Grouping of residues based on their contact interactions

Jun Wang and Wei Wang*
National Laboratory of Solid State Microstructure and Department of Physics, Nanjing University, Nanjing 210093, China
(Received 5 December 2001; published 28 March 2002)

Based on the concept of energy landscape a grouping method of residues for reducing the sequence complexity in proteins is presented. For the Miyazawa and Jernigan matrix, rational groupings of 20 kinds of residues with minimal mismatches, under the consideration of local minima and statistics on correlation between the residues, are studied. A hierarchical tree of groupings relating to different numbers of groups N is obtained, and a plateau around $N=8-10$ is found, which may represent the basic degree of freedom of the sequence complexity in proteins.

DOI: 10.1103/PhysRevE.65.041911 PACS number(s): 87.10.+e

Using a small set of amino acid residues to reduce the sequence complexity in proteins, i.e., reducing the naturally occurring 20 kinds of residues into several kinds, has been studied [1–3]. Some patterns of residues were discovered in the reconstruction of secondary structures, such as binary patterns in α helices and helix bundles [2] (see review [4], and references therein). These imply that the hydrophobic cores, the native structures and the rapid folding behaviors of proteins can be realized by some simplified alphabets of residues. Theoretically, the simplest reduction, the so-called H-P model including H group with hydrophobic residues and P group with polar residues, has been extensively used. Yet, the relation between different forms or levels of these reductions (such as the five-letter palette [3], or different H-P groupings [5,6]) relating to the original sequences is not generally established. To find out the physical origin of these reductions is of importance for the protein representation.

Based on the Miyazawa and Jernigan (MJ) matrix of contact potentials of residues [7], reductions by dividing residues into different groups are made in our previous paper [8]. Several simplified schemes from minimized mismatches between reduced interaction matrix and the original MJ one are found. However, the physical picture of the mismatch is not well clarified, and the physical reasons for the grouping of residues need to be further studied. It is also important to make a comparison between the grouping results of different interaction matrices, and to study the generality of our simplification method. The goal of this paper is in these aspects. In this paper, a general picture and simplified formula of mismatch, based on the concept of energy landscape, are presented. Some rational groupings are obtained. Statistics on correlation between the residues reveal that some residues tend to aggregate together or are friends to live in the same group. A plateau of mismatch around group number $N=8-10$ for three different interaction matrices is found, implying that groupings with $N=8-10$ may provide a rational reduction for the complexity of protein sequences. This coincides with a fact that proteins generally include more than seven types of residues [4].

To divide 20 types of residues into a number of groups, the basic principle may be that the residues in a group should be similar in their physical aspects, mainly the interactions. After grouping, the residues in a group could be represented by one of the residues from the group, thus the complexity of protein sequences is reduced. When a residue is replaced by another, the energy landscape of a protein [9] should not change its main feature (the shape) or the folding features are basically the same. This is the case, especially when the system is near the bottom of the funnel where a protein has the most compact conformations. The energy difference between two nearby conformations $(c1)$ and $(c2)$ is defined as

$$\Delta E=\sum_n [e_n^{(c1)}(s_i,s_j)-e_n^{(c2)}(s_k,s_l)], \tag{1}$$

where $e_n^{(c1)}(s_i,s_j)$ (or $e_n^{(c2)}(s_k,s_l)$) is the contact energy of the nth contact between two residues s_i and s_j (or s_k and s_l) in $c1$ (or in $c2$), s_i defines the residue type of the ith element in the protein sequence, and the number of contacts in two conformations are assumed to be the same. To keep the main feature of the energy landscape means that ΔE should not change its sign, i.e.,

$$\mathrm{sgn}[\Delta E^{new}]=\mathrm{sgn}[\Delta E^{old}], \tag{2}$$

when a residue $s_g(g=i,\ j,\ k,$ or $l)$ is substituted by one of its "friends" s_g' in the same group. Here ΔE^{old} and ΔE^{new} are the energy differences of the original sequence and its substitute, and $\mathrm{sgn}[X]=1,\ 0,$ or -1 for $X>0,\ X=0,$ or $X<0$. Any discrepancy of Eq. (2) may change the energy landscape, and a quantity "mismatch" is used to characterize the discrepancy. Thus, the mismatch acts as a quantitative nonfitness of substitutions of residues.

