A computational approach to simplifying the protein folding alphabet

Jun Wang and Wei Wang

National Laboratory of Solid-State Microstructure and Department of Physics, Nanjing University, Nanjing 210093, People's Republic of China.

What is the minimal number of residue types required to form a structured protein? This question is important for understanding protein modeling and design. Recently, an experimental finding by Baker and coworkers suggested a five-residue solution to this problem. We were motivated by their results and by the arguments of Wolynes to study reductions of protein representation based on the concept of mismatch between a reduced interaction matrix and the Miyazawa and Jernigan (MJ) matrix. We find several possible simplified schemes from the relationship of minimized mismatch *versus* the number of residue types (N = ~2–20). As a specific case, an optimal reduction with five types of residues has the same form as the simplified palette of Baker and coworkers. Statistical and kinetic features of a number of sequences are tested. Comparison of results from sequences with 20 residue types and their reduced representations indicates that the reduction by mismatch minimization is successful. For example, sequences with five types of residues have good folding ability and kinetic accessibility in model studies.

The heterogeneity of 20 types of amino acid residues and the diversity of different protein structures introduce complexity into protein folding[1–7]. Much effort has been made by considering 'minimalist' models with a small number of residue types to simplify this complexity [4–23]. To allow interpretation of some features of protein folding, the hypothetical interactions in these minimalist models are much simpler than the real ones[2–8,12–19,24–27]. Physically, this means that 20 types of residues (the natural set of residues) are grouped, approximately according to similarities in their physical and chemical properties. Each group can be regarded as a type of monomer, and can be represented by a letter. Consequently, the matrix of interaction between residues can be reduced to one with a small dimension (Fig. 1*a*). The simplest reduction scheme is the HP model (where H stands for hydrophobic and P for polar), which consists of only two 'letters' and considers hydrophobicity as the only driving force.

Experimentally, some specific patterns of amino acid composition have been discovered in the reconstruction of secondary structures, such as binary patterns in α-helixes and helix bundles[20–22]. Recently, Baker and colleagues successfully built the well-ordered SH3 domain with five types of amino acids[23]. These results indicate that it is possible to depict structural characteristics of proteins with a reduced set of amino acids. Furthermore, combinatorial studies by Baker *et al.*[23] suggest that five or more types of residues seem necessary for a foldable protein. Indeed, even in previous experimental studies of HP patterns[20–22], more than two types of residues were required for successful building of protein structure. Therefore, more detailed patterns than the simple two-letter HP forms should be explored for better description of proteins.

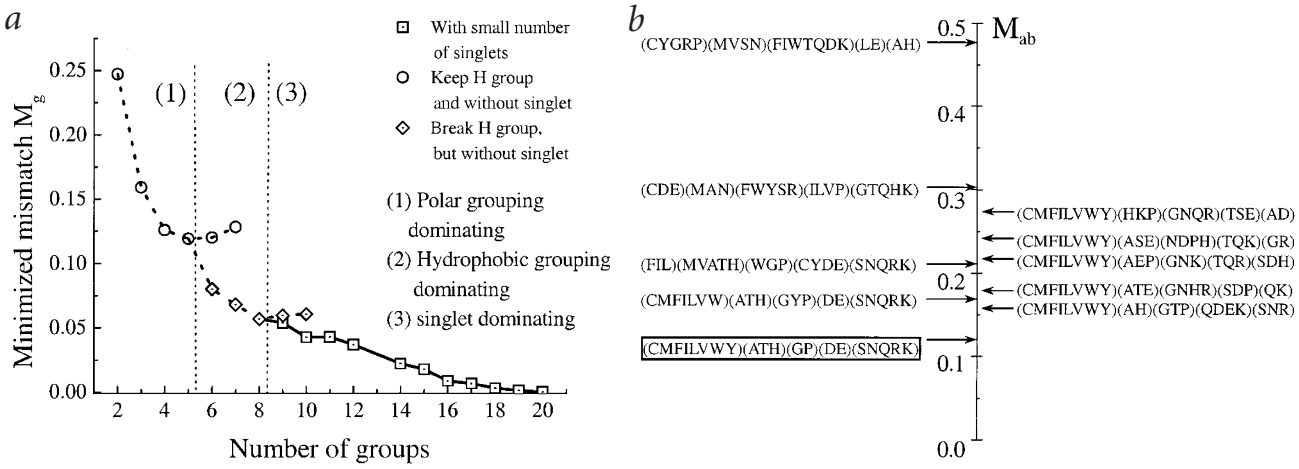
Different simplified schemes, such as the HP and multi-component models, in which each monomer has multiple choices of type, have been considered in theoretical studies[2–8,12–19,24–27]. As argued by Wolynes[6], the theoretical energy landscape for the two-letter cases has many local traps and does not have a steep funnel, features that could slow the folding process. The fact that many proteins have fast folding characteristics implies that some complexity of natural proteins is not embodied in the HP model[4–6,26]. To encode a wide range of structures and properties of proteins, multi-component schemes are recommended[7,8,19,27,28]. However, a number of questions remain unanswered[6]. First, how should one reduce the 20 types of residues in order to simplify the description? Second, how many types are necessary in the minimalist models? Third, what is the physical origin of the reduction in these minimalist models[6]?

