Science in China Series C: Life Sciences

Grouping of amino acids and recognition of protein structurally conserved regions by reduced alphabets of amino acids

LI Jing[1] & WANG Wei[1,2†]

[1] National Laboratory of Solid State Microstructure and Department of Physics, Nanjing University, Nanjing 210093, China;
[2] Interdisciplinary Center of Theoretical Studies, Chinese Academy of Sciences, Beijing 100080, China

Sequence alignment is a common method for finding protein structurally conserved/similar regions. However, sequence alignment is often not accurate if sequence identities between to-be-aligned sequences are less than 30%. This is because that for these sequences, different residues may play similar structural roles and they are incorrectly aligned during the sequence alignment using substitution matrix consisting of 20 types of residues. Based on the similarity of physicochemical features, residues can be clustered into a few groups. Using such simplified alphabets, the complexity of protein sequences is reduced and at the same time the key information encoded in the sequences remains. As a result, the accuracy of sequence alignment might be improved if the residues are properly clustered. Here, by using a database of aligned protein structures (DAPS), a new clustering method based on the substitution scores is proposed for the grouping of residues, and substitution matrices of residues at different levels of simplification are constructed. The validity of the reduced alphabets is confirmed by relative entropy analysis. The reduced alphabets are applied to recognition of protein structurally conserved/similar regions by sequence alignment. The results indicate that the accuracy or efficiency of sequence alignment can be improved with the optimal reduced alphabet with *N* around 9.

grouping of amino acids, structural recognition, sequence alignment

Protein sequence alignment is a widely used method for finding structural similarity[1,2]. When the sequence identities are above 30%, the structures of aligned sequences often have very high similarity. Sequence alignment is often used for the recognition of structurally conserved/similar regions. However, this is not workable for some sequences with sequence identities less than 30% since the aligned positions of sequence alignment and structural alignment for such protein pairs do not match well. However, it is well known that lots of these sequences could still share similar structures or conserved regions. For these sequences, the structural similarity may not be detected by the sequence alignment, i.e., could not be identified only by the sequence alignment based on 20 kinds of residues. Is there a way to increase the ability in recognition of structurally conserved/similar regions by sequence alignment for sequences with low identities?

It has been found that some residues are similar in their physicochemical features, and can be clustered into groups because they play similarly structural or functional roles in proteins[3−6]. After clustering, the identities of residues at various sites in the sequence pairs are increased. Then, it is possible that the structural similarity may be detected by sequence alignment with simplified alphabets when 20 kinds of residues are clustered into some groups. This would improve the ability in finding structurally conserved/similar regions and the

Received May 23, 2006; accepted September 19, 2006
doi: 10.1007/s11427-007-0023-3
†Corresponding author (email: wangwei@nju.edu.cn)
Supported by the National Natural Science Foundation of China (Grant Nos. 90403120, 10474041 and 10021001) and the Nonlinear Project (973) of the NSM

structural similarity of entire proteins. Of course, how to group residues will determine the accuracy of sequence alignment.

Previously, most of the substitution matrices used in sequence alignment are derived from databases of protein sequence alignment for sequences with high sequence identities. For example, BLOSUM matrices are based on blocks in BLOCKS database which results from the ungapped segments with similar structures[7]. However, the derived substitution matrices may not well describe the relationship between the residues for sequences with low identities[8,9]. In the Database of Aligned Protein Structures (DAPS), pairs of proteins with similar structures are collected. This database includes various sequences mostly with low identities. Substitution matrices based on such a database may reflect well the relationship among residues and could be used to find structurally conserved regions for proteins with low sequence identities. Actually, this is relevant to that the structural alignment describes the features and relationship of residues better than the sequence alignment does.

In this work, 20 kinds of naturally occurring residues are clustered into N groups based on a substitution matrix built from the DAPS database. Protein sequences with low identities are used, and the clustering of residues is made based on the substitution scores. From various features of clustering, it is found that the simplified representation of protein sequences retains the most important information of the proteins when 20 kinds of residues are clustered into more than 7 groups, i.e., $N \geq 7$. In particular, when N=6–9, the efficiency or accuracy of recognition of structurally conserved/similar regions for sequences with low sequence identities by sequence alignment is also improved. However, when $N \approx 20$ (or N<6), the accuracy of structural recognition is not so good since the sequence complexity cannot be effectively simplified (or too much information in the sequences is lost).

1 Materials and methods

1.1 Database

The database of aligned protein structures (DAPS) is used here for our purpose, and all details can be found from http://www.doe-mbi.ucla.edu/~parag/DAPS/. Briefly, DAPS is a database of structural alignments between proteins which have low sequence identities but similar folds. It was constructed based on the FSSP, DSSP, PDB and CATH databases. By using the set of all proteins from the release of the PDB (2001), an all-against-all 3D structural comparison for these proteins was done to find pairs which have similar structures. Note that one terminal or two termini of some sequence pairs were cut if these domains could not be superposable[10−12].

The sequence pairs in the DAPS database are found to have different sequence identities ranging from about 0 to 100%. In this work, these sequences were classified into 10 subsets S_1, S_2, …, S_{10}, such as the first subset S_1 including sequences with sequence identities S less than 10%, i.e., $S \leq 10\%$, and the second one S_2 with sequence identities S less than 20%, i.e., $S \leq 20\%$, and so on. For a given sequence subset, the numbers of residue pairs at all sites of aligned sequences were calculated, and were used to construct the substitution matrix.

