Protein Engineering vol. 16 no. 5 pp. 323–330, 2003
DOI: 10.1093/protein/gzg044

Reduction of protein sequence complexity by residue grouping

Tanping Li, Ke Fan, Jun Wang and Wei Wang[1]

National Laboratory of Solid State Microstructure, Institute of Biophysics and Department of Physics, Nanjing University, Nanjing 210093, China

[1]To whom correspondence should be addressed.
E-mail: wangwei@nju.edu.cn

It is well known that there are some similarities among various naturally occurring amino acids. Thus, the complexity in protein systems could be reduced by sorting these amino acids with similarities into groups and then protein sequences can be simplified by reduced alphabets. This paper discusses how to group similar amino acids and whether there is a minimal amino acid alphabet by which proteins can be folded. Various reduced alphabets are obtained by reserving the maximal information for the simplified protein sequence compared with the parent sequence using global sequence alignment. With these reduced alphabets and simplified similarity matrices, we achieve recognition of the protein fold based on the similarity score of the sequence alignment. The coverage in dataset SCOP40 for various levels of reduction on the amino acid types is obtained, which is the number of homologous pairs detected by program BLAST to the number marked by SCOP40. For the reduced alphabets containing 10 types of amino acids, the ability to detect distantly related folds remains almost at the same level as that by the alphabet of 20 types of amino acids, which implies that 10 types of amino acids may be the degree of freedom for characterizing the complexity in proteins.

Keywords: compositions of amino acids/protein fold recognition/reduced alphabet of amino acids/residue grouping/similarity matrix

Introduction

Proteins are the elementary blocks which execute biological functions in living organisms. There are many types of proteins in nature that carry out various complicated activities. Proteins are composed of 20 types of naturally occurring amino acids, and the majority of proteins are encoded by complex patterns of these 20 types of amino acids. That is, 20 types of amino acids introduce not only diversity and complexity into proteins, but also some specific propensities. For example, some amino acids are similar in physicochemical properties (Mathews and Van Holde, 1995) and mutations of amino acids can be tolerated in many regions of a sequence (Sinha and Nussinov, 2001). It has been discovered experimentally that some designed proteins with fewer than 20 types of residues can have stable native structures and contain nearly as much information as natural proteins (Regan and Degrado, 1988; Kamtekear *et al*., 1993; Davidson *et al*., 1995; Riddle *et al*., 1997).

Recently, a 57 residue Src SH3 domain with a β-barrel-like structure was studied (Riddle *et al*., 1997), and 38 out of 40 targeted residues in the domain could be replaced with five types of residues (Ile, Ala, Glu, Lys, Gly). From a physics viewpoint, this may imply that a 20 letter alphabet can be reduced into an *N* letter alphabet by clustering the similar amino acids into *N* groups, and then *N* letters can be chosen as the representative residues of these *N* groups (Chan, 1999; Wang and Wang, 1999). Obviously, the simplest reduction is the so-called HP model (Chan and Dill, 1989; Lau and Dill, 1989), where 20 types of amino acids are divided into two groups: H group and P group (H, hydrophobic residues; P, polar residues). Interestingly, such a type of simple two-letter HP model or the HP-like patterns could reproduce, to some extent, the kinetics and thermodynamics of protein folding and could be used to study the mechanism of folding (Regan and Degrado, 1988; Kamtekear *et al*., 1993; Davidson *et al*., 1995). Previously, a five-letter alphabet based on the statistical potential matrix by Miyazawa and Jernigan (MJ) [a pairwise interaction potential between amino acids (Miyazawa and Jernigan, 1996)] was studied (Chan, 1999; Wang and Wang, 1999). In that reduction, five representative residues were given as (Ile, Ala, Glu, Lys, Gly), which coincide with the experimental results of the 57 residue SH3 domain by Baker and coworkers (Riddle *et al*., 1997). (Hereafter, the residues are simply represented as single letters.) One of the advantages of such a reduction is that it reduces greatly the complexity of the protein sequences. It has been shown that sequences with these five types of letters have good foldability and kinetic accessibility in studies of protein-model chains (Wang and Wang, 2000). Some other simplified alphabets were also proposed (Reidhaar-olson and Sauer, 1988; Smith and Smith, 1990; Murphy *et al*., 2000; Soils and Rackovsky, 2000; Cieplak *et al*., 2001). For example, an alphabet studied by Murphy *et al.* (Murphy *et al*., 2000) was obtained from the similarity matrices of the amino acids that characterize the correlation between the amino acids. Cieplak *et al.* (Cieplak *et al*., 2001) simplified the folding alphabet based on a 'distance' of the hydrophobicity of the natural residues defined through the MJ matrix. The alphabet by Solis and Rackovsky (Solis and Rackovsky, 2000) was obtained by reserving the maximal information in proteins. In this work, the authors analyzed the relation between residues based on their similarities that are extracted from the interactions between the amino acids or amino acid sequence alignment, by using various clustering schemes. The residues were depicted as a vector in 20-dimensional space spanned with their inter-relationship. To some degree, however, these descriptions omit some possible correlations of the residues within the groups. Is the consideration on the detailed distribution or correlation of the residues in the groups helpful for producing useful groupings related to some specific proteins? Obviously, this is an important question for amino

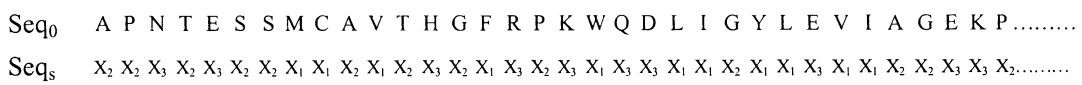

Fig. 1. Sketch map for the simplification of 20 types of residues for group number $N = 3$. The three groups are: (F, W, Y, C, M, I, L, V), (A, G, T, S, P) and (N, Q, D, E, H, R, K). The representative residues of three groups are set as X_1, X_2 and X_3, respectively. Seq_0 is the original protein sequence and Seq_s is the simplified one.

acid grouping studies. Its answer might promote the application of the grouping results.