Reduced representation by grouping of residues

Let us introduce a general method for reducing the representation of proteins by the grouping of residues. A reduction algorithm, which connects different representations of a protein generally rests on the idea that residues (or monomers) can be distributed into several groups, each of which has different physical and chemical properties, and thus, different interactions. For a successful reduction, the interactions between monomers of the two groups, say group A and B, should have similar characteristics.

		A			B			C			D			E		
		A(1)	…	A(n)	B(1)	…	B(n)	C(1)	…	C(n)	D(1)	…	D(n)	E(1)	…	E(n)
A	A(1) … A(n)	AA			AB			AC			AD			AE		
B	B(1) … B(n)	BA			BB			BC			BD			BE		
C	C(1) … C(n)	CA			CB			CC			CD			CE		
D	D(1) … D(n)	DA			DB			DC			DD			DE		
E	E(1) … E(n)	EA			EB			EC			ED			EE		

Fig. 1 A general view of reduction. Residues are put into N groups, such as A, B, C, D, E, and they are marked as A(i) ($1 \le i \le n_A$) and so forth. The matrix of interactions between residues (with 20×20 elements) is reduced to one with $N \times N$ blocks, which reflects the interaction between groups. Because of its symmetry, only the gray area is considered.



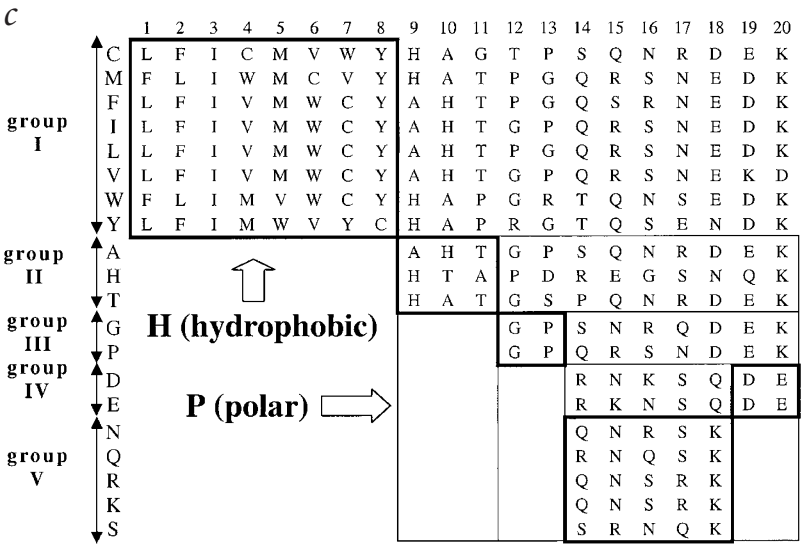


Fig. 2 The results of reduction. ***a***, Minimized mismatches M_g *versus* the number of groups *N* (or the number of residue types). For each *N*, the mismatch M_g is minimized over all possible groupings (see Methods). The circles show the minimized mismatches M_g by keeping hydrophobic groups unchanged and without singlets, and the diamonds show the minimized mismatches M_g by breaking hydrophobic groups, but without singlets. The squares show the minimized mismatches M_g taking into account of all the differences between the residues but with a small number of singlets during the minimization. These three regions are determined by the corresponding plateaus, which reflect optimal conditions under certain requirements (as stated in the text). ***b***, Mismatches M_{ab} of different reductions for group number *N* = 5. The reduction scheme with a black border at the bottom has the least mismatch. Variation of this reduction, even a little, may introduce an increase in mismatches. In particular, the reduction schemes on the left of the axis show an increased number of mismatches due to splitting of the hydrophobic group. The part on the right of the axis gives an illustration of gaining mismatch from mixing of polar groups. ***c***, Matrix reduction with ordering rule. For any residue (in the column labeled with CMFILVWY...), such as residue C, its interaction with any other residue is arranged in ascending order; that is, the interactions become less negative from left to right. The ordered interactions are listed in a line (in the horizontal direction) following each residue. Using this kind of arrangement, we determined that some residues, such as the eight residues CMFILVWY, form a group. That is, the interaction elements between these residues basically form a block due to the existence of jumps (or sharp changes) of the values of the interaction elements in the ascending order lines. These residues are taken as group I. By the same observation, we find group II with residues A, H, T, group III with G and P, group IV with D and E, and group V with N, Q, R, K and S, respectively. It is noted that any variation of the order of letters in each group in the label column will not affect the existence of blocks. To obtain more clear boundaries for grouping, the residues in the ascertained groups are excluded from the analysis for next ones. The trivial cases, in which all residues are in a group or only one is in a group, are skipped. The hydrophobic (group I) and polar (groups II, III, IV, and V) divisions can be seen clearly, which coincides with the MC results shown in Fig. 2*b*. As intermediate results, two-group schemes with group A (consisting of group I) and group B (consisting of group II,III,IV and V), and three-group schemes with group A (consisting of group I), group B (consisting of group II) and group C (consisting of group III, IV and V) are obtained.