In detail, 20 kinds of residues are partitioned into N groups as $G_1,\ldots,G_N$ with n_1 residues in group G_1, n_2 in G_2 and so on, where $n_1+n_2+\ldots+n_N=20$. For a given group number N, different values of n_i give different "sets" $(n_1,n_2,\ldots,n_N)$ of the partition, e.g., two sets (8,3,2,2,5) and (8,3,2,1,6) for $N=5$. [Actually, the "sets" relate to the partition of the number 20 into N groups, and the number of the sets L_N is 1, 10, 33, 64, 84, 90, 82, 70, 54, 42, 30, 22, 15, 11, 7, 5, 3, 2, 1, 1 for N from $N=1$ to 20, respectively.] The group assembly for a certain value of N could be represented as $\mathcal{G}_N=\{\{G_K^{(l)}(N),K=1,N\},l=1,L_N\}$ where $G_K^{(l)}(N)$ means the Kth group in the lth set among L_N. For a given set,

*Email address: wangwei@nju.edu.cn

different arrangements of residues in the groups represent different ''distributions'' of the residues, such as residue E in G_1 or in G_2. The mismatch will be minimized if the intragroup residues are friends for each group. [Residues that are not aggregated together finally in a group are not friends.] Due to the arbitrariness of contact index in ΔE and various possible distributions of residues, we define a strong requirement for a successful grouping: no change of the sign of each term in ΔE, i.e., $\lambda(s_i s_j s_k s_l) \equiv \mathrm{sgn}[e(s_i,s_j)-e(s_k,s_l)]$ equals to $\lambda(s_i' s_j s_k s_l) \equiv \mathrm{sgn}[e(s_i',s_j)-e(s_k,s_l)]$, when s_i is substituted by one of its friends s_i'. Here s_i, s_j, s_k, or s_l belongs to groups G_α, G_β, G_γ, or G_ν with $\alpha,\beta,\gamma,\nu \in 1,2,\ldots,N$, respectively. Generally, when a residue is substituted by another residue (friend or nonfriend) from the same group, one always has $\lambda(s_i' s_j s_k s_l)=1$ or 0 or -1. Then, all possible substitutions give a sum of related values of λ, i.e., $\Lambda_{\alpha\beta\gamma\nu}=\Sigma_{ijkl}\lambda(s_i s_j s_k s_l)$, which describes the total effects of substitutions of the residues from four groups G_α, G_β, G_γ, and G_ν. If $\lambda(s_i' s_j s_k s_l)$ is not the same as $\mathrm{sgn}[\Lambda]$, the substitution $s_i \rightarrow s_i'$ is not favorable or the grouping of s_i and s_i' in a group is a mismatch one. The average overall groups and residues gives out the total mismatch of this distribution

$$M_{ab}=\sum_{\alpha\beta\gamma\nu}\sum_{ijkl}\{1-\delta(\lambda(s_i s_j s_k s_l),\ \mathrm{sgn}[\Lambda_{\alpha\beta\gamma\nu}])\}/\sum_{\alpha\beta\gamma\nu}\sum_{ijkl}1, \tag{3}$$

where the summation runs overall possible combinations of α, β, γ, and ν and the index i runs overall residues in group G_α and so on, and the δ function is defined as $\delta(U,V)=1$ when $U=V$, 0 otherwise. For $\mathrm{sgn}[\Lambda]=0$, only the cases $\lambda(s_i s_j s_k s_l)>0$ are counted to avoid double counting.

