Protein Structure Comparison Using the Markov Transition Model of Evolution

Takeshi Kawabata and Ken Nishikawa*
Center for Information Biology, National Institute of Genetics, Yata, Mishima, Shizuoka, Japan

ABSTRACT **A number of automatic protein structure comparison methods have been proposed; however, their similarity score functions are often decided by the researchers' intuition and trial-and-error, and not by theoretical background. We propose a novel theory to evaluate protein structure similarity, which is based on the Markov transition model of evolution. Our similarity score between structures i and j is defined as log $P(j \rightarrow i)/P(i)$, where $P(j \rightarrow i)$ is the probability that structure j changes to structure i during the evolutionary process, and $P(i)$ is the probability that structure i appears by chance. This is a reasonable definition of structure similarity, especially for finding evolutionarily related (homologous) similarity. The probability $P(j \rightarrow i)$ is estimated by the Markov transition model, which is similar to the Dayhoff's substitution model between amino acids. To estimate the parameters of the model, homologous protein structure pairs are collected using sequence similarity, and the numbers of structure transitions within the pairs are counted. Next these numbers are transformed to a transition probability matrix of the Markov transition. Transition probabilities for longer time are obtained by multiplying the probability matrix by itself several times. In this study, we generated three types of structure similarity scores: an environment score, a residue–residue distance score, and a secondary structure elements (SSE) score. Using these scores, we developed the structure comparison program, Matras (MArkovian TRAnsition of protein Structure). It employs a hierarchical alignment algorithm, in which a rough alignment is first obtained by SSEs, and then is improved with more detailed functions. We attempted an all-versus-all comparison of the SCOP database, and evaluated its ability to recognize a superfamily relationship, which was manually assigned to be homologous in the SCOP database. A comparison with the FSSP database shows that our program can recognize more homologous similarity than FSSP. We also discuss the reliability of our method, by studying the disagreement between structural classifications by Matras and SCOP. Proteins 2000;41:108–122.**

Key words: Dayhoff substitution model; similarity score function; superfamily; fold; homology; analogy; SCOP; FSSP

INTRODUCTION

Recently, the number of entries in the Protein Data Bank surpassed 10,000, and it still increases by about 30–50 entries per week. This wealth of data has encouraged comparisons and classifications of known protein tertiary structures. Several motivations for structural comparisons exist: evolutionary biologists are interested in detecting a distant evolutionary relationship from structure resemblance,[1,2] physicists want to find general rules of protein architecture by studying common structure patterns,[3] structure genomists try to estimate the functions of hypothetical protein structure they have solved,[4] and structure predictors need to judge the accuracy of their predicted structures.[5] For all these purposes, the aid of automatic structure comparison is necessary to cope with the numerous and complicated structure data. Actually, several different programs have already been developed (for reviews, see Holm and Sander[6] and Orengo[7]).

In general, two problems must be solved to compare structure pairs automatically. The first one is to define similarity score functions between structure pairs. The distance between two residues and the vectors specifying the orientations of secondary structure elements (i.e., α-helix and β-strand, called SSEs) are often employed as the structural features, but there are various functions to quantify their differences. To take the case of the distance between two residues as an example, Taylor and Orengo[8,9] proposed a similarity score, $a/(D + b)$, where D is the difference between the two distances (or vectors), and a and b are arbitrarily defined constant values. Holm and Sander[10–13] defined a similarity score as $(a - D/\bar{D})\exp[-(\bar{D}/b)^2]$, where $\bar{D}$ is the averaged distance. Rossman and Argos[14] and Russel and Bartons[15] used a score, $\exp[-(D/a)^2]\exp[-S/b)^2]$, where S is the difference between neighboring residue distances. Gerstein and Levitt[16] used a different score, defined as $a/(1 + (D/b)^2)$. Other researchers defined their own functions.[17–20] In the case of SSE comparison, different similarity scores have also been proposed.[21–24] The second problem is to develop an efficient algorithm that finds the structural corresponding residues (alignment) with the largest similarity score. Although it has been proved that there is no perfect method to solve this problem,[25] many heuristic algorithms has been proposed so far, such as the double dynamic

*Correspondence to: Ken Nishikawa, Center for Information Biology, National Institute of Genetics, 1111, Yata, Mishima, Shizuoka 411, Japan. E-mail: knishika@genes.nig.ac.jp

Received 15 February 2000; Accepted 2 June 2000

programming method,[8,9,20] the Monte Carlo method,[10] the iterated dynamic programming method,[14–17] the path-extension method,[18,19] and the hierarchical alignment method.[9,11,12,23,24]