This has two meanings. On one hand, the interactions between any pairs of monomers, say A(i)B(j) and A(i')B(j'), are similar in their strength. On the other hand, the interaction between A(i) and B(j) is generally different from that between C(k) and D(l) (in groups C and D). For instance, in the HP reduction with two groups H and P, the interaction between hydrophobic monomers H(i) and H(j) is always stronger (more negative in strength) than that between hydrophilic ones P(k) and P(l) for any possible combination of i, j and k, l; the indices represent the sequence numbers in the corresponding groups (Fig. 1). Meanwhile, the strength of interaction between H(i) and H(j), and that of interaction between H(i') and H(j'), for any possible combination of i, j and i', j', do not appear to be significantly different. Therefore, after suitable arrangement of residues, the interaction matrix can be represented by a number of blocks (AA, AB, and so forth; Fig. 1) with elements (namely, contact interactions; equation 1 in Methods) in each block having similar properties in strength.

The goal of reduction, then, is to determine elements in each block from, for example, the MJ matrix (see below). In practice, we define a mismatch as the discrepancy of properties in strength between elements and blocks. For example, assuming the relative strength of block AB is weaker (or more positive; see Methods) than that of block CD, if the element A(i)B(j) in block AB is stronger (or more negative) than the element C(k)D(l) in block CD, the pair (A(i)B(j), C(k)D(l)) makes a mismatched pair (Fig. 1). A good reduction requires a minimized number of mismatches. Obviously, a trivial reduction of 20 groups with only one residue type in each group (which has null mismatch) is an example, but it shows no simplification. It is worthwhile to note that the simplest reduction, the HP model, actually includes two groups, the H group with the hydrophobic residues and the

P group with the polar residues[4–6]. However, the residue set is further simplified because two monomers (or letters) within a group have the same effective interactions. This reduced set shows limited heterogeneous features of proteins because of the very small number of different contact interaction energies. Thus, the complexity of proteins is simplified by such grouping of residues.

In this work, we present a reduction method based on an analysis of the statistical contact potentials of the MJ matrix (upper triangular part of Table 3 in ref. 29). By minimizing the mismatches between a reduced matrix and the MJ matrix *versus* the number of residue types, we find three ranges: (i) a polar dominated grouping (PDG) (ii) a hydrophobic dominated grouping (HDG) and (iii) a singlet dominated grouping (SD). In the range PDG, the differences between polar residues are important when the main force, the hydrophobic residues, are kept in a group. The mismatch minimization algorithm results in a reduction to five types of residues that gives a good description of proteins. In the range HDG, the differences between hydrophobic residues become important, and minimization of mismatches reaches another local minimum at around seven residue types. When all details of interaction between residues are considered, namely in the range SD, a 20-type representation gives a zero mismatch. In addition, for the five-type reduction, the statistical and kinetic folding characteristics of some reduced sequences (sequences with 20-type residues are represented by 5-type residues under an optimal reduction scheme) are found to be basically the same as that of the optimized sequences with 20 types of residues. These results suggest that a certain degree of heterogeneity is necessary but, at the same time, that some simplifications are also possible for the adequate description of natural proteins.

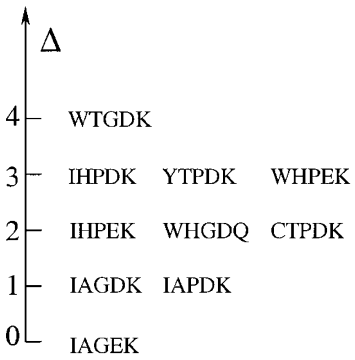


Fig. 3 Representative residues. Deviation of sets of representative residues for various reductions with $N = 5$ from the best set of **I,A,G,E,K** are shown. The **I,A,G,E,K** set has the lowest deviation from corresponding **P**, and its deviation is set as zero.

Reduction based on the MJ matrix

Based on the MJ matrix, some good reductions for different group numbers N are obtained by minimization of the mismatch (Fig. 2*a*). For a good reduction with a given number of groups, such as $N = 5$, any addition and subtraction of residues in a group or exchanges of residues between groups will make considerable increases in mismatches (Fig. 2*b*). Interestingly, except for the trivial minimum around $N = 20$, there are two other local plateaus: one near $N = 5$, and the other around $N = 7$ (Fig. 2*a*). Corresponding to these plateaus, there are three regions, and in each region, only one kind of factor dominates the dividing of groups, or grouping. In region (1) of Fig. 2*a*, the behavior of the minimized mismatch M_g is a result of separating polar residues into groups while the hydrophobic group is basically kept unchanged. In region (2) the hydrophobic residues are further divided into several groups, with the existing polar groups unchanged. In contrast, in region (3), the existence of groups containing single residues dominates the mismatch M_g, which relates to the effects of the detailed interactions between residues when the group number is large. The dividing of groups in each region makes a steep decrease of the mismatch M_g at first and then reaches a plateau. With these features, each local plateau reflects an ample consideration of certain kinds of interactions. If some groups are set unchanged manually following the increase of group number, a slight increase of the mismatch M_g may appear (see dashed lines in Fig. 2*a*).

