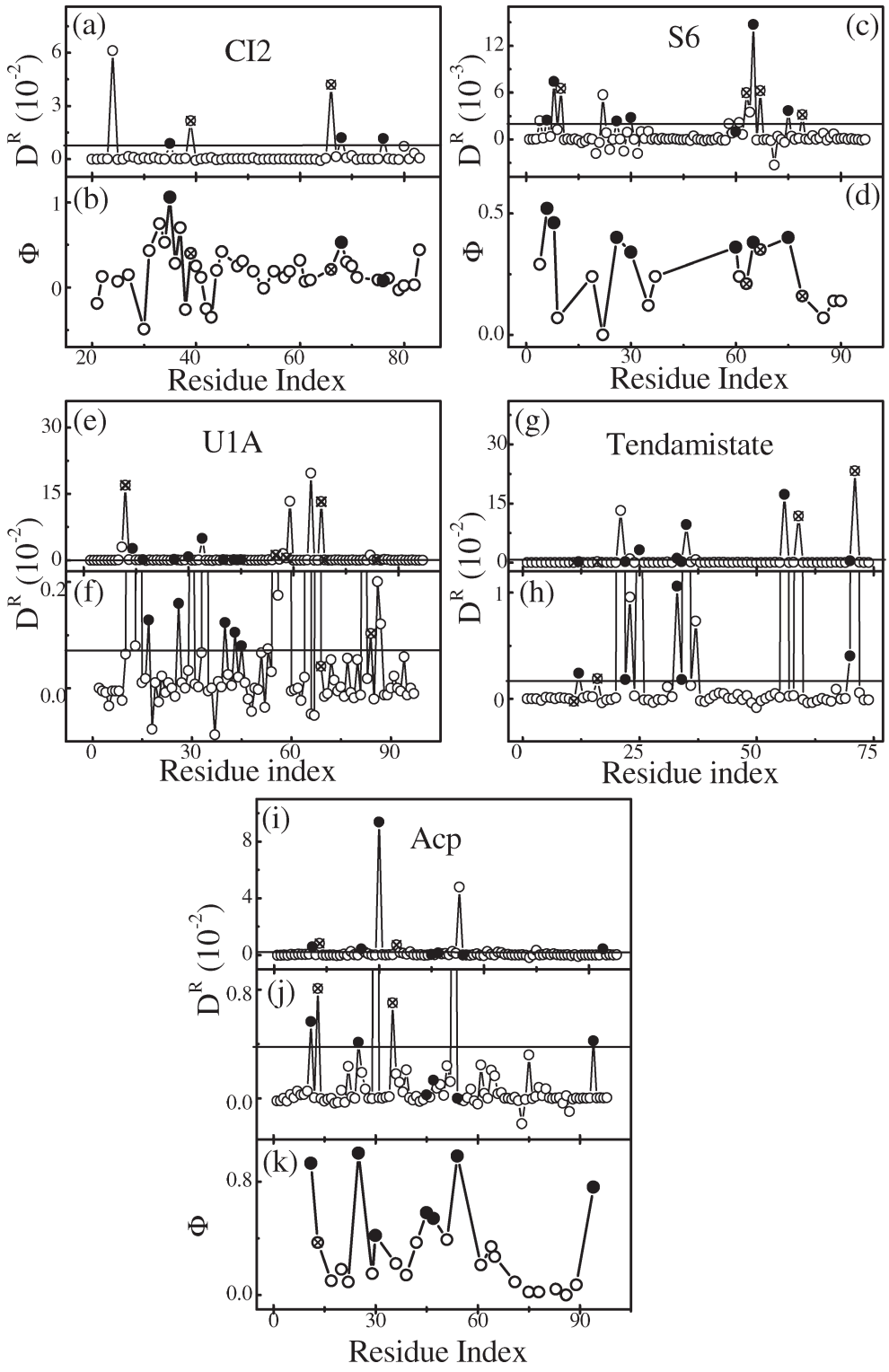

Figure 3

*D^R versus the residue index for CI2 (**a**), S6 (**c**), U1A (**e,f**), Tendamistat (**g,h**), and Acp (**i,j**), respectively. Note that the related ϕ-values for CI2, S6, and Acp are in (**b**), (d), and (k), respectively. Similar to Figure 2, (f), (h), and (j) show the detailed plotting around the reference lines.*