In the general scheme described here, we have noticed that the similarity score for comparing structural features has been differently formulated by various investigators, suggesting that the score function has been defined rather arbitrarily. Our idea originated at this point, and we wanted to define a similarity score on a more rational ground. We proposed a novel definition for structure similarity, which is similar to the amino acid substitution score.[26,27] Our similarity score between structures i and j is defined as follows:

$$S(i,j) = \log \frac{P(j \rightarrow i)}{P(i)} \tag{1}$$

where $P(j \rightarrow i)$ is the probability that structure j changes to structure i during the evolutionary process, and $P(i)$ is the probability that structure i appears by chance. This is a rational definition of structure similarity, especially for finding evolutionarily related (homologous) similarity. Although estimating the probability $P(j \rightarrow i)$ is a difficult task, an approximate value can be obtained using the Markov transition model, which is similar to Dayhoff's model of amino acid substitution.[26] A basic assumption of the model is that the structure of a protein gradually changes in the evolutionary process, as its sequence changes by amino acid substitutions, insertions, and deletions. Transition probabilities of an elementary step are estimated by closely related protein pairs, whose structures can be easily compared according to the sequence alignment. Actually, transition probabilities are designed for changes between structural features, such as secondary structure states of residues and distance between residues. Transition probabilities between distantly related structures are provided by repeatedly multiplying the probability matrix for an elementary step. Then, numerical values of the similarity score are obtained from Eq. (1). This model is a reasonable approximation at least for the "conserved core" region, although some regions change drastically in their long evolutionary history.[28]

The computer program developed in this study is called Matras (MArkovian TRAnsition of protein Structure), which automatically compares any two protein structures. The program Matras uses three kinds of similarity score functions derived by the Markov model, and employs a hierarchical alignment algorithm, in which a rough alignment is first obtained by SSEs, and then is improved with more detailed similarity functions. The Matras program was applied to all-versus-all comparisons of the protein structures stored in the PDB, and similar structures detected by Matras were compared with the SCOP database.[29] In the SCOP database, structural similarities were deliberately classified into the superfamily or fold level. The former assumes the evolutionary relationship (i.e., homology), but the latter does not (i.e., analogy). In view of our basic assumption, we expect that Matras may be good at detecting structural similarities corresponding to the superfamily (homology), although several researchers reported that distinguishing homologous and analogous relationship is difficult for automatic methods.[30–33] Inconsistencies between Matras and SCOP will be discussed.

METHODS

Dataset

To derive the parameters of our similarity score function, 3,261 representative proteins were selected from the Protein Data Bank (PDB) on April 14, 1999, with mutual sequence identity less than 95%. Small proteins with length less than 40 residues were removed from the database. The clustering by 40% sequence identity was then made against the representative, and maximum similarity trees were made for each cluster.

The performance of the method was evaluated using the SCOP database,[29] version 1.39. The dataset included protein domains with sequence identity of 30% or less, but excluded those in Class 6 of SCOP ("Membrane and cell surface protein") and of small domains with less than 40 residues. Domains composed of separated regions or different chains were also removed. As a result, 1,487 representative structures were obtained as a dataset.

Markovian Transition Model of Evolution

In this section, we briefly explain the theory of the Markov transition model and procedures for making our similarity score. An outline of the procedures is schematically described in Figure 1, which illustrates a simple case of the three-state secondary structure as structural features. In the first stage of the procedures, closely related proteins are grouped into the maximum similarity tree (Fig. 1a). This tree is assumed to be an approximation of the phylogenetic tree.[34] All of the connected protein pairs are aligned using the sequence similarity. Then, the number of structure transitions A_{ij} are counted for all of the connected protein pairs, where i and j are two structural features, such as α-helix and β-strand in the secondary structure, or different bins in the C^{β}-C^{β} distance histogram. The theory does not depend on which structure feature is used. To avoid the artificial population bias in the dataset, the numbers counted are normalized with the weight $1/N_c$, where N_c is the number of proteins in each cluster of the maximum similarity tree.