1.2 Grouping of residues based on the substitution matrix

For each sequence subset of the DAPS database, such as the subset S_3 with sequence identities $S \leq 30\%$, 20 kinds of natural occurring residues are grouped into N groups as follows. Similar to the work by Henikoff, an $N \times N$ substitution matrix is obtained by calculating observation frequencies $q_{ij}^{(N)}$, i.e., the number of residue pairs, and expected probability $e_{ij}^{(N)}$ between group i and group j with $1 \leq i, j \leq N$ [7]. Here N is the number of total groups and each group is described by an effective letter $G_i^{(N)}$ which represents the residues in the group, e.g., the effective letter $G_{10}^{(19)}$ represents residues I and V in the 10th group when the total number is N=19. Each element in the matrix describing the substitution score is defined as $s_{ij}^{(N)} = \log_2(q_{ij}^{(N)} / e_{ij}^{(N)})$, which is the logarithmic of odds ratio between the effective letters $G_i^{(N)}$ and $G_j^{(N)}$, and characterizes the substitution frequency of these two letters. Among all the elements, the maximal score $s_{kl}^{(N)}$ of two certain letters $G_k^{(N)}$ and $G_l^{(N)}$, which describes the best substitution between letters, corresponds to that these two letters can be clustered into a new group. This new group is assigned

with a new letter, and two old letters are replaced by the new letter. For this new alphabet, the observation frequencies and expected probabilities are calculated again. Thus a $(N-1)\times(N-1)$ matrix is obtained, and all 20 kinds of residues can be grouped step by step.

1.3 Relative entropy analysis

The relative entropy $H^{(N)}=\sum_{i=1}^{N}\sum_{j=1}^{i}q_{ij}^{(N)}s_{ij}^{(N)}$ is used to measure the averaged information of the matrix with N groups of residues[13, 14]. The larger the value of H, the more information content of the matrix is, and *vice versa*. $H_{\max}^{(N)}=-\sum_{i=1}^{N}P_i^{(N)}\times\log_2 P_i^{(N)}$ gives the maximal information value from the alphabets after clustering. Here, $P_i^{(N)}$ is the frequency of effective letter $G_i^{(N)}$ for group i. Obviously, when N=20, $P_i^{(20)}$ describes the compositions of 20 kinds of residues distributed in nature, and when N<20, $P_i^{(N)}$ means the summation of the compositions of all the residues in the i-th group. Here, $H_{\max}^{(20)}$ is the information of the composition of 20 kinds of natural amino acids distributed in nature.

1.4 Sequence alignment

In this work, the sequence alignment is used for the recognition of structurally conserved/similar regions for protein sequences. For the recognition of structurally conserved/similar regions, the ALIGN program in FASTA package is used to make the global comparison between two sequences. The parameters of gap insertion and elongation are set as −11 and −1. In practice, the sequences for alignment are directly extracted from the DAPS database. In the case of 20 letters, the program works as usual but with the residue substitution matrix obtained by our method, and in the case of reduced alphabets, the reduced sequences and reduced substitution matrices are simultaneously used. For comparison, sequence alignments with other programs, such as BLAST[15] and ClustalX[16], for the recognition are also made for the cases without reduction, namely the case of N=20.

1.5 Structural alignment

The structural alignment is used to find the equivalent sites or similarly structural profiles between proteins by superposition of their structures in a three-dimensional space. In this work all the results of structural alignments are extracted from the DAPS database. Since most of the alignments in the DAPS database were taken directly from the FSSP files based on the DALI algorithm[17], the elastic similarity score, structural similarity threshold 20% and other default parameters were used for the Dalilite program[18].

1.6 Comparison of results based on sequence and structural alignments

For a pair of proteins, if the aligned sites (with the same or not the same kinds of residues) found by sequence alignment match with the aligned (or equivalent) sites found by structural alignment, the sequence alignment can recognize the structurally equivalent sites[19]. Here, the equivalent sites found by structural alignment are taken as reference sites, and the aligned sites found by sequence alignment are matched to these reference sites to test the accuracy of sequence alignment. The fractions of correctly aligned sites in the total number of structural equivalent sites C_R=N_{correct} /N_{str} and in the total number of sequence aligned sites D_R=N_{correct} /N_{seq} are used to measure the match between two types of alignments. Here N_{correct} is the number of matched sites between the sequence alignment and the structural alignment, N_{seq} is the total number of aligned sites in sequence alignment, and N_{str} is the total number of structurally equivalent sites in structural alignment. Clearly, C_R and D_R measure the sensitivity and selectivity of sequence alignment, respectively. Therefore, the assessment of accuracy of the sequence alignment is performed by both C_R and D_R simultaneously.

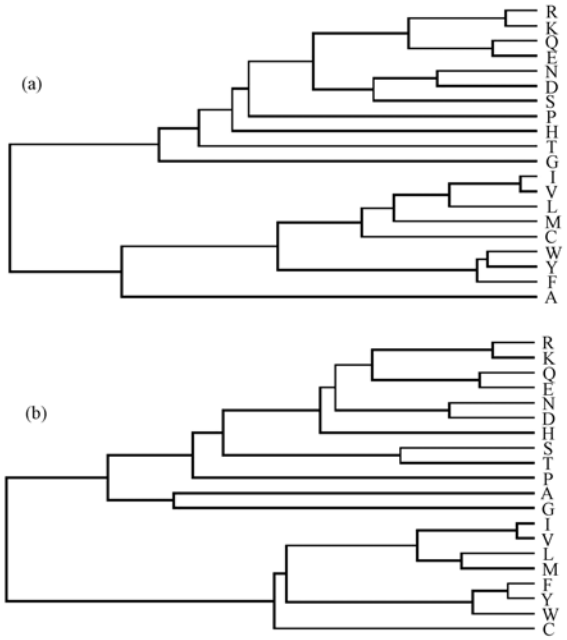
1.7 Principal components analysis

The substitution matrix of 20 kinds of natural occurring residues can be relevant to 20 vectors in a 20-dimensional space. A Euclidean distance matrix D between these 20 vectors can be obtained by $M_{ij}=(\sum_{k=1}^{20}(s_{ik}-s_{jk})^2)^{1/2}$. The cross-correlation matrix $R=MM^T$ is solved and the eigenvectors of two largest principal eigenvalues are projected onto a two-dimensional plot, which gives the main features among 20 kinds of residues[20, 21].