Among all distributions of a fixed set $(n_1,n_2,\ldots,n_N)$, the best distribution (or the best arrangement of the residues) gives a minimal mismatch among all M_{ab}, i.e., M_{abmin}. Thus, for this set, one obtains M_{abmin} and the related distribution of residues in every group. To find out M_{abmin}, a Monte Carlo minimization procedure is used, where a less value of M_{ab} is obtained after every random exchange of two residues between two groups is accepted with a Metropolis probability $\min[1,\exp(-\Delta M_{ab}/T)]$. Here ΔM_{ab} is the change of the mismatches and $T=0.1$ is an artificial ''temperature.'' An enumeration overall possible distributions of residues can also be made for small N. For each N, all minimal mismatches M_{abmin} of L_N sets can then be obtained. In principle, for each N we could choose the lowest M_{abmin} and the related grouping as the final result among all sets L_N. However, this is difficult for those sets with MGWSE or groups with singlets. For example, as shown in Fig. 1 the mismatch of set (1,19) is the lowest one among all ten sets (also the set (1,1,1,1,16) for $N=5$, and so on, see Fig. 5). Obviously, this kind of mismatches does not relate to the best or rational groupings of the residues. Therefore, we must consider a local minimum (or a plateau) among all sets as the rational global minimum M_g (see Fig. 1). Such a ''locality''

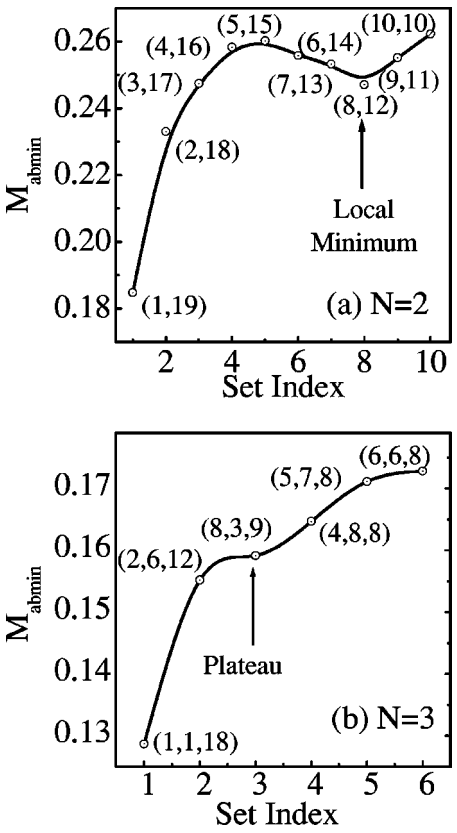

FIG. 1. M_{abmin} of different sets for $N=2$ (a) and $N=3$ (b). The set index represents the sets marked in the figure.

is motivated from the similarity between two groupings. Two groupings are regarded as a couple of neighbors when they can transform to each other just by exchanging two residues between two groups or by moving one residue from one group to another. With this, all local minima (or plateaus) are identified. Figure 1 shows such a local minimum (or a plateau) besides those with MGWSE. These local minima and plateaus represent better groupings, and reflect some intrinsic affinity between the residues. As a result, they are taken as the corresponding rational groupings with mismatches M_g.

The aggregation of some friendly residues into a group results from the correlation between these residues. Let us consider a two-residue correlation by counting the number of groups that include residues s_i and s_j, i.e.,

$$C(s_i,s_j)=\sum_{K=1}^{N}\sum_{l=1}^{L_N} I(s_i,G_K^{(l)}(N))I(s_j,G_K^{(l)}(N)), \tag{4}$$

where $I(s,G)=1$ when $s \in G$, or zero when $s \notin G$. Clearly, $C(s_i,s_j)$ is a quantitative scale of the affinity between two residues, or a probability of two residues being in a same group. It is worth noting that a weight average for groups with different mismatches is possible. For example, a probability with a Boltzmann-like distribution biased toward the small mismatches could be used. This might change the preference of the residues in some degree, but not largely. As we discuss the differences between different groups, the various definitions will not change the picture. Here we only discuss

