Construction and application of the weighted amino acid network based on energy

Xiong Jiao, Shan Chang, Chun-hua Li, Wei-zu Chen, and Cun-xin Wang*
College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100022, China
(Received 26 January 2007; published 4 May 2007)

A method is proposed to construct the weighted amino acid network. The weight of the link is based on the contact energy between residues. For the 197 proteins with low homology, the "small-world" property was studied based on this method. Additionally, analyses were carried out for the statistic characteristics of the network parameters, the influence of the weight on the network parameters, the network parameter difference of amino acids, and the links between the hydrophobic and hydrophilic residues. Using this method, we studied the network parameter change for the protein chymotrypsin inhibitor 2 (CI2) on its high-temperature unfolding pathway. It is found that the unfolding of the protein is mainly exhibited as the derogation of the hydrophobic core and the shortest path length rise in the unfolding process. This work is helpful for studies of protein folding and the relationship between structure and function using complex network theory.

DOI: 10.1103/PhysRevE.75.051903 PACS number(s): 87.14.Ee

I. INTRODUCTION

As the essential matter of life, a protein molecule can be treated as a complex network with each residue simplified as a node and the interaction between them as a link [1,2]. With the aid of research work on the complex network, some recent methods are used to study protein folding and the relation between structure and function. Vendruscolo *et al.* have looked for "key residues" through the analysis of the network parameter: betweenness [2]. By measuring the topology of the protein contact network, Dokholyan *et al.* have shown how the topological properties of the protein conformation determine its kinetic ability for folding [3]. Atilgan *et al.* have found that the average shortest path lengths are highly correlated with residue fluctuations through the analysis of the amino acid networks [4]. Amitai *et al.* have identified the active site residues by the network parameter: closeness [5]. Jacobs *et al.* have predicted the protein flexibility using the graph theory [6].

In the amino acid network, each residue is simplified to a single point to represent the network node and the links between these nodes are based on the distance between them. Generally, the C_α atom is chosen to be the network node and the link between a pair of nodes is determined by the distance between them. If the distance is less than the cutoff value 7.0 Å [4] (or 8.5 Å [2]), there exists a link. In the other method for constructing the amino acid network, the C_α atom is still chosen to be the network node, but the link between a pair of nodes is determined by the contacts between atoms of the two residues. A cutoff value 4.5 Å [7] (or 5.0 Å [8]) is used to judge the contacts between atoms. If atom contacts exist between two residues, there will be a link between the two nodes. As for the weighted amino acid network, the weight of the link can be based on the number of atom contacts between nodes [9]. Furthermore, when the diversity of amino acids is taken into account, these weights can be modified by the normalization factors [7]. Another way to add weight to the link is based on the probability of contact between amino acids of proteins [2].

For the amino acid network, especially the weighted amino acid network, related researches are just underway and many questions are needed to be explored, such as which model of the weighted amino acid network is more reasonable and how the network parameters change with different conformations of a protein molecule. This paper will do some pilot studies from these two points of view. A method is proposed to construct the weighted amino acid network. The link weight is based on the contact energies between residues. For the 197 proteins [7] with low homology, the unweighted and weighted amino acid networks are constructed and the statistic characteristics of the parameters of these networks are studied, including the average clustering coefficient (C), the average shortest path length (L), and the network parameter difference among different types of amino acids. Applying this weighted network, we have studied the changes of network parameters for the small protein chymotrypsin inhibitor 2 (CI2) on its high-temperature unfolding pathway [10,11].

II. THEORY AND METHOD

In the weighted amino acid network based on the contact energies between residues, the geometrical center of the side chain of an amino acid is chosen to represent the network node and the link between a pair of nodes is determined by the distance between them. If the distance between residues i and j, marked with r_{ij}, is less than the cutoff (r_c) value of 6.5 Å [12], there will exist a link between them. Thereby, the unweighted amino acid network is given and its adjacency matrix element a_{ij} can be expressed as follows:

$$a_{ij} = \begin{cases} 1 & i \neq j \quad \text{and } r_{ij} < r_c, \\ 0 & i = j \quad \text{or } r_{ij} \geqslant r_c. \end{cases} \tag{1}$$

Based on the contact energies between residues [12], the weighted network can be constructed and its adjacency matrix element a_{ij}^w can be expressed as