2 Results

2.1 Clustering of residues based on the DAPS database

Twenty kinds of residues are clustered into different groups with different levels of simplification based on our method. In order to analyze the validity of different clustering from various sequences with different sequence identities, 10 sequence subsets from the DAPS database are used. Figure 1 shows some examples of the tree-like clustering for two subsets S_3 and S_5. All 20 letters are clustered into the hydrophobic and polar divisions naturally at different levels N. Some similar features of residues are in common for two trees, suggesting a stable substitution relationship among these letters despite of the usage of different subsets of sequences. These results are basically similar to those obtained previously[22].



Figute 1 The tree-like distribution of residues using hierarchical clustering method directly based on substitution scores for sequence subset from DAPS database with sequence identities (a) $S \leq 30\%$ (subset S_3), (b) $S \leq 50\%$ (subset S_5). The letters denote the 20 types of naturally occurring residues, and the distance between groupings in tree describes the information loss of two groupings in information theory. Note that the horizontal distance in the tree between N = 20 and N = 2 is $L^{(20,1)} = D^{(20)} - D^{(1)}$. Here $D^{(1)} = 0$ for all the residues being clustered into one group is set as the root of the tree.

For the tree-like clustering, the order of the clustering is characterized by the degree of order of information $D^{(N)} = H_{\max}^{(N)} - H^{(N)}$. Here $D^{(1)} = 0$, i.e., for all the amino acids being clustered into one group, can be set as the root of the tree. A distance between groupings N and N', i.e., $L^{(N',N)} = D^{(N')} - D^{(N)}$ describes the information loss of two groupings. That is, the shorter the distance, the less the information loss between these two groupings. The horizontal spacing between different groupings in Figure 1 reflects such information loss between groupings.

For each sequence subset including high sequence identities, say the subset S_5, the relative entropy $H^{(N)}$ has a plateau when $N \geq 9$, and then decreases as the clustering level N decreases (see Figure 2 (a)). However, for the sequence subset S_3, the relative entropy fluctuates around $H^{(N)} \approx 0.1$. Note that the values of $H^{(20)}$ for different sequence subsets are different. The existence of the plateaus implies that the clustering retains the relationship of substitution between residues and results in minor information loss with respect to the case of 20 letters. Thus the effects of substitution between similar residues are well extracted for sequence subsets with various sequence identities. These can be seen more clearly from the plateau in the correlation between the relative entropy and the degree of information content for clustering (Figure 2(b)). Thus, using about 9 letters for the reduced sequences is enough to retain most of the information of original sequences with 20 kinds of residues. This feature is relevant to our previous results[23]. In addition, when the sequence identities are low, say $S \leq 30\%$, the difference between the relative entropy of reduced alphabet and that of the naturally occurring alphabet is very small.

2.2 Recognition of structurally conserved regions by sequence alignments

In general, when the sequence identities of a protein pair are low, say $S<30\%$, the aligned sites found by the sequence alignment and the equivalent sites found by the structural alignment are not consistent with each other mostly. However, as shown in previous studies, many proteins with low sequence identities still have similar structures. In our opinion, this depends on the identities of the residues. If the 20 letters are befittingly grouped based on their physicochemical features, the identities of sequences are enhanced and the differences between similar residues are decreased. Then more aligned sites could be found, enhancing the match with the structurally equivalent sites. Note that an over-reduced grouping

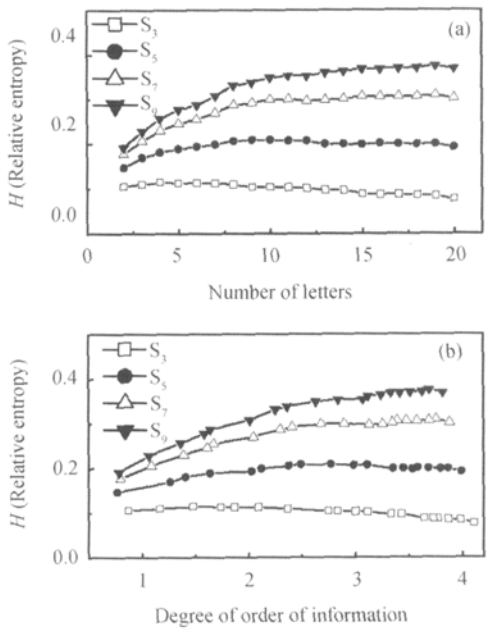


Figure 2 The relative entropy H versus the number of letters N in reduced alphabet (a), and the degree of order of information (b). Different sequence subsets from the DAPS database with different sequence identities $S \leq 30\%$ (subset S_3), $\leq 50\%$ (subset S_5), $\leq 70\%$ (subset S_7) and $\leq 90\%$ (subset S_9) are used.