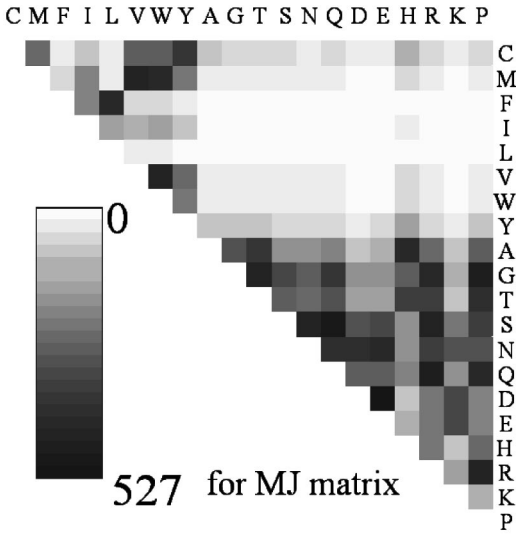

FIG. 2. A two-residue correlation $C(s_i,s_j)$ for the MJ matrix. Different shades of gray represent different values of the count $C(s_i,s_j)$ among all 84×5 groups for $N=5$.

the simple average with an equal weight. For all groups $\mathcal{G}_N$ with minimal mismatch M_{abmin}, it is found that the counts of some residue pairs are much large than those of other pairs (see Fig. 2). This means that some residues are friends and some are not, reflecting effective ''attraction'' between the residues in a group and ''repulsion'' between residues in different groups. Note that for the groupings with different N, we have similar patterns. The probability for finding a certain group G with specified residues among all minimal mismatch groups $\mathcal{G}_N$ can also be obtained by a count

$$C'(G)=\sum_{K=1}^{N}\sum_{l}^{L_N}\delta[G,G_K^{(l)}(N)], \tag{5}$$

where $\delta(G,G')$ is a δ function. As expected, different groups have different chances to appear (see Fig. 3). These differences result from not only the grouping affinity between residues but also the preference for the groups with a certain size. For comparison, the count $C'(G)$ is normalized by the total number of groups with the same size of group G in the group assemble $\mathcal{G}_N$. This normalized count is taken as a probability of the occurrence of group G, i.e.,

$$P(G)=C'(G)/\sum_{K=1}^{N}\sum_{l}^{L_N}\delta(\mathcal{S}(G),\mathcal{S}[G_K^{(l)}(N)]), \tag{6}$$

where $\mathcal{S}(G)$ defines the number of residues in group G, and $\delta(\mathcal{S}_1,\mathcal{S}_2)$ is also a δ function. From Fig. 3, it is found that some groups have large probabilities $P(G)$ and appear many times with large number of the counts $C'(G)$, implying that the residues in these groups have more chances to be in a group or that these groups have strong preference to appear in grouping. Thus, the grouping with these groups shows a better settlement of 20 kinds of residues than others. Note that some groups with large probabilities $P(G)$, but small counts $C'(G)$, are removed in our analysis because of lacking the statistical reliability. These correlation statistics are used in the grouping, especially in the selection of the best grouping among some competitive candidates.

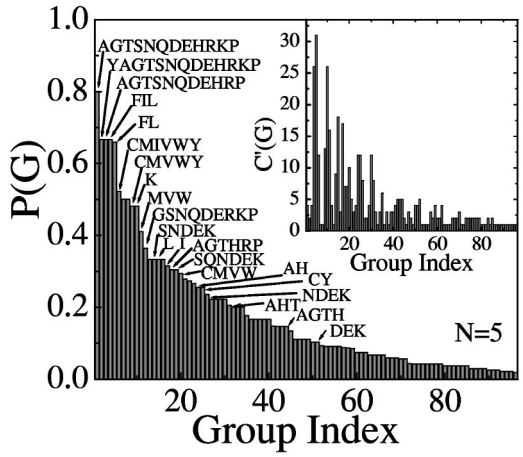

FIG. 3. Probabilities $P(G)$ and the counts $C'(G)$ for $N=5$ of the MJ matrix. The group index is arranged following the magnitude of the probability of the groups. Some groups are labeled.