The naturally occurring frequencies of 20 types of residues in proteins follow some type of pattern. The compositions of 20 types of amino acids in proteins may provide useful information for the simplification of the residue alphabet. In this paper, we integrate the information on compositions of residues into the reduction of the residue alphabet, and cluster similar amino acids into groups using a global alignment method. The representative residues for each group are also obtained. Then, the recognition tests with the reduced alphabets are discussed. By using a simplified BLOSUM matrix based on these schemes, we perform an 'all-against-all' sequence alignment and make coverage detection on the distantly related homologous proteins throughout the database SCOP40 (Brenner *et al*., 1998) for various levels of reduction. A platform of around 10 types of residues is obtained in a plot of the coverage versus the reduced alphabet size, indicating that 10 types of residues may be the minimum number of letters required to construct a rational folding model.

The paper is organized as follows. In the next section, we study the simplified alphabets by dividing the naturally occurring 20 types of amino acids into different numbers of groups. We make the simplification based on similarity score from the BLOSUM matrix. In the following section, we perform an evaluation of our simplified alphabets by comparing the coverage of the simplified matrix with the database of SCOP40. Finally, we present a brief summary.

Grouping residues based on reasonable simplification of protein sequence

Methods

As is already known, the so-called sequence alignment method is generally used to measure the degree of similarity between two protein sequences. To evaluate competing alignments, a substitution matrix is necessary, in which different scores are assigned to different exchanges of one amino acid with another. In general, a positive score indicates that two residues are similar, and substitution or mutation between them may be applicable. While a negative score implies that two residues are fairly different, and mutation between them may be unfavorable. There are many substitution matrices proposed according to the different scoring schemes, among which BLOSUM is the most widely used (Henikoff and Henikoff, 1992) and BLOSUM62 is usually the default choice for many sequence alignment programs, e.g. BLAST (Altschul *et al*., 1990).

In this work, we use BLOSUM62 as a starting point to simplify the amino acid alphabet, and try to find the related representative residues for each reduced alphabet. Our purpose here is to find an optimal grouping scheme with which the simplified sequence can reserve the maximal information on the original sequence. Our physical hypothesis is as follows. Suppose that a protein sequence, denoted Seq_0, is a specific arrangement of 20 types of amino acids (here we consider 20 types of residues) (see Figure 1). The sequence can be simplified by classifying the amino acids into different groups and by replacing the amino acids with the representative one in its group. For example, if 20 types of amino acids are classified into N groups, say $N = 3$ groups, we have group 1 with residues (F, W, Y, C, M, I, L, V), group 2 with (A, G, T, S, P) and group 3 with (N, Q, D, E, H, R, K). We could use X_1 to represent the whole residues in group 1, X_2 those in group 2 and X_3 those in group 3, i.e. X_1, X_2 and X_3 are the representative residues in group 1, group 2 and group 3, respectively. Thus, the parent sequence Seq_0 is simplified into Seq_s, and such a simplification is indicated in Figure 1. Our main task is to calculate the similarity scores between the two sequences Seq_0 and Seq_s for various assignments of residues in N groups, among which one assignment with the maximal similarity score will be the best grouping of the residues. Such a grouping is regarded as the most reasonable one since it may reserve maximally the information or the content of the parent sequence. In the following, we give more details on how to obtain the optimal grouping scheme.

Since X_i could be any one of the residues in the ith group, we define the jth residue in the ith group as $X_i(j)$. Specifically, the similarity score of $X_i(j)$ in the ith group in Seq_s to the kth residue R_k in the ith group in Seq_0 is

$$S(X_i(j), R_k) = Blosum(X_i(j), R_k) \tag{1}$$

where $Blosum(X_i(j),R_k)$ is the element in the substitution matrix for exchange of residue $X_i(j)$ with residue R_k. Then, the similarity score between all pairs of residue R_k and residue $X_i(j)$ in the ith group in the protein sequence is

$$S(X_i(j)) = \sum_{k=1}^{g(i)} m_i(k) S(X_i(j), R_k) \tag{2}$$

where $g(i)$ is the number of residue types in the ith group, $m_i(k)$ is the number of the residue R_k in Seq_0, and k runs over the whole ith group. Since $X_i(j)$ could be any one of the $g(i)$ residues in the ith group, we can use an average similar score for these different choices to describe the simplification:

$$S_i = \left[\sum_{j=1}^{g(i)} S(X_i(j))\right] / g(i) \tag{3}$$

Then, the total similarity score of the simplified sequence Seq_s to the parent sequence Seq_0 is calculated as the sum of the scores over all groups:

$$S = \sum_{i=1}^{N} S_i \tag{4}$$

where N is the number of groups. Clearly, S is a measure of the reservation of the information on the parent sequence. One can see that different assignments or distributions, denoted as groupings, of the residues in all groups lead to different values of S. For a given number of groups N and a set of numbers n_i of residues in each group ($n_1, n_2, \ldots, n_N$) (Wang and Wang, 1999) with $\sum_{i=1}^{N} n_i = 20$, we can always find a specific grouping of residues with a maximal value of S, i.e. $S = S'_{\max}$. Such a grouping is regarded as the best or the most reasonable one since it may reserve the maximal information on the parent sequence or has a maximal similarity to the parent sequence.