*Corresponding author. FAX: +86-10-67392837. Electronic address: cxwang@bjut.edu.cn

$$a_{ij}^{w} = \begin{cases} a_{ij}w_{ij}, & j \neq i \pm 1, \\ 2.55, & j = i \pm 1, \end{cases} \tag{2}$$

where w_{ij} is the link weight according to the magnitude of the contact energy between residues i and j suggested by Miyazawa and Jernigan [12], which is related to the types of the two amino acids. To avoid a negative weight value, the absolute value of the contact energy is used as the weight. For the covalent bond between residues i and $i\pm1$, the link weight is assumed as 2.55, which is the absolute value of the average collapse energy [12]. Thus, the distance matrix is constructed based on the weighted adjacency matrix and the definition of its element can be written as follows

$$d_{ij} = \begin{cases} 0, & i = j, \\ \infty, & a_{ij} = 0 \quad \text{and } i \neq j, \\ 2.55/(a_{ij}w_{ij}), & a_{ij} = 1. \end{cases} \tag{3}$$

The stronger the noncovalent interaction between two residues, the more the link between residues contributes to the stability of the whole protein. Thus, the link will get a greater weight and the distance between them will become shorter.

Additionally, a new network parameter, strength, is introduced into the weighted amino acid network. The strength of node i can be written as [9,13]

$$S_i = \sum_{j=1}^{N} a_{ij}^{w}, \tag{4}$$

where N is the number of network nodes and a_{ij}^{w} is an element of the weighted adjacency matrix. Furthermore, the strength of the whole network can be defined as $S=\frac{1}{2}\Sigma_1^N S_i$. The clustering coefficient of the weighted network can be calculated using the next expression [9,13]

$$C_i = \frac{1}{S_i(K_i - 1)} \sum_{j,h} a_{ij}a_{ih}a_{jh}\frac{w_{ij} + w_{ih}}{2}, \tag{5}$$

where S_i is the strength of the node i and K_i is its degree. The means of a_{ij} and w_{ij} are same as that of expression (2). The calculation of the shortest path length of the weighted amino acid network is based on the distance matrix of the network.

III. MATERIAL AND THE RESEARCH SYSTEM

The weighted amino acid networks were constructed based on a set of 197 proteins selected from the Protein Data Bank (PDB), including the four structure types α, β, $\alpha+\beta$, and $\alpha-\beta$. For all selected proteins, the resolution is better than 1.8 Å and the sequence identity is less than 20%. The sizes of proteins vary from 51 to 779 residues. For exploring the changes of network parameters with the changes of the protein conformations, the protein CI2 (PDB code 3CI2) was selected as a research object.

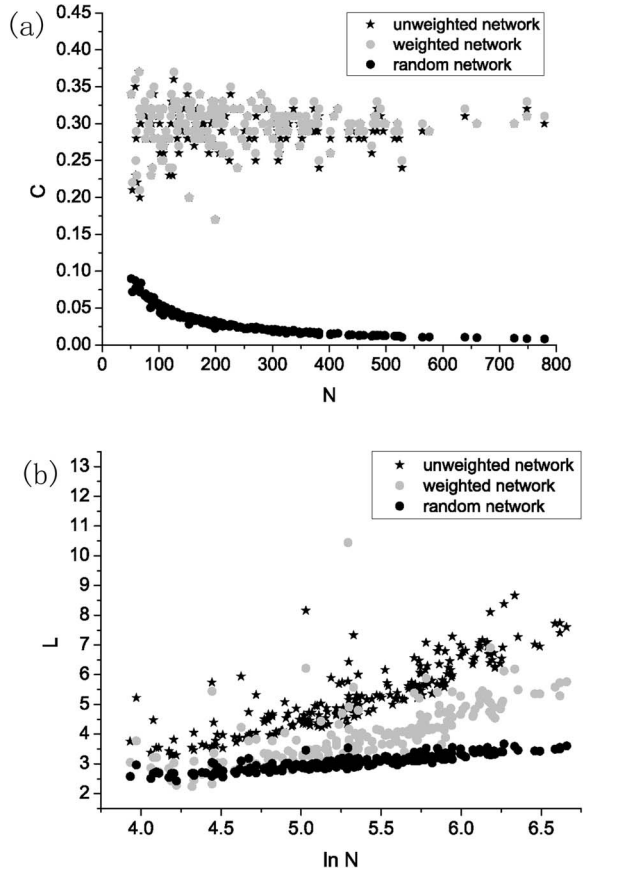

FIG. 1. For the weighted and unweighted amino acid networks of 197 proteins and the random networks with the same size: (a) the relation between the average clustering coefficient and the network size N and (b) the relation between the average shortest path length and the logarithm of N.