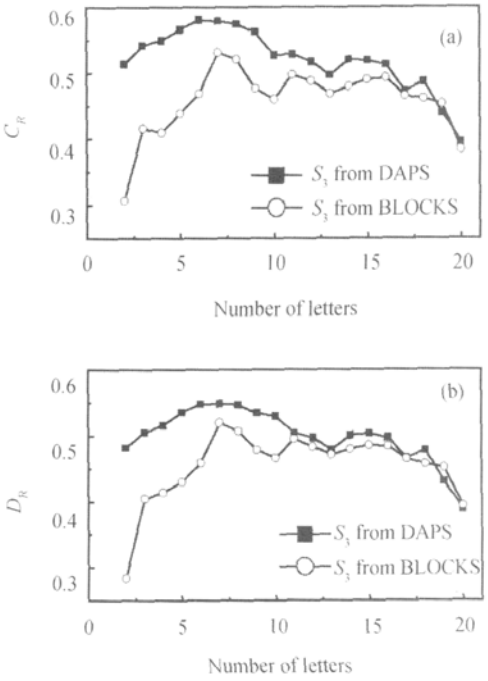


Figure 3 The sensitivity C_R and the selectivity D_R of the sequence alignment versus the number of letters N in reduced alphabet for the sequence subsets S_3 with sequence identities $S \leq 30\%$ from the DAPS database and the BLOCK database.

of residues will result in the loss of the differences between the residues. Thus the grouping of residues should be optimal.

Figure 3 shows the values of sensitivity C_R and selectivity D_R of the sequence alignment versus the number of groupings N for 200 protein pairs with sequence identities less than 30% taken randomly from the DAPS database. The reduced sequences are obtained by substituting 20 letters with related N letters. It is seen that both C_R and D_R have a maximum around N=6—9. Thus, a suitable clustering of residues can help to recognize the structural conservation/similarity by sequence alignment.

An example of increase in the ability to recognize two proteins (PDB codes: 1B8X_A and 2GSR_A) is shown in Figures 4—9. Figure 4 shows the three-dimensional structures of these two aligned proteins extracted from the DAPS database using the program MolMol[24]. It is seen that these two structures are very much similar since most parts are well superposed and have similar conformations and locations. However, there are five regions which have no corresponding local conformations, which are gaps in the structural alignment and marked in boxes (see Figure 4). Figure 5 shows the result of structural alignment where the structurally equivalent sites are connected with symbol "|" and five gaps are marked in boxes. These structural equivalent sites are taken as reference sites for the sequence alignment. The related results of sequence alignment using N=20, 8 and 2 letters are shown in Figures 6—8, respectively. From Figure 6—8, it is seen clearly that the number of aligned sites increases as the number of letters N decreases. Meanwhile, the number and lengths of gap boxes decrease. If the positions and lengths of the gap boxes resulting from the sequence alignment correspond to those from the structural alignment, the recognition by the sequence alignment matches well with the structurally equivalent sites. The pattern in Figure 7 for N =8 shows the best match, and the last three boxes, out of five, have similar or same positions and lengths with respect to those in Figure 5, indicating a good match between two types of alignments. Note that the first two boxes only have the same lengths with respect to those in Figure 5. This means that these two boxes are incorrectly placed. However, the effect of such incorrect boxes is small since the distance between them is small and the location of them is still related to that in Figure 5. Clearly when N =8 almost all the aligned sites in the

sequence alignment match well to those in the structural alignment, indicating the most accurate sequence alignment when N =8. Furthermore, with too many or too less letters, i.e., the case N=20 or N=2 shown in Figure 6 or Figure 8, the match between the sequence alignment and structural alignment is not as good. Thus, too much detailed clustering with large N and too coarseness clustering with small N work badly. Clearly, an optimal grouping around N=8 is obtained. This is consistent with the measure shown in Figure 3.

2.3 Structural recognition using other programs of sequence alignment

The sequence alignments shown in Figures 6—8 are worked out using FASTA program with related substitution matrices based on our clustering method. To make a comparison, the results of structural recognition characterized by the values of sensitivity and selectivity using three popular programs, namely FASTA, BLAST and ClustalX with the BLOSUM62 substitution matrix, are listed in Table 1. Note that the same 200 protein pairs are used. The parameters of gap insertion and elongation are all set as −11 and −1, and other filter options are

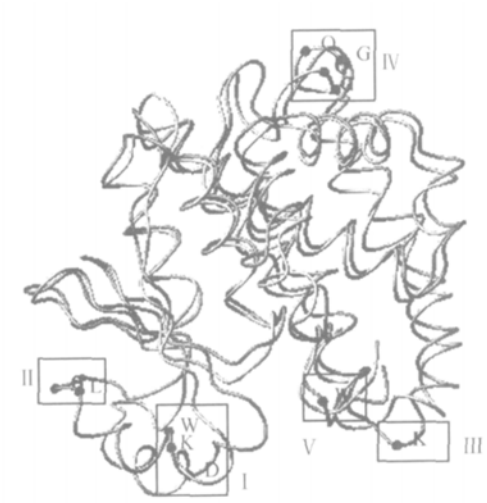


Figure 4 The example of the structural alignment and the sequence alignment between two proteins (PDB codes are 1B8X_A and 2GSR_A). The three-dimensional plot of the two protein based on structural alignment. The no-aligned positions in the proteins are marked with boxes and labeled by the residue names.