These results suggest that by considering hydrophobicity in this way, the $N = 5$ group reduction may work well. If more detailed differences among hydrophobic interactions are included, more groups, say $N = 7$, may be necessary. For even more detailed consideration of both hydrophobic and polar interactions between residues, more groups (with $N > 8$) may be required for modeling. As a result, the grouping with $N = 20$ takes into account all the details of contact interactions between the natural residues. Actually, the local plateaus mentioned above result from the compromise between the homogeneous simplicity and the heterogeneous details, and reflects a saturation of the grouping under a certain kind of interaction. Furthermore, the differences between polar residues seem to be more important than that between hydrophobic ones within a grouping. This shows up in the first stage after the HP reduction.

The residues with different functions fall naturally into different groups following the reduction procedures. For example, Asp and Asn, Glu and Gln, and Ser and Thr have different chemical properties and are in different groups in certain reductions. If these residues were mixed into a single group, many mismatches would occur. Such an obvious 'exclusivity' during reduction suggests that the mismatch consideration is a reasonable approach to this problem.

These results can also be partly obtained by a simple procedure. As mentioned above, the reduction to a block form of the MJ matrix requires the distinction of interaction elements between different blocks. Thus, the ascending order of the values of elements in each line of the MJ matrix may also relate to the intrinsic block feature. With this in mind, we can make a matrix decomposition (Fig. 2*c*) that yields five groups, while skipping the trivial groupings (such as all residues in one group, or only one residue in a group). More detailed reductions are not obtainable from this procedure. Interestingly, the results coincide well with the minimized mismatch considered above. They are consistent with not only five-group reduction, but also with two- or three-group reductions. Obviously, the HP reduction is an intermediate result of our five-group reduction by both methods in that the first group is represented by H, and the rest groups together by P (Fig. 2*c*). Such an HP division is the same as the results of Li *et al.*[14] and others[30].

Here, we take the MJ matrix as a reasonably accurate description of interactions between real residues. Although there are some intrinsic disadvantages in forms[24] and statistical methods[31], the MJ contact potentials are widely used in various applications and are believed to be qualitatively reliable. Fortunately, our reduction method depends only on the relation of matrix elements in strength (or differences of the contact potentials), which is generally maintained in the processes of obtaining statistical potentials[31]. Therefore, our

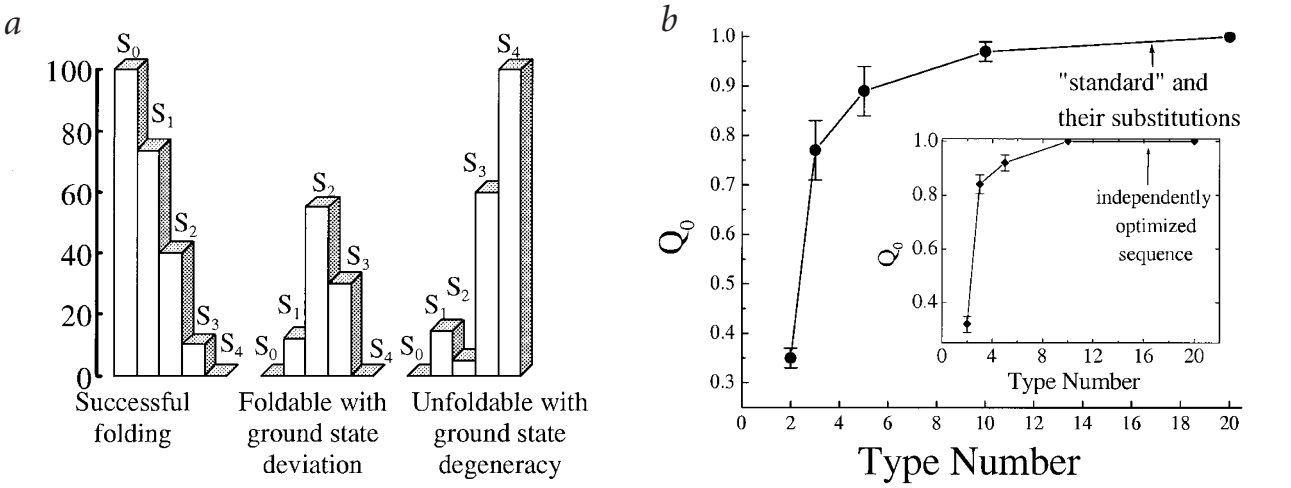


Fig. 4 Results of folding test for the reduced sequences. ***a***, The statistics of foldable and unfoldable ground states for different sequence schemes.