IV. RESULTS AND DISCUSSION

A. Small-world characteristic of the weighted amino acid network

When a complex network is compared with a random network with the same size (both the node numbers and the link numbers of these two networks are same, respectively), the complex network is of the "small-world" property [1], if the average clustering coefficient C and the average shortest path length L satisfy the next condition

$$C \geqslant C_r \quad \text{and} \quad L \geqslant L_r, \tag{6}$$

where C_r is the average clustering coefficient and L_r is the average shortest path length of the random network. They can be calculated with the expressions [1]

$$C_r \approx \langle K \rangle / N, \quad L_r \approx \ln N / \ln \langle K \rangle, \tag{7}$$

where N is the number of nodes of the random network and $\langle K \rangle$ is the average degree.

Figures 1(a) and 1(b) show the average clustering coefficients and the average shortest path lengths of the weighted and unweighted networks for 197 proteins. The two parameters of the random networks at the same size are also shown in Figs. 1(a) and 1(b). Obviously, all the weighted and un-

TABLE I. Comparison between the weighted network parameters of different type of amino acids in 197 proteins.

Amino acid	Strength-max[a]	Degree-max[b]	Strength-ave[a]	Degree-ave[b]
Cys	42.10	9	11.01	2.61
Met	48.81	9	10.80	2.34
Phe	51.06	9	13.36	2.44
Ile	57.16	10	14.80	2.81
Leu	61.19	10	15.01	2.64
Val	52.49	10	12.39	2.65
Trp	46.16	11	9.89	2.26
Tyr	36.13	11	8.63	2.18
Ala	34.94	10	6.49	2.03
Gly	27.18	10	4.66	1.84
Thr	27.46	9	4.95	1.91
Ser	25.65	9	3.93	1.74
Asn	21.08	8	3.48	1.61
Gln	24.14	9	3.58	1.52
Asp	20.36	9	3.16	1.56
Glu	23.38	9	2.66	1.29
His	28.01	9	5.65	1.91
Arg	21.31	9	3.79	1.54
Lys	18.71	9	2.12	1.16
Pro	31.4	10	4.28	1.73

[a]The maximum and the average value of strength residue obtained in the 197 proteins.
[b]The maximum and the average value of degree residue gained.

weighted amino acid networks present the "small-world" property. Comparing the result of this paper with that of other works in which the cutoff is selected to be 7.0 Å (or 8.5 Å) [2,4], we can see that the number of links and the average clustering coefficients all decrease when the cutoff is changed from 7.0 Å (or 8.5 Å) to 6.5 Å. But the "small-world" property of the networks is still remarkable. It shows that the construction method of the weighted amino acid network proposed in this paper is reasonable.

In the unweighted network, the clustering coefficient C scales the cohesiveness of the neighbors of a certain node only from the view of topology. But in the weighted network, it is a measure of the local cohesiveness not only including the topology factor, but also containing the information of interaction intensity of the local triplets. In the unweighted network, the shortest path length is only related to the link number of a pathway between two nodes and it is the least link number of all the pathways between two nodes. But in the weighted network, the link weights will affect the distance between nodes. Consequently, this will affect the calculation of the shortest path length. Therefore, the shortest path in the weighted network represents the path containing less links, as well as the path containing stronger contacts between nodes. These analyses show that the weighted amino acid network contains more information about the diversity of amino acids than that of the unweighted network.

As shown in Fig. 1(a), the average clustering coefficient of the weighted amino acid network has little changes compared with that of the unweighted network. This indicates that the weighted network model proposed in this paper has little effect on the clustering coefficient. When the atom contact number is taken as the link weight, the clustering coefficient will be obviously lower than that of the unweighted network [9]. The reason is that the weight of the covalent bond between residues i and $i\pm1$ is larger than that of the other noncovalent bond. Furthermore, there are two covalent bonds at the most in the case between a node and its neighbors. When the residue is located at the terminal of the peptide chain, there is only one covalent bond. Although the covalent bond has a large link weight, the probability to form local triplets is not high. Therefore, from Eq. (5), this will reduce the value of the clustering coefficient of the weighted network based on the atom contact number.