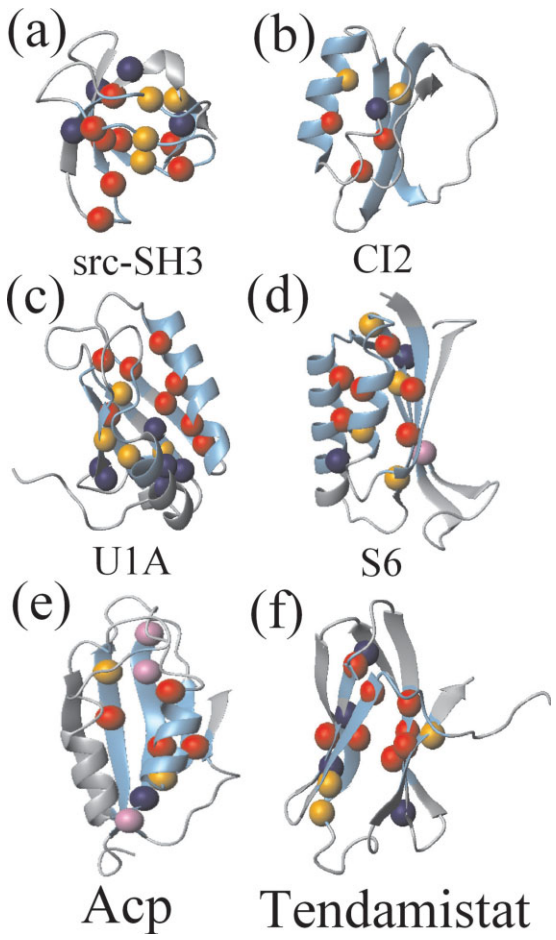

Figure 4

*Distribution of nuclei in the native structures for src-SH3 (**a**), CI2 (**b**), U1A (**c**), S6 (**d**), Acp (**e**), and Tendamistat (**f**). The nuclei are marked in sky blue. The identified (or unidentified) ANRs are plotted in red (or in violet), and the identified (or unidentified) UNRs are in orange (or in green), while the PNRs are in navy.*

They were not claimed as nuclei probably because they were not well characterized or/and their ϕ values are not the local maxima compared with their neighboring nuclei. For CI2, the only PNR, i.e., residue Trp24, is also in the area of the hydrophobic core and locates in the type-III reverse turn connecting with the N-terminal of β-strand-1 in the TSE,[4,5,10] and may not contribute to the formation or stability of the nuclei. However, it is important to the collapse of β-strand-1 in the transiting process when the nuclei expending to the postnuclei according to related simulations.[11] Additionally, an interesting case is for U1A. Seven PNRs, i.e., residues Thr11, Val57, Phe59, Ala65, Ala68, Tyr86, and Ala87, all located in the hydrophobic core are identified. It was argued that there are two groups of nuclei[12,13] [Fig. 4(c)]. Residues Val57 and Phe59, both close to the UNR Ile58, are in the area of the main nuclei (nuclei-I), affecting the formation and stabilization of nuclei-I. All other five PNRs are in the area of the small nuclei (nuclei-II), and correlate strongly to the related three UNRs in nuclei-II.[12] Therefore, it is reasonable to believe that these five PNRs play important role in the formation of nuclei-II by attracting each other strongly. It is worth noting that some PNRs, especially the ones neighboring to the related ANRs in sequence, may be the false positive FNRs. However, as shown in this work, our method provides an efficient way to coarsely identify the nuclei area of proteins.

CONCLUSIONS

In conclusion, only based on the residue contact network constructed from the native structures of proteins, the folding nuclei can be well identified without the knowledge of the TS. Such good identification using quantities, which characterize the topology of the residue contact network of the native structure of the protein, indicates that the nuclei are globally full-closeness centrals rather than degree or closeness centrals and play very important roles in the native structure formation and stabilization. Actually, the protein folding can be analogous as a building process of a network for communication between nodes. A general scheme is first to set some important nodes as full-closeness centrals, and then connect them together. Thus, the communication between all the nodes can be realized by linking the nodes to the centrals directly or indirectly. Our work shows an effective way to find such centrals in proteins. Finally, it is noted that although different definitions of contacts, such as using different cutoff distance r_c, will lead to different residue contact networks of a same protein, the accuracy of the identification in the FNRs is not changed basically. This may be due to that our identification is mainly based on the global property of the residue contact network, while different definitions of contacts only affect the residue contact network locally.

REFERENCES

1. Abkevich V, Gutin A, Shakhnovich E. Specific nucleus as the transition state for protein folding: evidence from the lattice model. Biochemistry 1994;33:10026–10036.
2. Pande V, Grosberg AY, Rokshar D, Tanaka T. Pathways for protein folding: is a "new view" needed? Curr Opin Struct Biol 1998;8:68–79.
3. Fersht AR. Nucleation mechanism of protein folding. Proc Natl Acad Sci USA 2000;97:1525–1529.
4. Fersht AR. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. Proc Natl Acad Sci USA 1995;92:10869–10873.
5. Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 2000;298:937–953.