S_0: Standard sequences with 20 residues: CMFILVWYATHGPDESNQRK
S_1: scheme 1 (five letters): (CMF**I**LVWY)(**A**TH)(**G**P)(D**E**)(SNQR**K**)
S_2: scheme 2 (variant five letters): (CMF**I**)(**L**VWY)(**A**TGS)(NQD**E**)(HPR**K**)
S_3: scheme 3 (three letters): (CMF**I**LVWY)(**A**THGPR)(DESNQ**K**)
S_4: scheme 4 (two letters): (CMF**I**LVWY)(**A**THGPDESNQRK).

The parentheses denote a group and the bold letter is the representative. 'Standard' contains a series of 27-residue sequences with 20 residue types uniformly distributed in the sequences. These sequences are well optimized for a certain compact structure taken as the native one at a low optimization temperature T = 0.8 by annealing in the sequence space[26] using the MJ matrix[29] as the energy matrix. The residues in the standard sequences are substituted by their representative letters according to different schemes, S_1–S_4. The ground states of these new sequences are tested by enumerating all possible compact structures on a cubic lattice[15,16]. The histograms are the statistical results of 200 optimized sequences. 'Successful folding' means that the substituted sequences share the same ground state or native structure of the standard sequences. 'Foldable with ground state deviation' denotes that these sequences are foldable, but their ground states are different from that of standard sequences. 'Unfoldable with ground state degeneracy' represents that they are unfoldable due to the degeneracy of ground states. ***b***, The average contact overlap $Q_0 = \langle Q_S \rangle$ *versus* the number of the residue types *N*. The contact overlap Q_S between the native structure and the lowest energy structure S is defined as the number of common contacts in both structures normalized by the maximal number of the native contacts — that is, 28 native contacts for the $3 \times 3 \times 3$ cubic lattice. In our simulation, 20 out of the 200 optimized standard sequences and their substitutes with the representative letters as in (*a*) are tested for their ability to fold by the MC procedures[36] at T = 1.6. For each N, $Q_0(N)$ is averaged over 20 different sequences (with 20 different MC trails for each sequence). Each run takes 10^9 MC steps. The native structure and the interactions are as in (*a*). The solid circles represent the results for the standard sequences and their direct substitutions, and the inset is for the independently optimized sequences using only the reduced alphabets. They show a similar monotonic increase in the values of $Q_0(N)$ as the number of residue types N increases, implying that there is a kinetic advantage in having more residue types.

results based on the analysis of the MJ matrix should be relevant to the physical origin of reduction, at least at a qualitative level. Interestingly, our five-letter palette is well in agreement with the results of Baker and colleagues[23] (see below). Hence, our reductions may capture the basic physics of protein interactions.

Representative letters for the reduction

As a further step in the reduction process, some residues should be identified as representatives of the groups. That is, each group will be represented by a type of residue. For example, the five-group reduction is represented by five types of residues (five letters). We select some residues from resultant groups by the following procedure. The interaction between the selected residues should be consistent with that between groups. Because of the similarity of residues in a group, the choice for representative residues seems arbitrary. However, we can determine them by a well-defined procedure (Fig. 3; see Methods). As a result, for the five groups with minimal mismatches, their representatives are taken as I, A, G, E and K, which is the same set that resulted from the experiments by Baker and coworkers[23]. With these residues, the MJ interaction matrix (210 independent elements) is reduced into a smaller matrix (15 independent elements). Similar to the matrix analysis procedure of Li *et al.*[14], our reduced matrix retains the general features of the MJ matrix.

Folding test for the reduced sequences

As a test, in this section we study the statistics of the folding properties of sequences realized by different reduction schemes. A number of 20-letter sequences, which are optimized using the method proposed by Shakhnovich *et al.*[28,32] and have good folding behavior, are collected as 'standard sequences'. The folding properties of their corresponding substitutes following different reduction schemes are tested. It is found (for example, by following scheme 1 in Fig. 4*a*) that a large number of five-letter substitutes have the same ground state as that of their original standard sequences, which implies that scheme 1 is a successful reduction. By comparison, other schemes lead to changes in the ground state conformations or degeneracy of ground states. Thus, the five-letter reduction of scheme 1 is better than those with fewer letters.

Besides the statistical view of the ground states, do the chains made of five types of residues have kinetic advantages? Here we report results of Monte Carlo (MC) simulations on the ability of the model chains to fold with different numbers of residue types. We use the average contact overlap — that is, the number of common contacts, $Q_0 = \langle Q_S \rangle$ between the native structure

and the lowest energy structure of a MC trial — as a measure of folding ability. This procedure is similar to that used by Dinner *et al.*[33] (also see ref. 5 and discussions and refs therein). A monotonic increase in the values of $Q_0(N)$ is obtained as the number of residue types N increases (solid circles in Fig. 4*b*). When $N = 5$, the value of $Q_0(5)$ reaches 0.90, which is ~90% of $Q_0(20)$, the value of the contact overlap for sequences with the natural set of residues. Physically, a large value of $Q_0(N)$ means that the sequences fold to the designed native state fast and also that there are more common contacts or greater similarity of contacts with the native state. That is, the monotonic increase in the value of $Q_0(N)$ implies that sequences with more residue types have better kinetic accessibility and folding ability. Thus, the large value of $Q_0(5)$ suggests that the sequences with five types of residues are kinetically similar to those with more types of residues or even 20 types. In other words, a five-type reduction is intrinsically of kinetic advantage and is a good reduction.