It is clear that this grouping scheme is related only to the substitution matrix and the compositions of residues in Seq_0, while the length of Seq_0 and the specific order of 20 types of residues in Seq_0 have no effect on our results. However, a lot of work indicated that the naturally occurring frequencies or the compositions of 20 types of residues in proteins show a specific distribution. It seems that the compositions of residues in proteins (distribution of residues) have reached their equilibrium during their long evolutionary history, and most proteins comply with this distribution of compositions for all residues. Thus, the average compositions of 20 types of amino acids from a radical protein sequence database SWISS-PROT (version 37) are used in the following studies, where over 80 000 sequences in this database are used for statistics (Z.P.Feng and C.T.Zhang, personal communication on the results over 80 000 protein sequences from database SWISS-PROT). The length of sequence is set as 300 in the following calculations.

The number of possible assignments for 20 types of amino acids is enormous. For example, there are more than 51×10^{12} ways to make simplified amino acid alphabets and more than 15×10^{12} simplified alphabets where 20 amino acids are represented by a reduced set of eight symbols (Cannata *et al.*, 2002). Approximately 17 years would be needed to complete the analysis of the 51×10^{12} possible simplified amino acid alphabets if we could calculate it on our computer with an average scanning rate of 6 million simplified alphabets per minute. Thus, an exhaustive search for the maximum of the scores is not feasible. It is also worth noting that the enumeration for the grouping is only for several cases, e.g. the cases of $N = 2, 3, 18$ and 19 (Wang and Wang, 1999, 2002). Thus, a heuristic Monte Carlo (MC) method is used to approach the optimal solution. The MC cycle for a case of $N = 3$ can be performed using the following steps:

(i) For a grouping with N groups, we enumerate the number of sets that characterize the number of residues in every group (Wang and Wang, 1999). For example, for $N = 3$, the number of sets is 33, such as (1, 1, 18), (1, 2, 17), (1, 3, 16), etc. For the set (1, 1, 18), we have one residue in group 1, one residue in group 2 and 18 residues in group 3. Twenty types of residues are randomly assigned in these three groups.

(ii) For a certain distribution of residues in these N groups, the score S can be obtained from Equation 4, which describes the feature of the present grouping.

(iii) Exchange two residues between two groups, i.e. make a move in a space spanned with various groupings. Clearly, this type of move keeps the set unchanged. Whether such a move is accepted or not depends on a Metropolis criterion (Metropolis *et al.*, 1953), i.e. the probability of accepting the move is $P = \exp(S_{\text{old}} - S_{\text{new}})/T_{\text{MC}}$. If P is larger than a random number in [0, 1], the move is accepted; otherwise, the move is rejected. Here, T_{MC} is an artificial temperature for the MC sampling, and is chosen to be 1.0. It is noted that different choices of the value of T_{MC} will not affect the result.

(iv) With this type of move, an MC search is used to find a maximal value of the score $S'_{\max}$. Due to the bias towards the high score case, generally, it may be unnecessary to exhaust the whole space. In our work, 10^7 MC steps are generally enough for the search of the maximal value of $S'_{\max}$.

(v) Go back to step (i), and repeat the simulations for another set, we find another $S'_{\max}$ for this set, i.e. change the numbers of residues in various groups and get the corresponding distribution of the maximum score.

(vi) After the simulations for all sets for a given N, we then find a global maximal $S_{\max}$ among all $S'_{\max}$. The grouping with the global maximum score $S_{\max}$ is taken as our final result of grouping for N groups.

As a further step for the reduction of the complexity of proteins, a set of representative residues for N groups should be identified. In every group, one residue should be set as a representative residue. Thus, for a given reduced grouping, the protein sequence made of 20 types of amino acids can be simplified by a set of representative residues X_i^P with $i = 1, 2, \ldots, N$. The similarity score of the simplified sequence to the parent sequence can be calculated by

$$S^P = \sum_{i=1}^{N} \left[\sum_{k=1}^{g(i)} Blosum(R_k, X_i^P) m(k) \right] \qquad (5)$$

Here, X_i^P could be one of the residues in the ith group. From Equation 5, one can see that different sets of representative residues as input have different values of the scores for the simplified sequences to the parent sequence for a given grouping of 20 types of residues. The most reasonable set of representative residues is the one with the largest score from Equation 5. It is this set of representative residues that can possess the largest similarity between the simplified sequences and the parent sequence for a given N.

Results

The reduced groupings obtained by the above procedures are listed in Table I. We can see that from $N = 12$ to 20 the groupings are basically continuous, i.e. there is no interlace of residues in the nearest levels of reduced groupings. If two residues were separated in two groups in a subtle classification, say N, they should be in the same group in a coarser classification with group number $(N - 1)$. However, there is one exception for the intermediate level of reduction. Residue H stays together with the hydrophilic residues from $N = 2$–4 and $N = 12$–19, but joins to the group (F, W, Y) from $N = 5$ to 9. This type of interlacing may result from the competition between detailed features of the residues between these two groups. This shows the underlying complexity in the grouping problem. Furthermore, those alphabets whose S values are slightly less than the largest one appearing in the MC processes are also recorded from $N = 2$ to 19. For $N = 2$–19, the differences between the largest score and the second largest score are large. This means that the corresponding reduced alphabets have greater superiority compared with the other alphabets, and they are the most reasonable reductions. For other cases, the differences in the S value between the largest