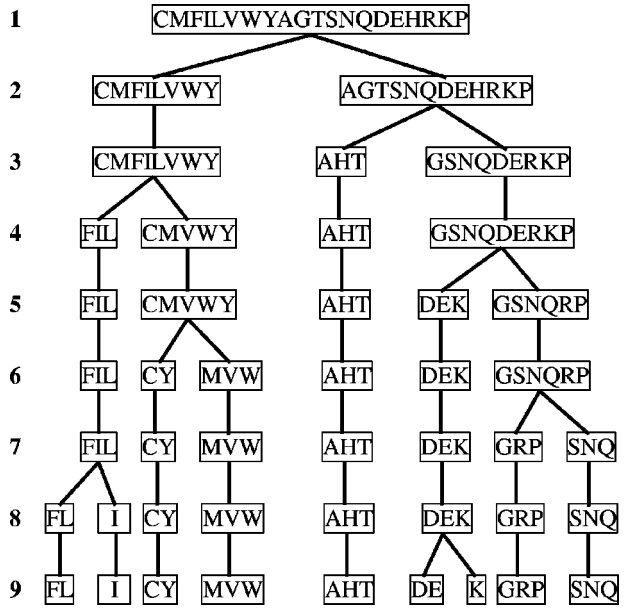

FIG. 4. The rational groupings of a hierarchically treelike structure for the MJ matrix for N up to 9.

With the method and requirements mentioned above, the reduction can be settled. For the MJ matrix, the groupings follow a hierarchically treelike structure (see Fig. 4). That is, 20 kinds of residues are firstly divided into two groups, i.e., an H group with residues (C,M,F,I,L,V,W,Y) and a P group with residues $(A,G,T,S,N,Q,D,E,H,R,K,P)$. Then these two groups are alternatively divided into two or more groups relating to different N, reflecting the detailed differences between the interactions of the H and P groups. In the case of $N=3$, to divide the P group (on the base of $N=2$) is obviously more rational than to divide the H group, suggesting a priority for dividing the P group first. Differently, for $N=4$, we should divide the H group first, and for $N=5$

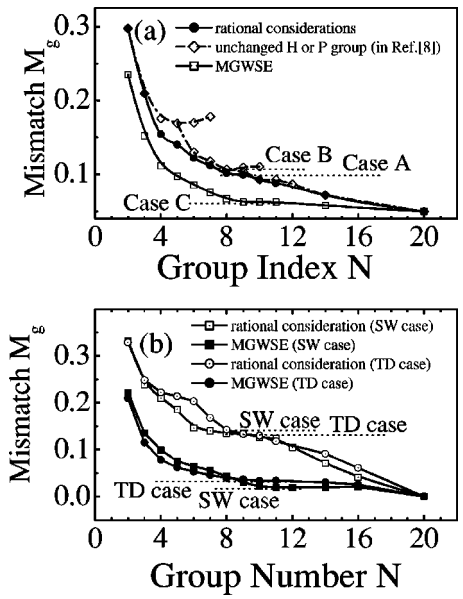

FIG. 5. M_g vs N: (a) for the MJ matrix; (b) for contact potentials in Ref. [6] (TD case) and in Ref. [10] (SW case). The plateaus are shown for different cases.

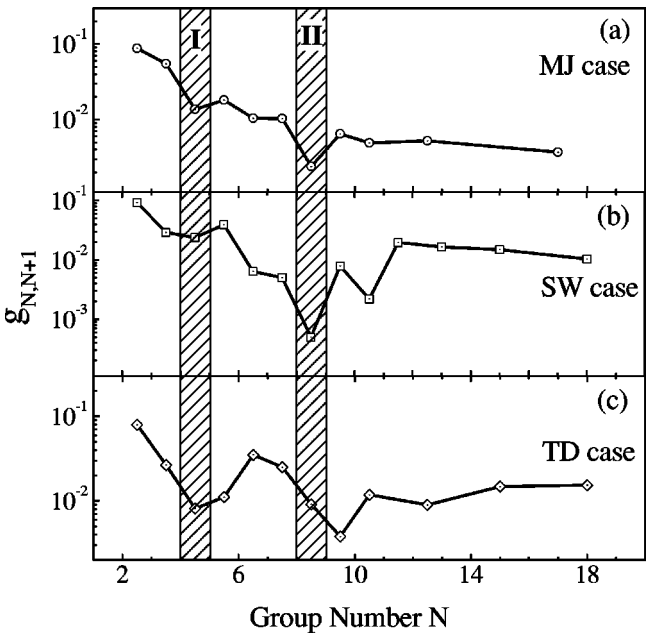

FIG. 6. The gradient $g_{N,N+1}$ vs group number N for (a) MJ case, (b) SW case and (c) TD case related to the rational considerations in Fig. 5, respectively. The grey regions highlight the common minima of $g_{N,N+1}$.