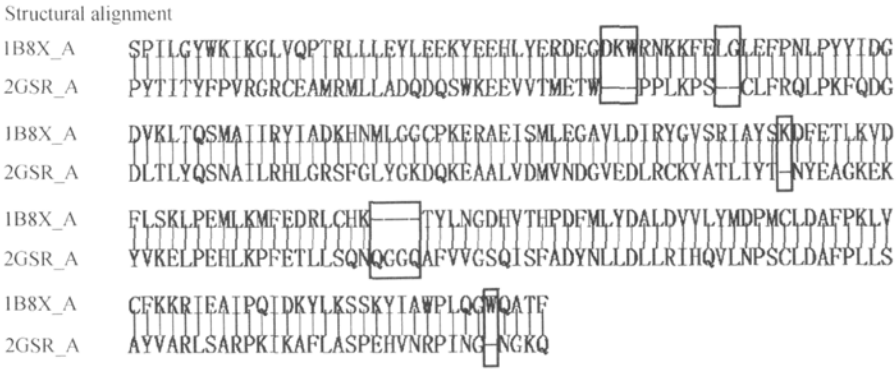


Figure 5 Structural alignment. It is extract from the DAPS database.

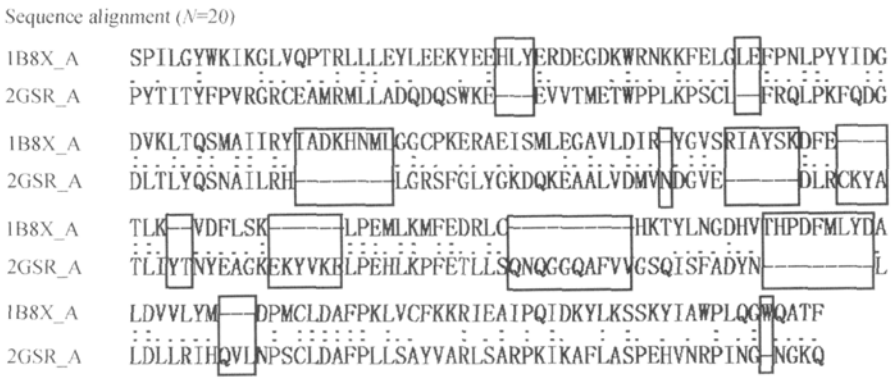


Figure 6 The sequence alignment using the program of FASTA. Twenty letters represent the 20 kinds of natural occurring residues.

Figure 7 The sequence alignment using the program of FASTA. Eight letters from "a" to "h" are used for 8 kinds of effective "residues".

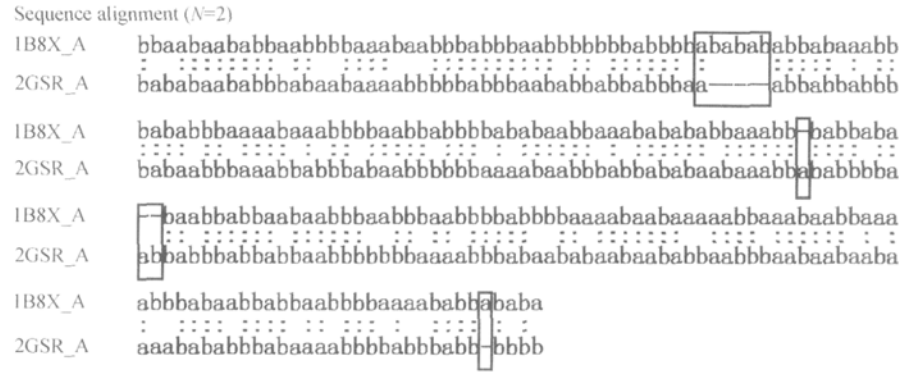


Figure 8 The sequence alignment using the program of FASTA. Two letters "a" and "b" are used for 2 kinds of effective "residues".

closed. Since it is difficult to make the alignment for cases of $N \neq 20$ directly using BLAST and ClustalX, the results are only for N=20. However, the results using FASTA with the substitution matrices from S_3 for cases of N=20 and 8 are also listed in Table 1 for reference. From Table 1, one can see that the alignment using FASTA with the substitution matrix from S_3 for N=8 is generally superior to that using other programs in the structural recognition for the sequence pairs with low sequence identities. The sequence alignment using ClustalX also obtains a high value of sensitivity and selectivity. An example of such sequence alignment for protein 1B8X_A and 2GSR_A is shown in Figure 9. From this figure, one can see that the alignment using ClustalX is not as good as that shown in Figure 7 using FASTA for the case of N=8. Thus, sequence alignments based on proper simplification carry out some important information encoded in the sequences. However, it is worthy to note that these three programs have different implementations when applied in the sequence alignment.

2.4 Reduced alphabets based on BLOCKS database

Many previous work about reduced alphabets from substitution matrix used the BLOSUM substitution matrix which was derived based on BLOCKS database[25, 26] constructed by aligned segments of highly conserved regions of protein family[27−29]. To check the difference between results from the DAPS database and the BLOCKS database, the groupings of residues for the BLOCKS database are worked out following the same method as used for the DAPS database, and then related sequence alignments for these groupings are made.

Figure 10(a) shows the tree-like grouping for the BLOCKS database. It is seen that all residues are also clustered into two main groups, i.e., the hydrophobic group and the polar group. The main feature of this grouping is similar to that in Figure 1 for the DAPS database, however, some detailed distribution of residues are slightly different.