B. Comparison of the network parameter for different types of residues

Table I shows four network parameters of 20 kinds of amino acids in the weighted networks based on the residue contact energy of 197 proteins. All parameters include the maximum degree value (degree-max), the maximum strength value (strength-max), the average degree value (degree-ave), and the average strength value (strength-ave) in the 197 networks. For different types of amino acids, there is little difference in the degree-max. The degree-max value is just about 10. Because of the steric hindrance, a node cannot get more links. The maximum degree value may be different when the construction method for the network is changed

TABLE II. The comparison of both the link number and the strength of the weighted amino acid network of protein CI2 at two different states—i.e., the native (0 ns) and denatured (10 ns) states on the unfolding pathway.

	Total[a]		HH[b]		HP[c]		PP[d]	
Time (ns)	LN[e]	ST[f]	LN	ST	LN	ST	LN	ST
0	108	404.69	47	248.64	43	126.46	18	29.59
10	71	253.72	26	133.15	34	101.84	11	18.73

[a]The total amino acid network.
[b]The connect between hydrophobic amino acids.
[c]The connect between hydrophobic and hydrophilic amino acids.
[d]The connect between hydrophilic amino acids.
[e]The uncovalent connect number between amino acids.
[f]The connect strength value between amino acids.

[4,9]. In this small range of degree value, the average clustering coefficient C and the average shortest path length L still satisfy the condition proposed by Watts and Strogatz, as mentioned above [1]. The degree-ave is between 1.16 and 2.81. The degree-max is the same in both the weighted and unweighted networks, and so is the degree-ave. Therefore, they have little discrimination power for different kinds of residues. But in the weighted network, there is a large contrast in the strength-max for different kinds of residues, in which the minimum and maximum values are 18.71 and 61.19, respectively. Similarly, a large contrast in the strength-ave also exists. It is found that the strength value is related with its hydrophobic property. The 20 kinds of amino acids are divided into two classes [14], in which the hydrophobic residues (H) include Ile, Leu, Val, Phe, Met, Trp, Cys, Tyr, Pro, and Ala and the hydrophilic ones (P) are Gly, Lys, Thr, Ser, Gln, Asn, Glu, Asp, Arg, and His. Accordingly, the links in the amino acid network can be classified into three types: the link between the hydrophobic residues (HH), the link between the hydrophilic residues (PP), and the link between the hydrophobic residue and hydrophilic one (HP). In the weighted network, the proportions of these three kinds of links HH, HP, and PP to all links in the whole network are 40%, 40%, and 20%, respectively. These proportions are same in the unweighted network. Because the different kinds of links get different weights, the three types of links (HH, HP, and PP) contribute differently to the strength of the whole protein and the corresponding proportions are 55%, 35%, and 10%, respectively. These results show that the link numbers of the HH and HP links obtain the same proportion (40%). But the interactions between hydrophobic residues are stronger than that between hydrophobic and hydrophilic residues, so the weight of the HH link is larger than that of the HP link and the strength of the HH link contributes more to the strength of the whole protein as comparing to that of the HP link.

For a real protein molecule, different types of amino acids are of different power to get links and the contact intensities of these links differ in thousands ways. In the unweighted amino acid network, the diversity of amino acids is not exhibited, so that the discrimination power of the network parameters between different kinds of residues is limited. But in the weighted amino acid network, because more information of the diversity of different kinds of amino acids is embodied in the network model, the corresponding discrimination power is enhanced greatly.

C. Changes of the network parameters on the unfolding pathway of a protein molecule

The protein CI2 was used to perform the unfolding molecular dynamic (MD) simulation at 498 K for 10 ns with the MD program GROMACS 3.3 [15]. The force field parameters were taken from GROMOS96 43a1 and the SPC/E water model was used. The weighted amino acid networks were constructed with trajectory data at different time points with an interval of 10 ps. Then, the change of the network parameters associated with the conformational changes was analyzed on the unfolding pathway.

1. Parameters of the unitary network

After the unfolding MD simulation, the root-mean-square deviation (RMSD) of the main chain atoms reaches 1.26 nm. Therefore, the protein conformation can be regarded as the unfolding state. Table II lists the comparison of both the link number and the strength of the weighted amino acid network of protein CI2 at two different states—i.e., the native (0 ns) and denatured (10 ns) states on the unfolding pathway.

From Table II, it can be seen that the unfolding of the protein is mainly exhibited as the decrease of the HH links. The decrease of the HH link number occupies a main part (56%) in the total decrease of the whole link number from the simulation time 0 ns to 10 ns. When the link weight is considered, the decrease of the HH link contributes a larger proportion (76%) to the strength lost of the whole protein. The destruction of the HH link leads to the derogation of the protein's hydrophobic core.