The number of structure transitions A_{ij} is transformed to the transition probability M^1_{ij} of an elementary step, that a structural state j changes to a new state i (Fig. 1b). In Dayhoff's model,[26] this transformation is performed by the following equations:

$$M^1_{ij} = \frac{\lambda m_j A_{ij}}{\sum_{k \neq j} A_{kj}} \tag{2}$$

$$M^1_{jj} = 1 - \lambda m_j \tag{3}$$

where λ is an arbitrary parameter and m_j is the mutability of state j, defined as:

$$m_j = \frac{\sum_{k(\neq j)} A_{kj}}{\sum_k A_{kj}}. \tag{4}$$

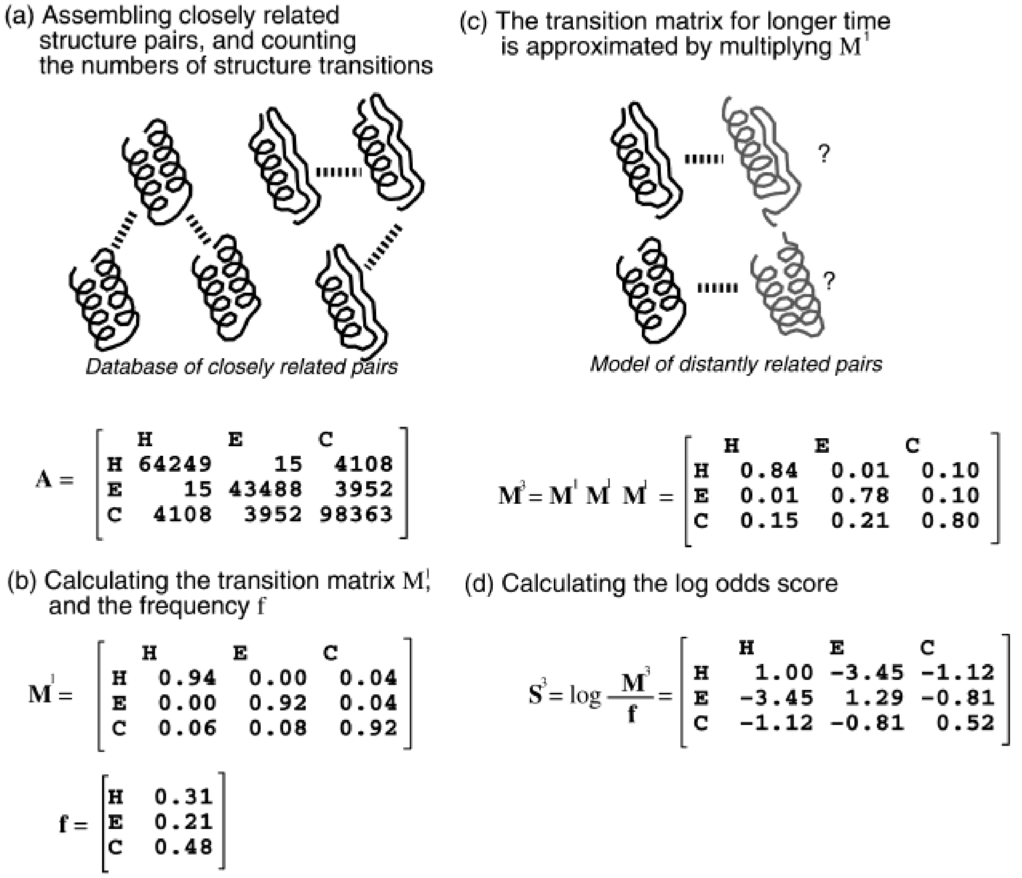

(a) Assembling closely related structure pairs, and counting the numbers of structure transitions
Database of closely related pairs
A =
H E C
H 64249 15 4108
E 15 43488 3952
C 4108 3952 98363
(b) Calculating the transition matrix M1, and the frequency f
M1 =
H E C
H 0.94 0.00 0.04
E 0.00 0.92 0.04
C 0.06 0.08 0.92
f =
H 0.31
E 0.21
C 0.48
(c) The transition matrix for longer time is approximated by multiplyng M1
?
?
Model of distantly related pairs
M3 = M1 M1 M1 =
H E C
H 0.84 0.01 0.10
E 0.01 0.78 0.10
C 0.15 0.21 0.80
(d) Calculating the log odds score
S3 = log M3/f =
H E C
H 1.00 -3.45 -1.12
E -3.45 1.29 -0.81
C -1.12 -0.81 0.52

Fig. 1. Procedures of deriving similarity scores for the three-state secondary structure features. “H,” “E,” and “C” represent α-helix, β-sheet, and coil, respectively.