In order to further interpret the complicated relationship among 20 kinds of residues, principal component

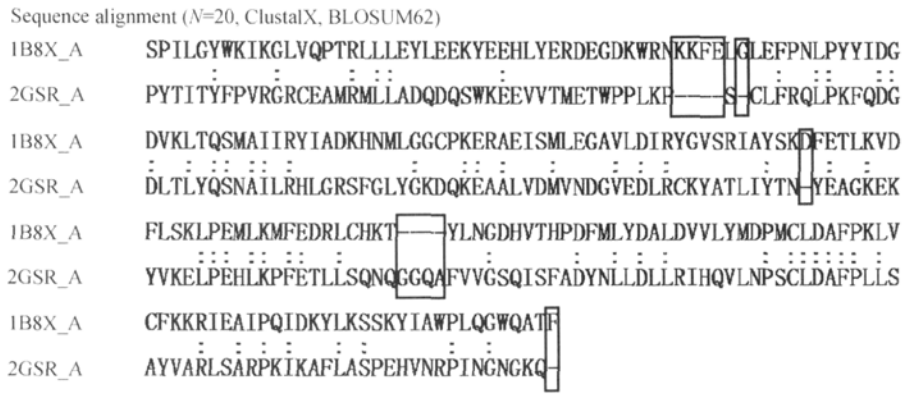


Figure 9 The sequence alignment using the program of FASTA. The pattern of the sequence alignment using the program of ClustalX with BLOSUM62 matrix.

Table 1 Structural recognition characterized by the sensitivity C_R and selectivity D_R based on sequence alignments for 200 protein pairs randomly taken from the DAPS database using different programs and matrices

Program	FASTA	FASTA	FASTA	BLAST	ClustalX
Matrix	reduced 8×8 matrix for S_3 from DAPS	20×20 matrix for S_3 from DAPS	BLOSUM62	BLOSUM62	BLOSUM62
Number of letters (N)	8	20	20	20	20
C_R	0.58	0.38	0.41	0.36	0.52
D_R	0.55	0.39	0.42	0.34	0.49

analysis is used to map the information of these residues from a 20-dimensional space to a 2-dimensional space. Figure 10(b) and (c) show the results of the principal component analysis of the matrices derived from both the BLOCKS and DAPS databases when sequence identities are lower than 30%, respectively. Obviously, the hydrophobic and polar feature can be seen more clearly for residues from the DAPS database than that from the BLOCKS database in a 2-dimensional space. This further indicates that the DAPS database better reflects the relationship of the residues in sequences with low sequence identity. This is because there are more sequences sharing the similarly structural profiles with low sequence identities in the DAPS database than those in the BLOCKS database.

Importantly, as shown in Figure 3, the values of the sensitivity and the selectivity of alignment based on matrices from the BLOCKS database are lower than those from the DAPS database. This implies that the substitution matrix from the DAPS database is better than the BLOSUM30 matrix in finding the structurally conserved/similar regions by sequence alignment for the sequences with low identities.

2.5 Reduced alphabets using other grouping methods

Our grouping method based on the substitution scores is one of the hierarchical clustering methods. As described above, this method works well when the sequence identities of sequences in the database are low. It is known that one of the most common approaches in the hierarchical clustering is the unweighted pair-group method using arithmetic average (UPGMA). In such an approach, every residue is mapped as a vector in a 20-dimensional space, and among them, two residues or two vectors with the shortest Euclidean distance are clustered into one group step by step. With this algorithm, we calculate the grouping for sequences with sequence identities lower than 30% (i.e., sequence subset S_3) in the DAPS database using the program of KITSCH in the Phylogenetic Inference Package PHYLIP[30]. From Figure 11(a), the grouping is not reasonable in many detailed aspects of the physicochemical features of residues. This is likely due to the effects of low identities of the sequences in the DAPS database.

Another commonly used approach for residue grouping is the dynamic clustering method. It also takes every residue as a vector with 20 components. When residues are clustered into some groups, the coordinate center of each group and the distance of every residue to the center can be calculated. The main idea of the dynamic clustering method is to obtain the minimized value of the summation of these distances by dynamic program-

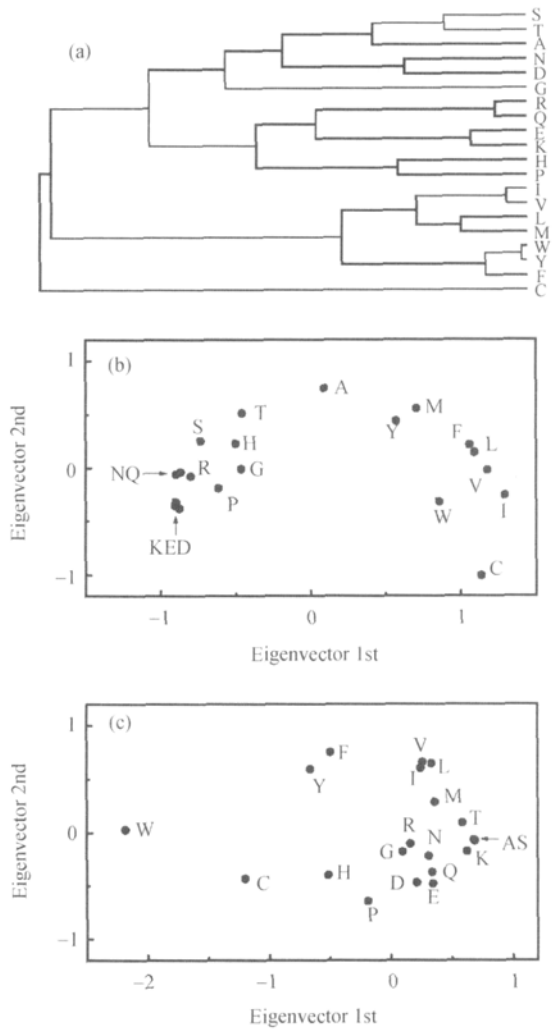


Figure 10 (a) The tree-like distribution of residues using hierarchical clustering method directly based on substitution scores for sequences in BLOCKS database with sequence identities $S \leq 30\%$. The principal component analysis for the substitution matrices for the sequence subset S_3 with sequence identities $S \leq 30\%$ from the DAPS database (b) and for the sequence subset with sequence identities $S \leq 30\%$ from the BLOCKS database (c). The 1st principal eigenvector and 2nd principal eigenvector related to the top two principal eigenvalues from the matrix are plotted. (b) and (c) show the distribution of related eigenvectors for 20 kinds of natural occurring residues in the substitution matrices.

ming algorithm[31]. However, it is worthy noting that the clustering tree cannot be obtained by such method. For the sequences in subset S_3 from the DAPS database, a clustering table based on such a method is also obtained as shown in Figure 11(b). Compared to Figure 1(a), this grouping is also not reasonable due to the effects of low identities of the sequences.