Table I. Clustering of amino acids by the maximal score for different levels of reduction

	Group
2	CMFILVWY AGTSNQDEHRKP
3	CMFILVWY AGTSP NQDEHRK
4	CMFWY ILV AGTS NQDEHRKP
5	FWYH MILV CATSP G NQDERK
6	FWYH MILV CATS P G NQDERK
7	FWYH MILV CATS P G NQDE RK
8	FWYH MILV CA NTS P G DE QRK
9	FWYH ML IV CA NTS P G DE QRK
10	FWY ML IV CA TS NH P G DE QRK
11	FWY ML IV CA TS NH P G D QE RK
12	FWY ML IV C A TS NH P G D QE RK
13	FWY ML IV C A T S NH P G D QE RK
14	FWY ML IV C A T S NH P G D QE R K
15	FWY ML IV C A T S N H P G D QE R K
16	W FY ML IV C A T S N H P G D QE R K
17	W FY ML IV C A T S N H P G D Q E R K
18	W FY M L IV C A T S N H P G D Q E R K
19	W F Y M L IV C A T S N H P G D Q E R K
20	W F Y M L I V C A T S N H P G D Q E R K

Table II. Clustering of amino acids where there is no interlacing for different levels of reduction

	Group
2	CFYWMLIV GPATSNHQEDRK
3	CFYWMLIV GPATS NHQEDRK
4	CFYW MLIV GPATS NHQEDRK
5	CFYW MLIV G PATS NHQEDRK
6	CFYW MLIV G P ATS NHQEDRK
7	CFYW MLIV G P ATS NHQED RK
8	CFYW MLIV G P ATS NH QED RK
9	CFYW ML IV G P ATS NH QED RK
10	C FYW ML IV G P ATS NH QED RK
11	C FYW ML IV G P A TS NH QED RK
12	C FYW ML IV G P A TS NH QE D RK
13	C FYW ML IV G P A T S NH QE D RK
14	C FYW ML IV G P A T S N H QE D RK
15	C FYW ML IV G P A T S N H QE D R K
16	C FY W ML IV G P A T S N H QE D R K
17	C FY W ML IV G P A T S N H Q E D R K
18	C FY W M L IV G P A T S N H Q E D R K
19	C F Y W M L IV G P A T S N H Q E D R K
20	C F Y W M L I V G P A T S N H Q E D R K

score and some others are small. Thus, there may not be much priority for the alphabet of maximal scores compared with those alphabets whose S values are close to the maximal one, and some other parameters should be considered. Non-interlacing of residues for various N may be an important factor to decide the grouping. Next, we propose a method to solve the interlace problem in grouping.

In order to determine the final result of the groupings, two considerations should be made: (i) the gap of scores between the largest one and the others is large for N = 2 and 19, respectively; (ii) the groupings corresponding to the maximum scores in Table I have no interlacing of amino acids from N = 2 to 3 and N = 12 to 19; these groupings should be kept. Therefore, we select the groupings in such a way that for N = 2–3 and N = 12–19, the detailed classifications of amino acids are the same as in Table I. For N = 4–11, we select the groupings in which there is no interlacing of residues between

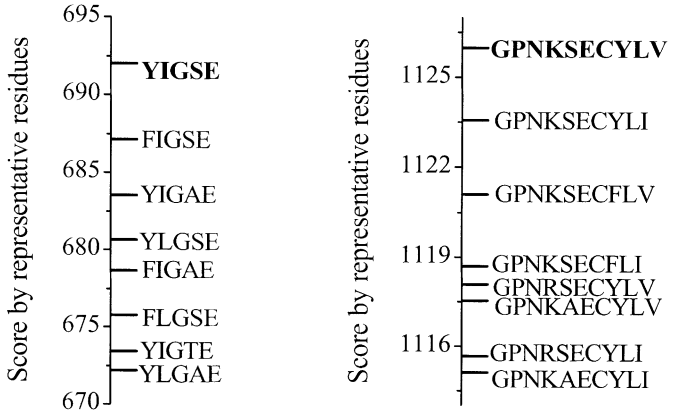

Fig. 2. The representative residues for N = 5 and N = 10 for the no-interlace alphabet. The letters in bold are the best sets from Equation 5.

the different number of groups for N from 2 to 19; meanwhile, the scores obtained from Equation 4 should be as large as possible. The grouping result under such considerations is listed in Table II. For N = 2,3 and N = 12–20, the groupings all have the largest scores. For N = 4–11, the scores are slightly less than their related largest ones. Because all the groupings have no interlacing for N = 2–19, such a grouping is called a no-interlace grouping. It should be noted that most of the classifications for different N in Tables I and II are similar, and there is only a minor difference in the location of the residue H. One should also note that these results are almost the same as different versions of BLOSUM.

The representative residues relating to two groupings in Tables I and II are shown with bold letters. In Table II, the set of representative residues for the no-interlace alphabet of a five-letter alphabet are: Y, G, I, S, E, while the results found experimentally (Riddle *et al.*, 1997) or argued theoretically (Chan, 1999; Wang and Wang, 1999) are I, A, G, E, K. Four of them are coincident, I, A, E, G, since the score for residue A is nearly the same for residue S (see Figure 2 for spectra of the values for the score S^P by different sets of representative residues for N = 5 and 10). The representative residues of the 10-letter alphabet are C, Y, L, V, G, P, S, N, E, K. Note that the score for residue I is also basically the same as for residue V. As will be discussed in the next section, the five residues (I, A, E, K, G) may not be so good for simplifying the complexity of proteins. Nearly saturated information for natural protein, when compared with the whole set of 20 types of residues, can be obtained by including an additional five residues: C, Y, L, P, and N. This means that a 10-letter alphabet may be the minimum number of letters for the simplification of protein complexity. However, this requires further statistical study of the folds of proteins.