The relationship between the parameter λ and the ratio of the accepted transition ρ (that is called the percentage of accepted point mutation “PAM” in Dayhoff’s model) is described as the following equation:

$$\rho = 1 - \sum_i f_i M^1_{ii} = \lambda \sum_i f_i m_i \tag{5}$$

where f_i is the frequency of state i:

$$f_i = \frac{\sum_l A_{il}}{\sum_k \sum_l A_{kl}}. \tag{6}$$

If a value of ρ is decided, then the parameter λ is obtained using Eq. (5).

The only difficulty in applying this framework to structure comparison is how to define the arbitrary parameter ρ. Dayhoff used ρ = 0.01 (1 PAM) for her substitution model, but there is no theoretical background supporting 0.01 as the best choice. Because we plan to make several kinds of transition matrices that focus on different structure features (such as secondary structure or distance between residues), the different probabilistic transition models should be modified to the same time scale. Therefore, we use λ = 1.0 for all of the transition matrices. It means that the mutation probability matrix is defined by the following simple equation, instead of Eq. (2) and (3):

$$M^1_{ij} = \frac{A_{ij}}{\sum_k A_{kj}}. \tag{7}$$

The mutation probability matrices for longer time scales are obtained by multiplying the matrix M^1 itself (Fig. 1c). A matrix of N evolutionary steps, M^N, is obtained by $M^N = (M^1)^N$. It is assumed that the protein structure changes gradually in the course of the evolutionary process. Using the probability M^N and f_i, the log-odds similarity score $S(i, j)$ defined in Eq. (1) can be approximated as follows (Fig. 1d):

$$S(i, j) = \log \frac{P(j \to i)}{P(i)} \simeq \log \frac{M^N_{ij}}{f_i}. \tag{8}$$

The number of evolutionary steps N is determined to distinguish the SCOP superfamily relationship with the maximum coverage (see Results).

Practically, for a rarely observed state i, values of Eq. (7) are strongly influenced by small changes in the number of observations. To avoid this problem, we use the following corrected probability M'^N_{ij}, instead of the probability M^N_{ij}:

$$M'^N_{ij} = \left(1 - \frac{m\sigma}{1 + m\sigma}\right) f_i + \frac{m\sigma}{1 + m\sigma} M^N_{ij} \tag{9}$$

$$m = \min[M^N_{ij}, M^N_{ji}] \tag{10}$$

where σ is an arbitrary parameter.[35] In this study, the value of σ is 10000.0 for an environmental score, and is 100.0 for secondary structure element (SSE) and distance scores (see next section).

Three Kinds of Similarity Score

Three kinds of similarity scores: an environment score (S_{env}), a distance score (S_{dis}), and a secondary structure element score (S_{sse}), were devised as follows.

Environment Score S_{env}

The environment score was defined for ten environmental states, which are a combination of local structure and solvent accessibility. There are five kinds of local structures: α-helix ("H"), β-strand ("E"), coil whose φ, ψ angles are in region "G," region "B" and region "L." The α-helix and the β-strand are defined using the DSSP program.[36] The region "G" is $-180 \le \phi \le 0$ and $-120 \le \psi \le 60$, the region "L" is $0 < \phi \le 180$ and $-120 \le \psi \le 120$, and the region "B" is the rest of the φ-ψ plane. There are two kinds of solvent accessibility: exposed and buried, which are defined using the accessibility (ACC) value obtained by the DSSP program. If the value of ACC is larger than 20% of the value for the extended conformation, then its accessibility state is defined as "exposed," otherwise, it is "buried." The ten kinds of "environment" are defined by combining the five local structures and the two accessibility classes. Using Eq. (8), the environment score S_{env} is defined as follows:

$$S_{\text{env}}(E_i, E_j) = \log \frac{M^N(E_i, E_j)}{f(E_i)} \tag{11}$$

where E_i is the ith residue environment of a protein and E_j is the jth residue environment of another protein. The probabilities $M^N(E_i, E_j)$ and $f(E_i)$ are estimated by the procedures described in the previous section. This score is only used for seeking a tentative alignment.