6. Dokholyan N, Buldyrev S, Stanley H, Shakhnovich E. Identifying the protein folding nucleus using molecular dynamics. J Mol Biol 2000;296:1183–1188.
7. Gruebele M, Wolynes PG. Satisfying turns in folding transitions. Nat Struct Biol 1998;5:662–665.
8. Grantchanova VP, Riddle DS, Santiago JV, Baker D. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src-SH3 domain. Nat Struct Biol 1998;5: 714–720.
9. 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.
10. Itzhaki L, Otzen D, Fersht AR. The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. J Mol Biol 1995;254:260–288.
11. Li L, Shakhnovich EI. Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations. Proc Natl Acad Sci USA 2001;98:13014–13018.
12. Ternström T, Mayor U, Akke M, Oliveberg M. From snap-shot to movie: phi-value analysis of protein folding transition states taken one step further. Proc Natl Acad Sci USA 1999;96:14854–14859.
13. Shen TY, Hofmann CP, Oliveberg M, Wolynes PG. Scanning malleable transition state ensembles: comparing theory and experiment for folding protein U1A. Biochemistry 2005;44:6433–6439.
14. Vendruscolo M, Dokholyan NV, Paci E, Karplus M. Small-world view of the amino acids that play a key role in protein folding. Phy Rev E 2002;65:061910.
15. Hubner IA, Oliveberg M, Shakhnovich EI. Simulation, experiment, and evolution: understanding nucleation in protein S6 folding. Proc Natl Acad Sci USA 2004;101:8354–8359.
16. Qin M, Zhang J, Wang W. Effects of disulfide bonds on folding behavior and mechanism of the β-sheet protein Tendamistat. Biophys J 2005;90:272–286.
17. Ozkan SB, Bahar I, Dill KA. Transition states and the meaning of Phi-values in protein folding kinetics. J Mol Biol 2001;8:765–769.
18. Sänchez IE, Kiefhaber T. Origin of unusual ϕ-values in protein folding: evidence against specific nucleation sites. J Mol Biol 2003;334:1077–1085.
19. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. Proteins 2001;44:150–165.
20. Jung J, Lee J, Moon HT. Topological determinants of protein unfolding rates. Proteins 2005;58:389–395.
21. Greene LH, Higman VA. Uncovering networsk systems within protein structures. J Mol Biol 2003;334:781–791.
22. Atilgan AR, Akan P, Baysal C. Small-world communication of residues and significance for protein dynamics. Biophys J 2004;86:85–91.
23. Amitai G, Shemesh A, Sitbon E, Shklar M, Venger DNI, Pietrokovski S. Network analysis of protein structures identifies functional residues. J Mol Biol 2004;344:1135–1146.
24. Munõz V, Eaton W. A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc Natl Acad Sci USA 1999;96:11311–11316.
25. Takada S. Go-ing for the prediction of protein folding mechanisms. Proc Natl Acad Sci USA 1999;96:11698–11700.
26. Alm E, Baker D. Prediction of proteinfolding mechanisms from free-energy landscapes derived from native structures. Proc Natl Acad Sci USA 1999;96:11305–11310.
27. Klimov DK, Thirumalai D. Native topology determines force-induced unfolding pathways in globular proteins. Proc Natl Acad Sci USA 2000;97:7254–7259.
28. Ozkan SB, Dill KA, Bahar I. Computing the transition state populations in simple protein models. Bioploymers 2003;68:35–46.
29. Chen H, Zhou X, Liaw CY, Koh CG. Kinetic analysis of protein folding lattice models. Mod Phys Lett B 2004;18:163–172.
30. Bai YW, Zhou HY, Zhou YQ. Critical nucleation size in the folding of small apparently two-state proteins. Proten Sci 2004;13:1173–1181.
31. Simler BR, Levy Y, Onuchic JN, Matthews CR. The folding energy landscape of the dimerization domain of *Escherichia coli* trp repressor: a joint experimental and theoretical investigation. J Mol Biol 2006;363:262–278.
32. Albert R, Barabási A-L. Statistical mechanics of complex networks. Rev Mod Phys 2002;74:47–96.
33. Albert R, Jeong H, Barabási A-L. Error and attack tolerance of complex networks. Nature 2000;406:379–382.
34. Selvaraj S, Gromiha MM. Importance of hydrophobic cluster formation through long-range contacts in the folding transition state of two-state proteins. Proteins 2004;55:1023–1035.
35. Koradi R, Billeter M, Whrich K. MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graphics 1996; 14:51–55.