divide the P group again. For example, in the case of $N=5$, the H group is divided into (F,I,L) and (C,M,V,W,Y), and the P group is divided into (A,H,T), (D,E,K), and (G,S,N,Q,R,P). Similar results are obtained for N up to 9 with a sequential order of hydrophobicity without any overlap between the hydrophobic branch and the hydrophilic one following the H/P dividing. This relates to a clear picture of the rational groupings. The difference between the present study and previous one in Ref. [8] is that there are alternant dividings of the H and P groups in the new groupings, which gives out a little decreasing in the mismatches, and also slightly different representative residues. The former results under some restrictions, such as to fix the H group (with eight residues) unchanged, may relate to somewhat rough dividing and the grouping space for searching the local minimal is a little bit limited.

Figure 5 shows a decrease in the mismatch as the group number N increases, implying, in general, the more groups the better. However, there is a plateau near $N=8$ (case A), which characterizes the saturation of the grouping. This means that more groups will not further decrease the mismatch or more groups might not greatly enhance the efficiency of the complexity reduction. Thus, the number $N=8$ may indicate the minimal number of residue types to reconstruct the natural proteins, or a basic degree of freedom of the complexity for protein representation. This, in a sense, relates well to the argument in Ref. [4]. Noted that the former plateau at $N=5$ ceases due to the canceling of the grouping restriction. Interestingly, in Fig. 5, we also plot all the lowest mismatches relating to the groupings with MGWSE, which generally are not the local minima. An example is the grouping with groups (1,1,1,1,16), which has the lowest mismatch among all sets of $N=5$. However, it is noted that even including all these trivial groups, the curve still shows a plateau around $N=9$ with eight groups with single residue of C,M,F,I,L,V,W,Y and one group with the rest twelve residues as well [see case C in Fig. 5(a)]. Clearly, this plateau relates again to the saturation of the H or P grouping or the detailed differences between the interactions of the residues, and also gives out a support on the discussion for the $N=8$ plateau above. In addition, similar results for two other interaction matrices [6,10] are also obtained [see Fig. 5(b)].

To see the plateaus more clearly, we derive the gradient of mismatch M_g from N groups to $N+1$ groups for above rational cases. Here, the gradient $g_{N,N+1}$ is defined as $g_{N,N+1}=|M_g(N+1)-M_g(N)|$. It is obvious that there are minima of gradient $g_{N,N+1}$ vs N, implying a small variation of mismatch as the group number N increases. These minima may correspond to plateaus or shoulders of the curve of the mismatch vs group number. For our results, the values of gradient $g_{N,N+1}$ of different datasets of contact potentials basically are minimal around $N=5$ (gray region I in Fig. 6), which correspond shoulders around $N=5$, and also are minimal around $N=8$ (gray region II in Fig. 6), which relate to plateaus about $N=8$ (see Fig. 5). That is to say, the contact potentials of different sources all favor the eight-type grouping. Such an independence of detailed forms of interactions suggests that the grouping with eight-type residues might be a common feature of residues in the protein systems.

It is worth noting that for each N the representative residues have been found for the MJ matrix [11], e.g., (I,A,D) for $N=3$, (I,A,C,D) for $N=4$ and (I,A,G,E,C) for $N=5$. These residues are selected based on the rational groupings by minimizing the mismatch among all other choices. The foldability of the reduced sequences and the effectiveness of the reduced alphabet have also been studied. All these details will be reported elsewhere.