Distance Score S_{dis}

Residue–residue distance is one of the most popular features employed for structure comparison. This score focuses on the distance D_{ij} between the C^{β} atoms of the ith and jth residues. For glycine, the virtual C^{β} atom is generated according to the standard geometry. The distance is transformed into a discrete histogram, because the theory cannot deal with continuous features. The interval of the histogram is set to 1 Å, and the range of distance is $0 < D_{ij} < 50$(Å). The distance score S_{dis} between the distance D_{ix} between the ith and xth residues of a protein, and the distance D_{jy} between the jth and yth residues of another protein, is defined as follows:

$$S^k_{\text{dis}}(D_{ix}, D_{jy}) = \log \frac{M^N_k(D_{ix}, D_{jy})}{f_k(D_{ix})} \tag{12}$$

where k is the number of the residue separation along a chain, defined as $|i - x|$. The score is independently prepared for each residue separation k, which strongly affects the distribution of distance. Short range separations ($1 \le k \le 20$) are treated independently, while long range separations ($k > 20$) are represented by a single matrix. When gapped alignments are performed, $|i - x|$ is not always the same as $|j - y|$. In such cases, the residue separation k is defined as a larger value among them ($k = \max[|i - x|, |j - y|]$). Figure 2 shows the probabilities and the similarity score in the case of the residue separation $k = 5$. This score is used in the final stage of alignment of our program, because it is the most sensitive to detect structural similarity among our three scores. We confirmed that the distance score between the C^{β} atoms produces better alignment than that between the C^{α} atoms, especially for a β-strand region.

Secondary Structure Element Score S_{sse}

The use of secondary structure elements (SSE) is also popular for structure comparison, because of its simplicity.[9,11,12,21–24] A secondary structure element is a continuous residue group that is defined as an α-helix or a β-strand. It is represented by a single vector defined by the principle inertial axis with the smallest moment. The spatial arrangement of a pair of SSEs is described by six parameters: the number of residues L_1, L_2, the closest distance between SSE pair d, the bond angles θ_1, θ_2, and the dihedral angle φ (see Fig. 3), and four kinds of scores S^{θ}_{sse}, S^{ϕ}_{sse}, S^{d}_{sse}, and S^{L}_{sse} are introduced. These similarity functions between the ith and xth SSE of a protein, and the jth and yth SSE of another protein, are defined as follows:

$$S^{\theta}_{\text{sse}}(\theta_{ix}, \theta_{jy}) = \log \frac{M^N(\theta_{ix}, \theta_{jy})}{f(\theta_{ix})} \tag{13}$$

$$S^{\phi}_{\text{sse}}(\phi_{ix}, \phi_{jy}) = \log \frac{M^N(\phi_{ix}, \phi_{jy})}{f(\phi_{ix})} \tag{14}$$

$$S^{d}_{\text{sse}}(d_{ix}, d_{jy}) = \log \frac{M^N(d_{ix}, d_{jy})}{f(d_{ix})} \tag{15}$$

$$S^{L}_{\text{sse}}(L_i, L_j) = \log \frac{M^N(L_i, L_j)}{f(L_i)}. \tag{16}$$

The bond angle θ, the dihedral angle φ, and the distance d are transformed to the histograms. The interval for the angles θ and φ is 10°, that for the distance d is 1 Å. These scores are independently prepared for three combinations of secondary structure types (α-helix vs. α-helix, β-strand vs. β-strand, and α-helix vs. β-strand). The total SSE score is represented as the sum of the following six terms:

$$\begin{aligned} S_{\text{sse}}(i, x, j, y) = {} & S^{L}_{\text{sse}}(L_i, L_j) + S^{L}_{\text{sse}}(L_x, L_y) + S^{\theta}_{\text{sse}}(\theta_{1,ix}, \theta_{1,jy}) \\ & + S^{\theta}_{\text{sse}}(\theta_{2,ix}, \theta_{2,jy}) + S^{d}_{\text{sse}}(d_{ix}, d_{jy}) + S^{\phi}_{\text{sse}}(\phi_{ix}, \phi_{jy}). \end{aligned} \tag{17}$$

This score is used for making tentative SSE alignments in our program.

Alignment Strategy

Because finding the optimal structural alignment between two proteins requires vast computational resources, an efficient heuristic method is required to solve this problem. We use the most popular heuristics, hierarchical