This conclusion is supported both by molecular design experiments[20–23] and by some other physical insights[4,6–8,12,13]. In addition, $Q_0(N)$ for some sequences which are optimized using only the reduced alphabets are also obtained (the inset in Fig. 4*b*). They show a similar monotonic increase as that of direct substitution of residues by the representative letters for the standard sequences, which implies the dependence of folding ability of sequences on the number of residue types for the reduction rather than on the specific choice of sequences. It is worth noting that the value of $Q_0(2) = 0.35$ for $N = 2$ is approximately equal to the average common contact overlap between the designed native structure and all possible compact structures of a 27-monomer chain placed on a $3 \times 3 \times 3$ cubic lattice.

This may mean lower kinetic accessibility or lower folding ability for the sequences substituted (or optimized) by only two types of residues (see scheme 4 in Fig. 4*a*). Note that for the optimized sequences with two types of residues, the compositions are fixed when we optimize them. If the compositions are not fixed, the optimization process would produce sequences with good folding behaviors similar to those using effective H and P residues[34], which have higher $Q_0(2)$ values. A more detailed discussion of these aspects of our results will be reported elsewhere.

Finally, we have also examined the funnel characteristics of the energy landscape for the different reductions based on a coarse-grained model[17,35]. We find that the obvious change of steepness of the funnel shows a transition between those proteins that fold well (sequences with five or more types of residues) and those that do not (sequences with few types of residues). These will be presented elsewhere. In short, the five-letter scheme may be a form of simplified representation of natural proteins.

Conclusions

Minimized mismatch as a function of the number of residue types presented in this paper gives us a conceptual way to understand the process of reduction to schemes with different sets of residues. The plateaus in three regions give suitable representations of proteins relevant to different interactions, such as hydrophobicity or polarity. The five-letter case corresponds to the condition with coarse-grained consideration on the hydrophobic interaction; the seven-letter case reflects the basic properties of both hydrophobic and polar residues; and the trivial 20-letter case, which includes all particular interactions between residues, makes a full description of proteins. These results suggest that the description of proteins can be simplified in different coarse-grained levels.

Methods

Reducing interaction matrix. Given a group number N, say $N = 5$, L residues (L is the number of residues in a hypothetical protein to be reduced, and in this work, L = 20) are arbitrarily assigned into N groups such as $\{A(i), 1 \le i \le n_A\}$ in group A, $\{B(j), 1 \le j \le n_B\}$ in B and so on, where element numbers n_x satisfy $n_A + n_B + n_C + n_D + n_E = L$. Each group is represented by a new letter, say group A by R_A. Then, with a new alphabet, $\{R_x, X = A,B,C,D,E\}$, an $L \times L$ matrix of contact potentials between residues is reduced to $N \times N$ blocks (Fig. 1). Each block XY in reduced matrix contains a number of elements $\{Z(X(i),Y(j)), 1 \le i \le n_X, 1 \le j \le n_Y\}$, in which Z(X(i),Y(j)) is the contact interaction between residues X(i) and Y(j). Here $X,Y \in \{A,B,C,D,E\}$.

Relation between two blocks. The relation between two blocks XY and X'Y' can be obtained by considering the values of elements in corresponding blocks. We define a relative strength **P** between blocks XY and X'Y' as:

$$\mathbf{P}(XY,X'Y') = \text{sign}\,[\Sigma_{ijkl}\, p_{XYX'Y'}(i,j,k,l)/(n_{XY}n_{X'Y'})] \qquad (1)$$

where $p_{XYX'Y'}(i,j,k,l) = \text{sign}[Z(X(i),Y(j)) - Z(X'(k),Y'(l))]$ is the relative strength between elements Z(X(i),Y(j)) and Z(X'(k),Y'(l)). The 'sign' function is assigned a value according to the following rule: sign[x]=1, 0 and -1 for x > 0, x = 0 and x < 0, respectively, and $n_{XY} = n_X n_Y$, $n_{X'Y'} = n_{X'}n_{Y'}$ are the numbers of elements in blocks XY and X'Y'. When **P** is positive, zero or negative, the interaction between elements in block XY is weaker (or more positive), equal or stronger (or more negative; note that the elements of the MJ matrix are all negative) than that in block X'Y', respectively. For XY=X'Y', the relative strength is defined to be zero, **P**(XY,XY)=0.

Mismatch pairs. Under a certain reduction, when $p_{XYX'Y'}(i,j,k,l)$ is different from the corresponding **P**, the Z-pair {Z(X(i),Y(j)), Z(X'(k),Y'(l))} is a mismatch pair.