Thus, these two methods mentioned above could not be applied to the clustering of residues for datasets with low sequence identities. This is due to the fact that both methods are based on the calculation of the Euclidean distances between the residues that are taken as vectors, and the differences of these distances are too small to distinguish the residues as also implied in Figure 11(b). This is really the case when the sequence identities are low, especially for those sequences in subset S_3 from the DAPS database with sequence identities lower than 30%. More detailed comparison and discussion for these methods will be presented elsewhere.

3 Discussion

Recognition of protein structurally conserved/similar regions by sequence alignment is a useful approach to analyze the relationship between protein structures and sequences. It is well known that the most important application is to predict the three-dimensional structures of proteins by homology modeling. However, this method could not work well for those sequences with low sequence identities, say lower than 30%. Can we improve the ability in the recognition of structurally conserved/similar regions by sequence alignment for these sequences? In this work, we have proposed a way to solve this problem.

Previous work showed that the complexity of protein sequences could be simplified using less than 20 types of natural occurring residues without losing the information of the sequence and structural features. In the simplification of protein complexity by residue grouping based on protein sequences, choosing a suitable database is very important for characterizing the relationship of residue substitution. The protein structural alignment database DAPS chosen in this work reflects well the relationship of residues for protein sequences with low identities. Because of low identities of sequences in this database, the relative entropy H analysis showed that the relative information between residues was weak. The clustering methods based on the geometric distances between residues could not well reflect the detailed differences. Thus, a grouping method based on the statistics of residue substitution was proposed in this work, and a hierarchical grouping tree of residues was obtained. Comparing to previous work, the distribution of residues in various groups is similar, e.g., all residues are clustered into the hydrophobic residues and the polar groups. Such a detailed distribution of residues corresponds to the physicochemical features of the residues. This

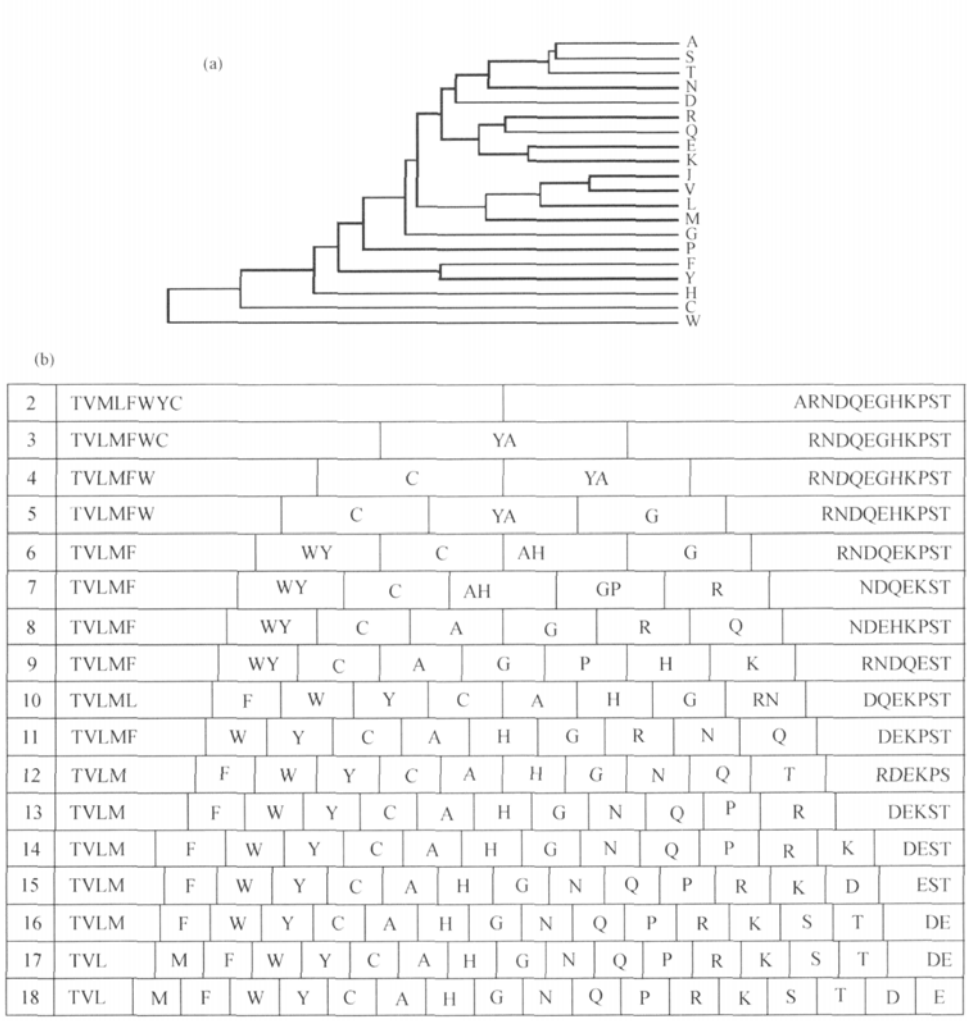


Figure 11 (a) The tree-like distribution of residues using the UPGMA clustering method for sequence subset S_3 with sequence identities $S \leq 30\%$ from the DAPS database. Note that the clustering is based on BLOSUM30 matrix using the program of KITSCH in PHYLIP. (b) The clustering of residues using dynamic clustering method for the sequence subset S_3 from the DAPS database. Note that the tree-like distribution cannot be obtained.

method can be used to find the relative information between the residues.