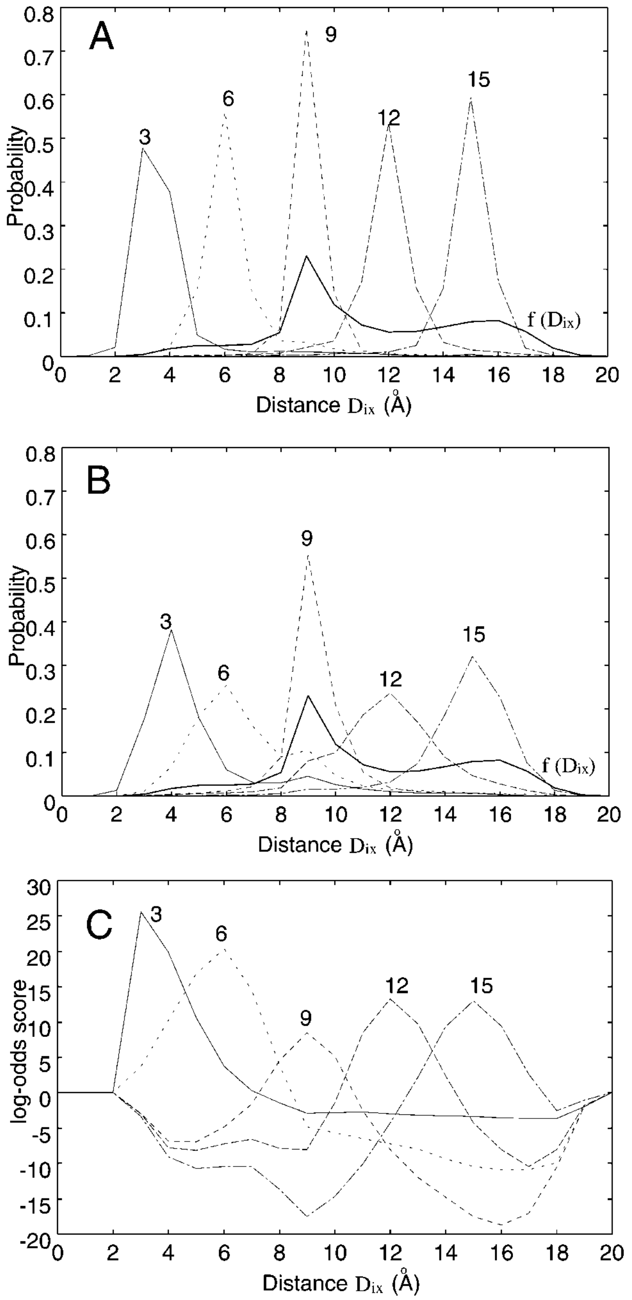


Fig. 2. Transition probabilities and similarity scores for the $C^\beta - C^\beta$ distance between two residues separated by five residues (k = 5) along the chain. (**A**) Transition probabilities $M^1_{k=5}(D_{ix}, D_{jy})$ for evolutionary step N = *1* and observed frequency $f_{k=5}(D_{ix})$. The probabilities for D_{jy} = 3, 6, 9, 12, and 15 Å are shown. A peak at 9 Å of $f_{k=5}(D_{ix})$ corresponds to an α-helix conformation. (**B**) Transition probabilities $M^3_{k=5}(D_{ix}, D_{jy})$ for longer evolutionary steps N = 3. The distribution spreads more than that of N = 1. (**C**) Log-odds similarity scores $S_{k=5}(D_{ix}, D_{jy})$. for N = 3. A peak value for 9 Å is lower than any other distance. This means that an α-helix conformation has little importance in suggesting an evolutionary relationship, because it is a frequently observed conformation in the database.

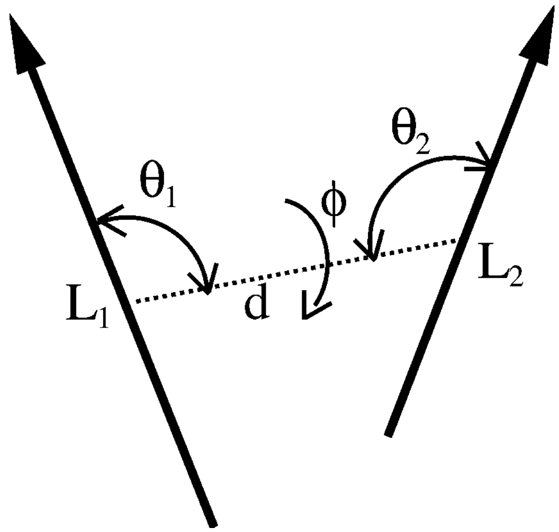


Fig. 3. The six parameters for representing the geometrical arrangement of an SSE pair.