The average clustering coefficients of the unweighted and weighted amino acid networks on the unfolding pathway are very similar, approximately between 0.2 and 0.3. The average clustering coefficient is less sensitive to the structural change. The reason is that when the protein unfolds, most of the secondary structures lose and the RMSD reaches 1.26 nm, but the protein still keeps a global random coil state

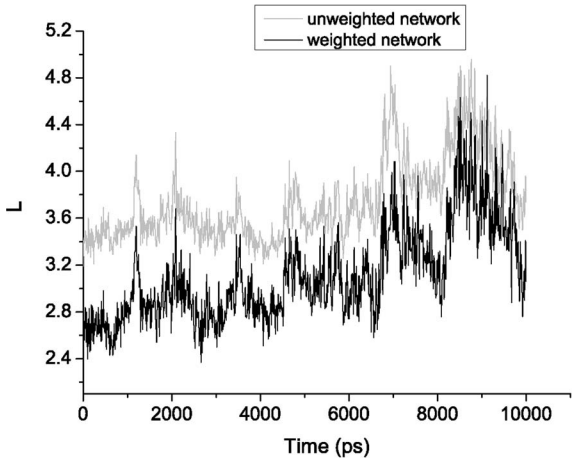

FIG. 2. The average shortest path length of the weighted and unweighted amino acid networks of protein CI2 on the unfolding pathway.

and there are still many nonbonding interactions in the protein. In addition, with the decrease of the node degree, the clustering coefficient will rise [9]. When the node degree is 2, the clustering coefficient will librate between 0 and 1 and the value will be 0 when the node degree becomes 1.

The average shortest path lengths of the weighted and unweighted networks on the protein unfolding pathway become longer with the structure looser, which can be seen from Fig. 2. Comparing the lengths of the two kinds of networks, we find that the length of the weighted network is more sensitive to the structural changes. This is mainly due to the destroying of the *HH* link, which has a distance less than 1. Therefore, while the hydrophobic core of protein derogates, the shortest path length will rise more obviously than that of the unweighted network.

2. Parameter of the folding core

Through the conformational cluster analysis [10] of the structures on the unfolding trajectory, it is found that the transition state of unfolding of CI2 was reached about 2.35 ns. The structure of the transition state is very similar to that proposed by Best and Vendruscolo [16]. The weighted amino acid network of the transition state was constructed, and the corresponding betweenness of the nodes is shown in Fig. 3.

For the shortest pathways, the more the ways passing through node i, the larger the betweenness value of node i will be and the status of the node i in the whole network becomes more significant. In Fig. 3, the betweenness values of the folding cores *A*16, *L*49, and *I*57 [17,18] locate at the local maximum [2]. Therefore, these residues are of remarkable functions during the folding process. The comparison of the betweenness value calculated from the weighted network with that of the unweighted network shows that the weighted network has an obvious higher power to recognize the folding cores.

At different time, we calculated the difference between the average strength values of the folding cores and that of

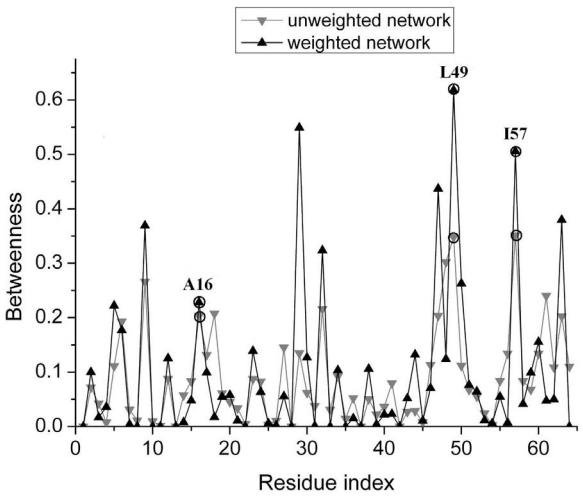

FIG. 3. The betweenness of residues was calculated from the weighted and unweighted amino acid networks for the transition state of protein CI2. The folding cores were marked by black circles.

all amino acids of the protein. The calculation formula is given by

$$s^t_{core} = (s^t_{16} + s^t_{49} + s^t_{57})/3 - \overline{s^t}, \tag{8}$$

where s^t_{16}, s^t_{49}, and s^t_{57} are the strength values of the folding cores at different time t and $\overline{s^t}$ is the average strength value of all amino acids of the protein at the same time. s^t_{core} is the difference between these two average values. The results are shown in Fig. 4. On the unfolding trajectory, with the destruction of the *HH* link, the hydrophobic core becomes derogated. But the strength of the folding core is higher than the strength average of all amino acids, especially before the transition state occurs. This indicates that the folding core plays a key role during the folding process. More and stronger links will be established between the folding cores and between the cores with other residues.