Discussion

As expected, residues with similar physicochemical properties are generally grouped together, such as the large hydrophobic residues (L, V, I, M), the large and mainly hydrophobic aromatic residues (F, Y, W), the long-chain positively charged residues (K, R), the alcohols (S, T) and the charged/polar residues (E, D, N, Q). From Table I, some of our results are consistent with those obtained by others (Murphy *et al.*, 2000; Jonson and Petersen, 2001). When N = 2, our results are consistent with the two-letter HP model where H represents the

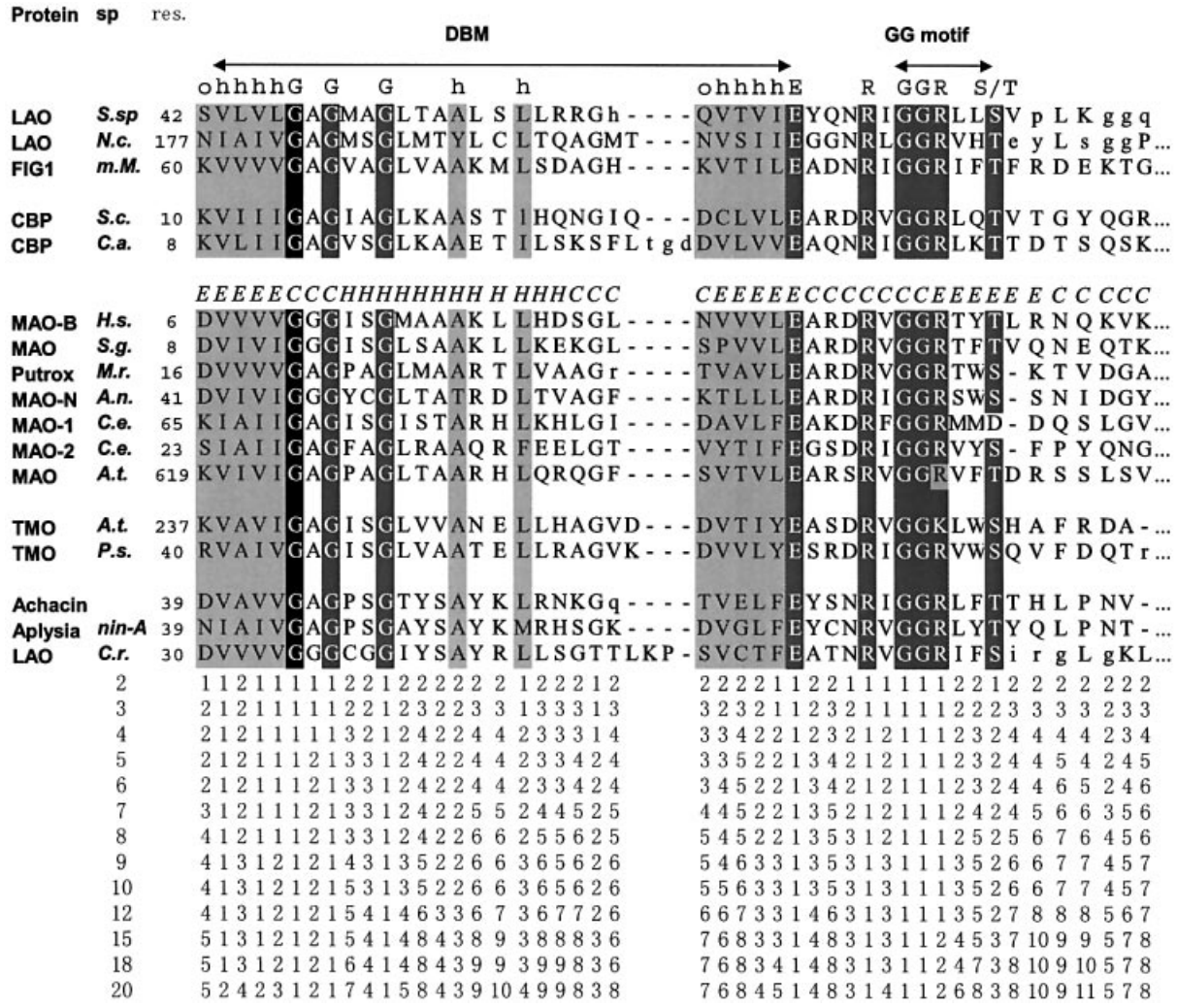

Fig. 3. One region of sequence alignment in the paper by Vallon (Vallon, 2000). The gray shaded regions are the conserved residues and those in the black background are the most conserved. Many conserved residues coincide with our groupings. For each column, the numbers of groups occurring for the residues are listed with respect to different grouping schemes, i.e. different values of N in Table II. The numbers of the occurring groups for alignment positions in the gray and black background are less than those of the no-conserved region.

hydrophobic group (C, M, F, I, L, V, W, Y) and P the hydrophilic group (A, G, T, S, N, Q, D, E, H, R, K, P). With further division for $N = 3$, the hydrophilic group is divided into two parts, namely the residues (A, G, T, S, P) and (N, H, Q, E, D, R, K), while the hydrophobic group with a stronger interaction is reserved. For $N = 4$, the hydrophobic group is divided into two parts, i.e. the aromatic residues (F, W, Y) are separated from the more hydrophobic residues (C, M, I, L, V). In the intermediate levels of the reduction after $N = 5$, some residues are singled out, such as the residues G, P and C, respectively. Furthermore, in the later level of reductions, some residues with very similar chemical propensities are still kept in the same group, such as the residues S and T, the residues I and V, the residues R and K, etc.