Mismatch between two blocks. Mismatch between two blocks XY and X'Y' is defined as the ratio of mismatch pairs to all possible Z-pairs. That is:

$$M(XY,X'Y') = \Sigma_{ijkl}\,[1 - \delta(p_{XYX'Y'}(i,j,k,l) - \mathbf{P}(XY,X'Y'))] / (n_{XY}n_{X'Y'}) \qquad (2)$$

where $\delta(x)$ is 1 when x=0, and 0 otherwise. It is noteworthy that for two identical blocks X'Y', XY, there are Z-pairs that will be compared twice. For example, for a Z-pair with two elements a and b, there are $p_{XYXY}(i,j,k,l) = \text{sign}[a - b]$ and $p_{XYXY}(k,l,i,j) = \text{sign}[b - a]$. However, the information embedded in $p_{XYXY}(i,j,k,l) > 0$ (or <0) is the same as that in $p_{XYXY}(k,l,i,j) < 0$ (or >0). Thus, only $p_{XYXY}(i,j,k,l) > 0$ needs to be considered in counting the mismatch pairs.

Mismatch of all blocks. Once the group number N and the number of elements in each group (that is, a set $\{n_X, n_Y \ldots\}$, are given) the residues can be distributed into these groups. Then, the mismatch of all blocks can be obtained by $M_{ab} = \sum_{XY,X'Y'} M(XY,X'Y') / \sum_{XY,X'Y'} 1$ with the summation over the upper triangle of the matrix (Fig. 1).

Minimization of mismatch with fixed element number in each group. When given a group number N and a set $\{n_X, n_Y \ldots\}$, such as a set {8,3,2,2,5} for $N = 5$, the exchange of two residues in two groups results in different mismatches M_{ab}. The new grouping due to such an exchange is accepted with a Metropolis probability $\exp(-\Delta M_{ab}/T)$, in which ΔM_{ab} is the change of the mismatches and T = 0.1 is an artificial 'temperature'. After a number of iterations, a minimum of mismatch M_{abmin} relating to a given N and a given set $\{n_X, n_Y \ldots\}$ is obtained. Generally, such a minimization process (10^7 MC steps), say for $N = 5$ with the set of element numbers {8,3,2,2,5}, may cost several hours on a DEC workstation (au-433MHz).

Minimization of mismatch for a fixed group number. For a given group number N, there are many sets with various numbers of elements in N groups, $\{n_X, n_Y ...\}$. The numbers of the sets can be enumerated as: 1, 10, 33, 64, 84, 90, 82, 70, 54, 42, 30, 22, 15, 11, 7, 5, 3, 2, 1, 1 for group number from N = 1 to 20, which gives a total of 627. For different sets of $\{n_X, n_Y ...\}$, we may have different M_{abmin}. Then a global minimized mismatch M_g from all M_{abmin} relating to different sets $\{n_X, n_Y ...\}$ can be obtained, and a corresponding set $\{n_X, n_Y ...\}$ is selected as the optimal reduction for a given N (Fig. 2*b*).

Dividing of groups with single residues. Due to the definition of the mismatch, the reduction that includes groups with a single residue often has little mismatch. Thus, these reductions create competing minima, and affect the dividing of groups, especially when the number of groups $N < 9$. However, for a small number of groups, physical considerations imply that the dominating factor for the minimization of the mismatch comes from the dividing of the polar groups or the hydrophobic groups. Thus, for $N < 9$ the occurring of groups with a single residue (singlets) is restricted. That is, in finding the global minimum M_g from all minimized mismatches M_{abmin} relating to various sets $\{n_X, n_Y ...\}$, groups with a single residue are excluded. By using this restriction, a local minimum for a certain set $\{n_X, n_Y ...\}$ that lacks groups with a single residue is taken as the global minimum of the mismatch M_g. When $N \geq 9$, there is no local minimum without singlets. Thus we have to consider groups with a single residue. It is found that, in general, a set of small number of groups with a single residue has a local minimum mismatch. This set is taken as the grouping of the residues. The physical details will be presented elsewhere.

Determination of representative residues. In any reduction scheme, such as that of Baker and coworkers, each group, for example group A, is represented by a residue R_A. An optimal representative residues $\{R_X, X = A,B...\}$ may be determined by the minimization of the deviation $\Delta=\sum_{XYX'Y'}[1-\delta(p_{XYX'Y'}(R_X,R_Y,R_{X'},R_{Y'})-\mathbf{P}(XY,X'Y'))]$. That is, the selected residues $\{R_X, X = A,B...\}$ must satisfy the relative strengths between blocks, or has little deviation from **P**. For example, when the strength of block AB is weaker (or more positive) than that of block CD (that is, **P**(AB,CD) > 0) the interaction between R_A and R_B should also be weaker (or more positive) than that between R_C and R_D — that is, $Z(R_A,R_B) > Z(R_C,R_D)$. A similar sampling method as that in the section entitled "Minimization of mismatch with fixed elements in each group" is used to minimize Δ, and a set of residues can be obtained such as the set **IAGEK** in Fig. 3.

Acknowledgments
We thank D. Baker, P. G. Wolynes and K. A. Dill for comments on our manuscript, and H. S. Chan and D. Thirumalai for valuable suggestions. W.W. thanks the support by the Outstanding Young Research Foundation of the National Natural Science Foundation, the National Natural Science Foundation, the Nonlinear Project of the National Science and Technology Committee, and the National Laboratory of Solid State Microstructure.