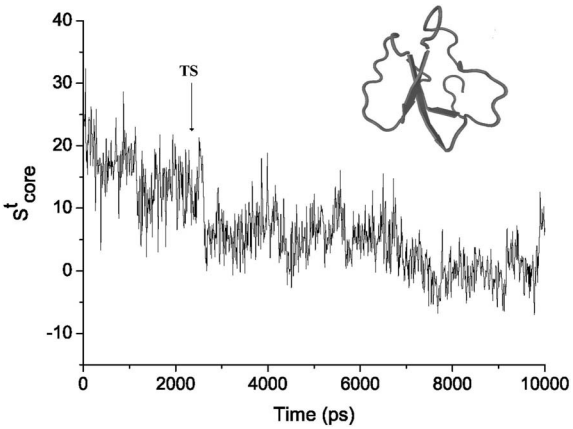

FIG. 4. The difference between the average strength values of the folding core and that of all the amino acids of protein CI2 on the unfolding pathway. The inset is the structure of the transition state model of CI2.

V. CONCLUSION

A method is proposed to construct the weighted amino acid network based on the contact energy. Through the analysis of the network parameters of the weighted amino acid networks for a set of 197 proteins, it is found that the weighted amino acid network is of an obvious "small-world" property. By an analysis of the influence of different weight modes on the network parameters, it is revealed that the weighted amino acid network contains more information about amino acids types than that of the unweighted network. The parameters of the weighted network conferred stronger discrimination power for different types of residues. Additionally, through the analysis of the changes of the weighted network parameters on the unfolding pathway of the protein CI2, it is observed that the unfolding of the protein is mainly exhibited as the derogation of the hydrophobic core. The shortest path length of the weighted network will rise increasingly with the protein unfolding, but the average clustering coefficient is less sensitive to the change of secondary structure. The betweenness values of the folding core at the transition state are local maximum, and it is easy for the betweenness of the weighted network to distinguish the folding core from other residue.

In summary, the weighted network based on the contact energy is reasonable and this work is helpful for studies of protein folding and the relationship between structure and function using complex network theory.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (Grant Nos. 10574009 and 30400087), Specialized Research Fund for the Doctoral Program of Higher Education (Grant No. 2004005013), and Beijing Excellent Person Sustentation Fund (Grant No. 20061D0501500192).

[1] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[2] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus, Phys. Rev. E **65**, 061910 (2002).
[3] N. V. Dokholyan, L. Li, F. Ding, *et al.*, Proc. Natl. Acad. Sci. U.S.A. **99**, 8637 (2002).
[4] A. R. Atilgan, P. Akan, and C. Baysal, Biophys. J. **86**, 85 (2004).
[5] G. Amitai, A. Shemesh, E. Sitbon, *et al.*, J. Mol. Biol. **344**, 1135 (2004).
[6] D. J. Jacobs, A. J. Rader, L. A. Kuhn, *et al.*, Proteins: Struct., Funct., Genet. **44**, 150 (2001).
[7] K. V. Brinda and S. Vishveshwara, Biophys. J. **89**, 4159 (2005).
[8] L. H. Greene and V. A. Higman, J. Mol. Biol. **334**, 781 (2003).
[9] M. Aftabuddin and S. Kundu, Physica A **369**, 895 (2006).
[10] A. J. Li and V. Daggett, J. Mol. Biol. **257**, 412 (1996).
[11] J. H. Wang, Z. Y. Zhang, H. Y. Liu, and Y. Shi, Phys. Rev. E **67**, 061903 (2003).
[12] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[13] A. Barrat, M. Barthelemy, R. Pastor-Satorras, *et al.*, Proc. Natl. Acad. Sci. U.S.A. **101**, 3747 (2004).
[14] S. J. Sun, R. Brem, H. S. Chan, *et al.*, Protein Eng. **8**, 1205 (1995).
[15] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. **7**, 306 (2001).
[16] R. B. Best and M. Vendruscolo, Structure (London) **14**, 97 (2006).
[17] L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, J. Mol. Biol. **254**, 260 (1995).
[18] L. Mirny and E. Shakhnovich, Annu. Rev. Biophys. Biomol. Struct. **30**, 361 (2001).