It is noted that the residues (G, P, C) are easily singled out rather than clustered with other residues after N reaches a certain number. Specifically, the residues G, P and C are singled out at $N = 5$, 6 and 10, respectively. This is in accordance with the results by Solis and Rackovsky (Solis and Rackovsky, 2000), and can be reasonably understood from protein structure. The influence of three residues (G, P, C) on folding of proteins is especially different. The amount of residue G is large in most proteins, especially in loop or turn regions. Thus, to simplify protein sequence by reducing the alphabet and to reserve the information on the original sequence as much as possible, it is rational that the residue G is reserved rather than mutated. The reason for the residue P being separated in the grouping is simple. The side chain of residue P forms a ring with the peptide backbone and the adjacent backbone conformations are severely affected. Thus, it has a high preference in the β region. The residue C is quite different since it can form a disulfide bond that is very important in stabilizing the protein structure (Solis and Rackovsky, 2000).

Now, let us make a comparison of the groupings in a real case. It is well known that sequence motifs are those regions of proteins where variability is limited, usually because of structural constraints linked either to folding or to interactions with cofactors, substrates or other proteins. It is found that some regions of motifs are basically conserved. The dinucleotide-binding motif (DMB) is one of the classical motifs, and it is constituted of beta–alpha–beta (Vallon, 2000). As described

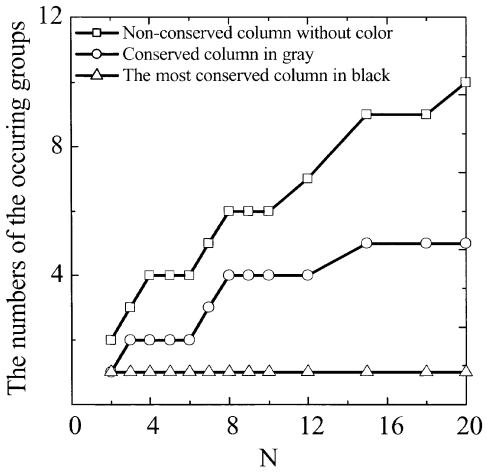

Fig. 4. The numbers of occurring groups versus the reduced group number *N* are plotted for three representative positions in Figure 3. For the most conserved columns (in the black background), the numbers of groups occurring are almost equal to one. For the columns with less conservation (in the gray background), the numbers of groups occurring are also small. This is because the properties of the amino acids in these conserved columns are always similar, and they are grouped in the same groups. For the non-conserved columns without color, the numbers of groups occurring are apparently larger than those in the above two cases.

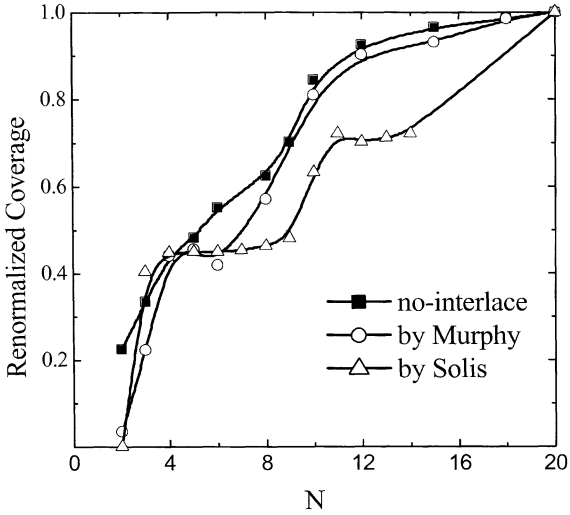

Fig. 5. The coverage scaled with its value for $N = 20$ for database SCOP40 versus the reduced letter *N*. The reduced no-interlace alphabets derived by the method described in the text are shown (filled squares). As a comparison, two curves of the alphabets by Murphy *et al.* (Murphy *et al*., 2000) (open circles) and Solis and Rackovsky (Solis and Rackovsky, 2000) (triangles) are also plotted.

in Vallon's work, classical DBM motifs can be found in the N-terminal part of the sequence in most cases of flavoproteins. The 'GG motif' (RxGGRxxS/T) is found in L-amino acids oxidase (LAOs) and in a wide variety of other flavoprotein families. The conservation of some regions in the DMB and GG motifs is shown in figure 1 of the paper by Vallon (Vallon, 2000). The grouping results are given in Figure 3, where the gray shaded regions are conserved parts and the most conserved residues are shown with a black background. From Figure 3, one can see that most of the residues for alignment in conserved regions coincide with the two-letter HP model from our grouping. That is, the residues are either all hydrophobic or all polar in the conserved positions. Some residues are coincident with that in the groups of our grouping in Table II, such as the residues with large hydrophobic side chains (L, V, I, M), long-chain positively charged residues (K, R) and alcohol residues (S, T). These residues have similar chemical properties. The residues in each conserved position could be classified into several groups according to our grouping in Table II for different *N*. The occurring numbers of the groups for each position are shown below the sequences. One can see that for the most conserved position in the black background, the number of occurring groups is quite small. That is, the number of groups is almost equal to one for all those columns. The regions in the gray positions follow the same pattern, but the conservation becomes slightly weaker. For the positions without color, the numbers of the occurring groups are larger than those of the above two regions. From Figure 3, we can also see some typical examples of the similarity between the residues (Miyata *et al*., 1979). Residues I and V play the same role and they are exchangeable since they are basically the same according to our grouping. One more example is that the buried residue V or I can be replaced by residue L or V and the structures and functions are the same when a family of CEA-like protein sequences is aligned (Bates *et al*., 1992). Such examples are also for residues M and L, and residues A and G, and so on. In addition, many experimental findings also indicate similar results (Reidhaar-olson and Sauer, 1988). In Figure 4, the numbers of the occurring groups versus the reduced group number *N* is plotted. One can see that the curves really relate to the above-mentioned results for three types of positions, i.e. the smaller the number of occurring groups for an aligned column, the larger is the similarity of these aligned residues.