alignment, in which a rough alignment is first obtained by the SSEs, then the alignment is improved with more detailed similarity functions. Many investigators have employed this approach, although the details are different among them.[9,11,12,24] Our procedure of hierarchical alignment consists of the following three stages. First, using the score S_{sse}, the SSEs are aligned using the "build-up" method,[22] which is a kind of branch-and-bound method. The algorithm starts with the SSE pair clusters, and then the list of the score is sorted and the first N_{keep} clusters are taken. In the next step, one SSE is added to the clusters, and then the score list is sorted and the best N_{keep} clusters are taken again. This procedure is repeated until better score clusters do not appear any more. In the second stage, a dynamic programming (DP) alignment using the score S_{env} is performed. We use a local alignment algorithm with affine gap penalty $\alpha k + \beta$, where k is the number of gapped residues.[37,38] A positive score γ is added to the region of previously aligned SSEs. In the third stage, a DP alignment, with the distance score S_{dis}, is iteratively performed using the alignment determined in the previous stage. In the DP alignment, the similarity score of the ith residue of a protein and the jth residue of another protein is defined as follows:

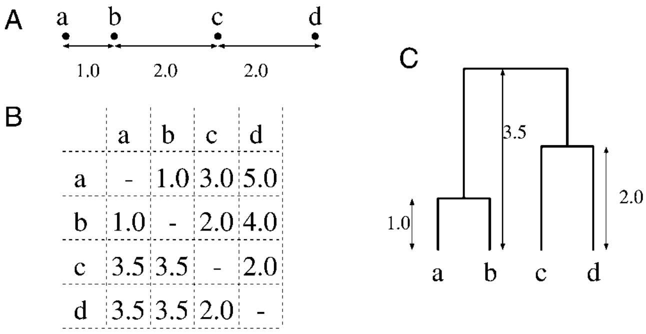
$$S(i,j) = \frac{1}{2}\sum_{a} S^{k}_{\mathrm{dis}}(D_{i,x(a)}, D_{j,y(a)}) \qquad (18)$$

where a residue pair $x(a)$ and $y(a)$ is determined by a previous alignment. This procedure is repeated until convergence or repeating N_{repeat} times. Finally, the similarity is assessed by the distance score.

In this study, the gap penalties α are -6 for the environment score, and -100 for the distance score. The parameter β is defined as 4α, and the offset parameter γ is defined as $-(\alpha + \beta)$. The numbers N_{keep} and N_{repeat} are 30 and 10, respectively.

Score Normalization

The significance of the structure similarity is evaluated by the following Z-score:



	a	b	c	d
a	-	1.0	3.0	5.0
b	1.0	-	2.0	4.0
c	3.5	3.5	-	2.0
d	3.5	3.5	2.0	-

Fig. 4. Example of calculating tree-distance. Supposing that four proteins, *a, b, c,* and *d,* sit in one-dimensional space (**A**), the distances between the four proteins are tabulated as the upper part of the matrix (**B**). The tree shown in (**C**) is generated by the average-distance clustering. Tree-distances, shown in lower part of matrix B, are obtained by summing up lengths of branches. The distances between different cluster proteins, such as $a - c$ and $a - d$, change to averaged values.

$$Z(p, q) = \frac{S(p, q) - E(p)}{\sigma(p)} \qquad (19)$$

where p and q represent proteins, $S(p, q)$ is the similarity score of proteins p and q, and $E(p)$ and $\sigma(p)$ are the average value and the standard deviation of the score of protein p over the database. This normalization is simple among the other normalization schemes proposed previously, considering the length of protein or number of compared residues.[9,13,39] Our distance score S_{dis} correlates with the square of N_{comp}, which is the number of compared residues. We therefore employed a quadratic normalization, in which E_p and $\sigma(p)$ in Eq. (19) are determined by the least-square fitting of the similarity score. The regression line $S_p^{\text{reg}}(p, q) = A_p N_{\text{comp}}^2(p, q) + B_p$ is calculated for the score $S(p, q)$ of each protein p by the fitting parameters A_p and B_p. Then, the $E(p)$ in Eq. (19) is replaced by S_p^{reg}, and $\sigma(p)$ is obtained by the averaged error of the regression line, $\sigma(p) = \sqrt{\sum_q^{N\text{pro}} [S(p, q) - S_p^{\text{reg}}(p, q)]^2 / N_{\text{pro}}}$, where N_{pro} is the number of protein chains in the database.

We also tried a normalization using the clustering tree of the structures. A tree-distance is obtained by the following three steps. First, an all-versus-all comparison using Matras was performed, and the Z-score was calculated. Second, the average-distance clustering (UPGMA method) was performed using the Z-score, in which the distance of the clustering is defined by the following equation:

$$D(p, q) = \begin{cases} 20 - \bar{Z}(p, q) & \bar{Z}(p, q) \leq 20 \\ 0 & \text{otherwise} \end{cases} \qquad (20)$$

where $\bar{Z}(p, q)$ is an averaged Z-score defined by $\{Z(p, q) + Z(q, p)\}/2$. The tree-distance $D_{\text{tree}}(p, q)$ is newly obtained by summing up the length of all of the branches that connect protein p to protein q. Figure 4 illustrates how to obtain the tree-distance.