Reduced alphabets and their related substitution matrices are applied to recognition of structurally conserved/similar regions. For the recognition of structurally conserved/similar regions using the sequence alignment based on the reduced alphabets, it is found that when N=6—9 the accuracy of the recognition is improved. This shows that the accuracy of sequence alignment can be improved if an optimal reduced alphabet is used, i.e., the ability of recognition of structurally conserved/similar regions can be improved by optimally grouping residues.

Finally, it is worthy to note that the available information needed for estimating substitution matrices based on structural alignments might be not superior to the information available based on sequence alignments. This depends on the database used. If the database, such as the DAPS database, used is good enough, we do have more information than we have from the database based on the sequence alignments. This is confirmed by the comparison between the results from the DAPS database and from the BLOCKS database. Thus, the substitution matrices based on the DAPS database well reflect the relationship between the residues. It is also worthy to note that the improvement of accuracy or efficiency of sequence alignment depends on the reduction of the residue alphabets or their related substitution matrices

rather than on the alignment algorithms. The accuracy or efficiency can be improved further if better alignment algorithms are used.

We thank Y.Q. Zhou for his help with this manuscript.

1 Bowie J U, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science, 1991, 253: 164—170
2 Jones D T, Taylor W R, Thornton J M. A new approach to protein fold recognition. Nature, 1992, 358: 86—89
3 Regan L, Degrado W F. Characterization of a helical protein designed from first principles. Science, 1988, 241: 976—978
4 Kamtekar S. Protein design by binary patterning of polar and nopolar amino acids. Science, 1993, 262: 1680—1685
5 Plaxco K W. Simplified proteins: Minimalist solutions to the "protein folding problem". Curr Opin Struct Biol, 1998, 8: 80—85
6 Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nature Struct Biol, 1999, 6: 1033—1038
7 Henikoff S, Henikoff J G. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA, 1992, 89: 10915—10919
8 Ogata K, Ohya M, Umeyama H. Amino acid similarity matrix for homology derived from structural alignment and optimized by the Monte Carlo method. J Mol Graph Model, 1998, 16: 178—189
9 Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins, 2005, 58: 321—328
10 Friedberg I, Kaplan T, Margalit H. Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. Protein Sci, 2000, 9: 2278—2284
11 Mallick P, Weiss R, Eisenberg D. The directional atomic solvation energy: An atombased potential for the assignment of protein sequences to known folds. Proc Natl Acad Sci USA, 2002, 99: 16041—16046
12 Kleiger G. PFIT and PFRIT: Bioinformatic algorithms for detecting glycosidase function from structure and sequence. Protein Sci, 2004, 13: 221—229
13 Karlin S, Altschul S F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA, 1990, 87: 2264—2268
14 Altschul S F. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol, 1991, 219: 555—565
15 Karlin S, Altschul S F. Applications and statistics for multiple high-scoring segments in molecular sequences. Proc Natl Acad Sci USA, 1993, 90: 5873—5877
16 Higgins D G, Sharp P M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene, 1988, 73: 237—244
17 Holm L, Sander C. Mapping the protein universe. Science, 1996, 273: 595—602
18 Holm L, Sander C. Dictionary of recurrent domains in protein structures. Proteins, 1998, 33: 88—96
19 Blake J D, Cohen F E. Pairwise sequence alignment below the twilight zone. J Mol Biol, 2001, 307: 721—735
20 Dosztanyi Z, Torda A E. Amino acid identity matrices based on force fields. Bioinformatics, 2001, 17: 686—699
21 Johnson M S, Overington J P. A structural basis for sequence comparisons an evaluation of scoring methodologies. J Mol Biol, 1993, 233: 716—738
22 Li T. Reduction of protein sequence complexity by residue grouping Protein Eng, 2003, 16: 323—330
23 Fan K, Wang W. What is the minimum number of letters required to fold a protein. J Mol Biol, 2003, 328: 921—926
24 Koradi R, Billeter M, Whrich K. MOLMOL: A program for display and analysis of macromolecular structures. J Mol Graphics, 1996, 14: 51—55
25 Henikoff S. Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene, 1995, 163: GC17—GC26
26 Pietrokovski S, Henikoff J G, Henikoff S. The blocks database–A system for protein classification. Nucleic Acids Res, 1996, 24: 197—200
27 Clarke N D. Sequence "minimization": Exploring the sequence landscape with simplified sequences. Curr Opin Biotech, 1995, 6: 467—472
28 Riddle D S. Functional rapidly folding proteins from simplified amino acid sequences. Nature Struct Biol, 1997, 4: 805—809
29 Akanuma S, Kigawa T, Yokoyama S. Combinatorial mutagenesis to restricted amino acid usage in an enzyme to a reduced set. Proc Natl Acad Sci USA, 2002, 99: 13549—13553
30 Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. Evolution, 1985, 39: 783—791
31 Liu X. Simplified amino acid alphabets based on deviation of conditional probability from random background. Phys Rev E, 2002, 66: 021906-1—021906-4