Performance evaluation of reduced alphabets

Methods

So far, we have studied the simplification of residue alphabets by comparing the similarity scores between the sequences through the alignment. In this section, let us put these simplified alphabets to real protein sequences and evaluate the effectiveness of these simplifications by comparing the coverage of the homologous sequences detected via a simplified matrix to that denoted in the database of SCOP40 (Brenner *et al*., 1998). SCOP is a database of structural classification of proteins and provides a detailed and comprehensive description of the relationship of the known protein structures (Murzin *et al*., 1995). The classification of the proteins in SCOP includes four levels: the classes, the folds, the superfamilies and the families. Proteins in the same superfamily and family are believed to share the same ancestor, that is, they are homologous. Furthermore, SCOP40 is a sub-database extracted from SCOP (version 1.36). It consists of 1323 sequences that share no more than 40% sequence identity and represents all distantly related proteins in Protein Data Bank (PDB). These sequences are divided into 639 homologous superfamilies. The total number of aligned sequence pairs in the database SCOP40 is 1 750 329, in which 9044 homologous pairs are marked by the database SCOP40 itself. Our evaluation of the effectiveness of the simplifications will be based on the alignment between all the sequence pairs in this database.

For the detection of protein homology from protein sequences, a program named BLAST (Altschul *et al*., 1990)

is a widely used tool. In BLAST, the method of alignment of two protein sequences is to seek an equal-length segment that has a maximal aggregate score by a similarity matrix, such as BLOSUM62. Thus, the homology of proteins can be detected using a program called BLASTP, and an 'all-against-all' alignment between the sequences can then be worked out for the database SCOP40. The coverage is defined as the number of protein sequence pairs M, with the aligned score larger than an expectation threshold value (E-value) divided by the total number of the homologous pairs in SCOP40 (i.e. 9044 pairs):

$$C = M/9044 \tag{6}$$

In addition, a function of error per query (EPQ) is defined as the total number of non-homologous sequence pairs detected by the BLAST program above the same threshold divided by the total number of the queries, i.e. 1 750 329. Here the value of EPQ is set as 0.001, which means that only 0.1% errors occurred for the homologous detection. By varying the threshold of the E-value, we choose the value of coverage C as our final result by keeping the related EPQ value <0.1%. The gap insertion and elongation parameters for the alignment are set to be –11 and –1, respectively.

For the 'all-against-all' sequence alignment in BLASTP, a similarity matrix, usually the BLOSUM62, is needed. Now in our case, for various levels of simplifications of the residues, the BLOSUM62 matrix should be simplified according to the reduced residue alphabets. The simplification for the elements in the BLOSUM is to replace the original elements of BLOSUM62 with an average if these related residues are in the same group, but not to make the substitutions of residues falling into the same group. For the N-letter simplification of similarity matrix, the similarity score between residues belonging to the ith group and residues in the jth group are averaged by

$$Blosum_{ij}^{N} = \sum_{k=1}^{g(i)} \sum_{l=1}^{g(j)} Blosum_{kl} / \sum_{k=1}^{g(i)} \sum_{l=1}^{g(j)} \tag{7}$$

Here k runs over the residues of the ith group, l runs over the residues of the jth group, $g(i)$ is the number of proteins in the ith group and $g(j)$ is the number of proteins in the jth group.

Results and discussion

The coverage for database SCOP40 versus the reduced letter N corresponding to the no-interlace alphabet (Table II) for EPQ = 0.001 is shown in Figure 5 (here the coverage values are normalized by the coverage at N = 20). From Figure 5, it is clear that with different levels of reduction, the coverage decreases as N decreases, i.e. the capability of the reduced alphabets to recognize the protein sequence pattern decreases when compared with that of the 20 letter alphabet (solid squares). From N = 10 to 20, the values of the coverage are >0.85, and for N = 8, the coverage decreases to ~0.56. Further decreasing the value of N decreases the coverage rapidly. For N = 2, the coverage is relatively small, 0.22, which is for the two-letter HP case. Obviously, there is a plateau for $N \geqslant 10$, which characterizes the saturation of the coverage to its value of N = 20. This means that groups more than N = 10 will not further increase the efficiency of the description of the complexity of proteins from the aspect of the sequence alignment. Thus, a number around N = 10 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 by Baker and co-workers (Plaxco *et al*., 1998), and also relates to our previous work on the reduction of the complexity from the aspect of residue–residue interaction (Wang and Wang, 2002).