Correspondence should be addressed to W. W.
email: wangwei@netra.nju.edu.cn

Received 22 December, 1998; accepted 10 August, 1999.

1. Ron E. *Recent developments in theoretical studies of proteins.* (World scientific publishing Co., Singapore; 1996).
2. Sfatos, C. D. & Shaknovich, E. I. *Phys. Rep.* **288**, 77–108 (1997).
3. Pande, V. S., Yu Grosberg, A. & Tanaka, T. *Biophys. J.* **73**, 3192–3210 (1997).
4. Dill, K. A. & Chan, H. S. *Nature Struct. Biol.* **4**, 10–19 (1997).
5. Chan, H. S. & Dill, K. A. *Proteins: Struct. Funct. Genet.*, **30**, 2–33 (1998).
6. Wolynes, P. G. *Nature Struct. Biol.* **4**, 871–874 (1997).
7. Sali, A., Shakhnovich, E. I. & Karplus, M. *Nature* **369**, 248–251 (1994).
8. Socci, N. D., Onuchic, J. N. & Wolynes, P. G. *J. Chem. Phys.* **104**, 5860–5868 (1996).
9. Pain, R. H. *Mechanisms of protein folding.* (Oxford University Press, Oxford; 1994),
10. Shakhnovich, E. I. & Gutin, A. M. *Biophys. Chem.* **34**, 187–199 (1989).
11. Bryngelson, J. D. & Wolynes, P. G. *J. Chem. Phys.* **93**, 6902–6915 (1989).
12. Onuchic, J. N., Wolynes, P. G., Luthey-Schulten Z. & Socci, N. D. *Proc. Natl. Acad. Sci. USA* **92**, 3626–3630 (1995).
13. Goldstein, R. A., Luthey-Schulten, Z. & Wolynes, P.G. *Proc. Natl. Acad. Sci. USA* **89**, 9029–9033 (1992).
14. Li, H., Tang. C. & Wingreen, N. S. *Phys. Rev. Lett.* **79**, 765–768 (1997).
15. Chan, H. S. and Dill, K. A. *J. Chem. Phys.* **92**, 3118–3135 (1990); *Erratum* **107**, 10353 (1997).
16. Shakhnovich, E. I. & Gutin, A. M. *J. Chem. Phys.* **93**, 5967–5971 (1990).
17. Camacho, C. J. *Phys. Rev. Lett.* **77**, 2324–2327 (1996).
18. Klimov, D. K. & Thirumalai, D. *Phys. Rev. Lett.* **76**, 4070–4073 (1996).
19. Veitshans, T., Klimov D. & Thirumalai, D. *Folding & Design* **2**, 1–22 (1997).
20. Regan, L. & Degrado, W. F. *Science* **241**, 976–978 (1988).
21. Kamteker, S., Shiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. *Science* **262**, 1680–1685 (1993).
22. Davidson, A. R., Lumb, K. J. & Sauer, R. T. *Nature Struct. Biol.* **2**, 856–863 (1995).
23. David, S. R. *et al. Nature Struct. Biol.* **4**, 805–809 (1997).
24. Chan, H. S. & Dill, K. A. *Proteins: Struct. Funct. Genet.* **24**, 335–344 (1996).
25. Li, H., Helling, R., Tang, C., & Wingreen, N. S. *Science* **273**, 666–669 (1996).
26. Yue, K. *et al. Proc. Natl. Acad. Sci. USA* **92**, 325–329 (1995).
27. Mirny, L. A., Abkevich, V. I., & Shaknovich, E. I. *Proc. Natl. Acad. Sci. USA* **95**, 4976–4981 (1998).
28. Shakhnovich, E. I. *Phys. Rev. Lett.* **72**, 3907–3910 (1994); *Erratum* **74**, 2618 (1995).
29. Miyazawa, S. & Jernigan, R. L. *J. Mol. Biol.* **256**, 623–644 (1996).
30. Jernigan, R. L. & Ting, K.-L. In *Statistical mechanics, protein structure, and protein substrate interactions.* (ed. Doniach, S.) 317–326 (Plenum Press, New York; 1994).
31. Thomas, P. D. & Dill, K. A. *J. Mol. Biol.* **257**, 457–469 (1996).
32. Shakhnovich, E. I. & Gutin, A. M. *Proc. Natl. Acad. Sci. USA*, **90**, 7195–7199 (1993).
33. Dinner, A., So, S. S. & Karplus, M. *Proteins: Struct. Funct. Genet.* **33**, 177–203 (1998).
34. Micheletti, C., Seno, F., Maritan, A. & Banavar, J. *Phys. Rev. Lett.* **80**, 2237–2240 (1998).
35. Xing, Z. W., Wang, J. & Wang, W. *Phys. Rev.* E **58**, 3552–3556 (1998).
36. Socci, N. D. & Onuchic, J. N. *J. Chem. Phys.* **103**, 4732–4744 (1995).