Finally, as a remark, we note that we use the pair-wise contact potentials as the starting point of our approach. Actually, the effective interactions between residues in folding processes are of many body due to their complicated interplay with solvent. The pair-wise interactions between the residues are the average ones under some approximations, and are believed possessing the basic ingredients of the driving forces in the folding in general [5–8]. Recently, it is pointed out that the many-body effect may have their important roles for the recognition of the correct folds and the thermodynamics and kinetics of the folding processes [12–19]. To consider the many-body effect would be appealing for the grouping problem. Generally, the preferences between some certain residues may be enhanced, while some fragile connection between residues might be broken due to the competition of the many-body perturbation. However, the basic pattern of residue grouping will be maintained though the relation between some residues may become vague and complex. The detailed schemes deserve further investigation.

In conclusion, we have shown a grouping method of residues based on a requirement that the energy landscape should be basically kept in reduction. A quantity, the mismatch, is taken as the measurement of the reduction. Our results imply that the residues do have some similarities in their interaction properties and can be put together into groups. By choosing a residue for each group, the complexity of proteins can be reduced or the proteins can be represented with reduced compositions. Especially, a basic degree of freedom of the complexity with 8–10 types of residues is found.

This work was supported by the Foundation of NNSF (Nos. 10074030, 90103031, and 10021001) and the Nonlinear Project (973) of the NSM. J.W. thanks the Ke-Li Research Foundation. We thank C. Tang, C. H. Lee, and H. S. Chan for comments and suggestions.

[1] K. A. Dill, Biochemistry **29**, 7133 (1990); H. S. Chan and K. A. Dill, Macromolecules **22**, 4559 (1989); H. Li *et al.*, Science **273**, 666 (1996); E. D. Nelson and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **95**, 10 682 (1998); G. Tiana, R. A. Broglia, and E. I. Shakhnovich, Proteins: Struct., Funct., Genet. **39**, 244 (2000); H. S. Chan and K. A. Dill, *ibid.* **30**, 2 (1998); R. A. Goldstein *et al.*, Proc. Natl. Acad. Sci. U.S.A. **89**, 9029 (1992); P. G. Wolynes, Nat. Struct. Biol. **4**, 871 (1997); L. R. Murphy *et al.*, Protein Eng. **13**, 149 (2000).

[2] L. Regan and W. F. Degrado, Science **241**, 976 (1988); S. Kamteker *et al.*, Science **262**, 1680 (1993); A. R. Davidson *et al.*, Nat. Struct. Biol. **2**, 856 (1995).

[3] D. S. Riddle *et al.*, Nat. Struct. Biol. **4**, 805 (1997).

[4] K. W. Plaxco *et al.*, Curr. Opin. Struct. Biol. **8**, 80 (1998).

[5] H. Li *et al.*, Phys. Rev. Lett. **79**, 765 (1997).

[6] P. D. Thomas and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **93**, 11 628 (1996).

[7] S. Miyazawa and R. J. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[8] J. Wang and W. Wang, Nat. Struct. Biol. **6**, 1033 (1999).

[9] P. G. Wolynes *et al.*, Science **267**, 1619 (1995).

[10] B. Shoemaker and P. G. Wolynes, J. Mol. Biol. **287**, 657 (1999).

[11] It is found that the percentage of overlap of the representative residues is larger than 75% for $N \geqslant 3$ for three interaction matrices used in this paper.

[12] K. A. Dill, J. Biol. Chem. **272**, 701 (1997).

[13] M. Vendruscolo, R. Najmanovich, and E. Domany, Proteins: Struct., Funct., Genet. **38**, 134 (2000).

[14] H. S. Chan, Proteins: Struct., Funct., Genet. **40**, 543 (2000).

[15] H. Kaya and H. S. Chan, Proteins: Struct., Funct., Genet. **40**, 637 (2000).

[16] H. Kaya and H. S. Chan, Phys. Rev. Lett. **85**, 4823 (2000).

[17] S. Takada, Z. Luthey-Schulten, and P. G. Wolynes, J. Chem. Phys. **110**, 11 616 (2000).

[18] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **90**, 6369 (1993).

[19] K. Fan, J. Wang, and W. Wang, Phys. Rev. E **64**, 041 907 (2001).