For comparison, we also calculate the coverage related to the clustering alphabets by Murphy *et al.* (see figure 1 in the paper by Murphy *et al*., 2000) and by Solis and Rackovsky [see table 1(b) in the paper by Solis and Rackovsky, 2000]. For the alphabets by Solis and Rackovsky, most of the values of the coverage are smaller than those of our alphabets (see the curve with the open triangles). The curve by the alphabets of Murphy *et al.* gives similar results to ours except for N = 2 and 6. One can see that the retained coverage for N = 2 is ~0.22 by our alphabet, compared with the values almost being zero for the alphabets by Murphy *et al*. and by Solis *et al*. Our result for N = 2 indicates that there is still some structure information encoded in the protein sequences even for the two-letter simplification if the clustering or grouping of the amino acids is reasonable. This is consistent with the results of some work on protein folding by the HP model (Regan and Degrado, 1988; Kamtekear *et al*., 1993; Davidson *et al*., 1995).

Summary

Previously, many studies on the simplification of amino acid alphabets have been obtained according to different criteria such as physicochemical properties. The general properties of various residues and the protein sequence features could be suggested from these reduced amino acid alphabets. For example, for the case of N = 2, i.e. the simplest case, the HP model was derived in these studies. In addition, some residues with similar chemical properties tend to be grouped together, such as the hydrophobic residues (I, V) and the aromatic residues (F, W, Y). Nevertheless, different simplifying schemes embody different propensities of residues. For example, the reduced alphabet derived from the MJ matrix may be useful for folding since the simplification is based on the interaction between the residues. The reduced alphabets in this work are derived from amino acid substitutions by scoring similarities via the similarity matrix, which may be helpful for the recognition of protein folds.

In summary, in this work we obtain reduced amino acid alphabets by using a sequence alignment. The selected alphabet reserves mostly the maximum information on the original protein sequence. The alphabet is similar in some letters to studies by others. However, our results are more reasonable from many aspects. Our conclusion is that 10 types of amino acids may be the degree of freedom for characterizing the complexity in proteins. With 10 types of amino acids, the information in the protein could make the protein closer to that consisting of 20 amino acids. Some further work on protein folding and design is necessary to clarify this point.

Acknowledgements

This work was supported by the NNSF of China (grant no.s 10074030, 90103031 and 10021001) and the Nonlinear Science Project (973) of the NSM.

References

Altschul,S.F., Gish,W., Miller,W., Myerse,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.

Bates,P.A., Luo,J.C. and Sternberg,M.J.E. (1992) *FEBS Lett.*, **301**, 207–214.

Brenner,S.E., Chothia,C. and Hubbard,J.P. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
Cannata,N., Toppo,S., Romualdi,C. and Valle,G. (2002) *Bioinformatics*, **18**, 1102–1108.
Cieplak,M., Holter,N.S., Maritan,A. and Banavar,J.R. (2001) *J. Chem. Phys.*, **114**, 1420–1423.
Chan,H.S. (1999) *Nat. Struct. Biol.*, **6**, 994–996.
Chan,H.S. and Dill,K.A. (1989) *Macromolecules*, **22**, 4559–4573.
Davidson,A.R., Lumb,K.J. and Sauer,R.T. (1995) *Nat. Struct. Biol.*, **2**, 856–863.
Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
Jonson,P.H. and Petersen,S.B. (2001) *Protein Eng.*, **14**, 397–402.
Kamtekear,S., Schiffer,J.M., Xiong,H., Babik,J.M. and Hecht,M.H. (1993) *Science*, **265**, 1680–1685.
Lau,K.F. and Dill,K.A. (1989) *Macromolecules*, **22**, 3986–3997.
Mathews,C.K. and Van Holde,K.E. (1995) *Biochemistry*. Benjamin Cumming, San Francisco, CA.
Metropolis,N., Rosenbluth,A.W., Rosenbluth,M.N., Teller,A.H. and Teller,E. (1953) *J. Chem. Phys.*, **21**, 1087–1092.
Miyata,T., Miyazawa,S. and Yasunaga,T. (1979) *J. Mol. Evol.*, **12**, 219–236.
Miyazawa,S. and Jernigan,R.L. (1996) *J. Mol. Biol.*, **256**, 623–644.
Murphy,L.R., Wallqvist,A. and Levy,R.M. (2000) *Protein Eng.*, **13**, 149–152.
Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
Plaxco,K.W., Riddle,D.S., Grantcharova,V. and Baker,D. (1998) *Curr. Opin. Struct. Biol.*, **8**, 80–85.
Regan,L. and Degrado,W.F. (1988) *Science*, **241**, 976–978.
Reidhaar-olson,J.F. and Sauer,R.T. (1988) *Science*, **241**, 53–57.
Riddle,D.S., Santiago,J.V., Bray,S.T., Doshi,N., Grantcharova,V.P., Yi,Q. and Baker,D. (1997) *Nat. Struct. Biol.*, **4**, 805–809.
Sinha,N. and Nussinov,R. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 3139–3144.
Smith,R.F. and Smith,T.F. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 118–122.
Solis,A.D. and Rackovsky,S. (2000) *Proteins: Struct. Funct. Genet.*, **38**, 149–164.
Wang,J. and Wang,W. (1999) *Nat. Struct. Biol.*, **6**, 1033–1038.
Wang,J. and Wang,W. (2000) *Phys. Rev. E*, **61**, 6981–6986.
Wang,J. and Wang,W. (2002) *Phys. Rev. E*, **65**, 041911_1–041911_5.
Vallon,O. (2000) *Proteins: Struct. Funct. Genet.*, **38**, 95–114.

Received November 20, 2002; revised March 10, 2003; accepted April 4, 2003