Computational Implementation

The program Matras is implemented in C language for UNIX workstations. Using an SGI O2 workstation (MIPS R10000 175 MHz), a pairwise alignment can be generated in 2 sec for 150 residue protein pairs (e.g., globins), and in 12 sec for 300 residue protein pairs (e.g., TIM-barrels). For the all-versus-all comparison of the SCOP database, we used a Fujitsu VPP-500 supercomputer. Using 16 PEs, it took about 24 h to compare all 1487 chains.

RESULTS AND DISCUSSION

We regarded the SCOP database[29] as the standard for homologous relationship. The SCOP database classifies protein similarity in four hierarchical levels: "class," "fold," "superfamily," and "family." In these four levels, protein pairs within the same "superfamily" or "family" are classified as homologous, by structural or functional similarity. The taxonomy of each protein is represented by a taxonomy number. For example, the PDB entry "1btc" has the taxonomy number "3.1.1.2," which means it belongs to the α/β class (3), the β/α (TIM)-barrel fold (3.1), the glycosyltransferase superfamily (3.1.1), and the β-amylase family (3.1.1.2). We evaluated the performance of our program in terms of whether it can correctly discriminate protein pairs in each superfamily or family from other nonhomologous pairs. The dataset to be examined contains 1,487 domains taken from the SCOP database, including 4,670 homologous pairs at the level of either superfamily or family.

Coverage-Reliability Plots for Distinguishing SCOP Superfamily

To evaluate the ability to detect homologous pairs, we used a coverage-reliability plot. Coverage and reliability depend on a threshold score value S, and are defined as follows:

$$\text{Coverage } (S) = \frac{N_{tp}(S)}{N_t} \qquad (21)$$

$$\text{Reliability } (S) = \frac{N_{tp}(S)}{N_p(S)} \qquad (22)$$

where $N_{tp}(S)$ is the number of homologous protein pairs that have a similarity score greater than S, N_t is the number of homologous protein pairs, and $N_p(S)$ is the number of protein pairs that have a similarity score greater than S. All of the protein pairs are sorted according to the score, then the coverage and reliability are calculated and plotted against all of the observed scores in increasing order. The curves plotted at the upper right are better than those at the lower left. In this plot, the score must be symmetric for proteins, and an averaged value Z-score, $\bar{Z}(p, q)$ ($= \{Z(p, q) + Z(q, p)\}/2$) is employed as the similarity score. Many researchers used similar but slightly different plots to evaluate their recognition methods.[40–43] Our plot is the same as the "sensitivity-specificity plot" introduced by Rice and Eisenberg.[42] However, because other researchers differently defined "sensitivity" and "specificity" (for example, see Russel et al.[43]), we use the term "coverage-reliability plot" to avoid confusion.

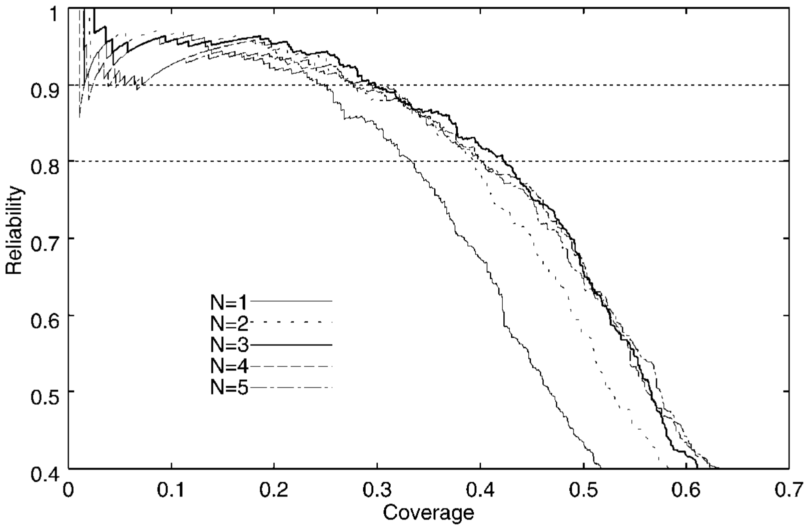


Fig. 5. Coverage-reliability plots for different evolutionary steps *N* = 1, 2, 3, 4, and 5.

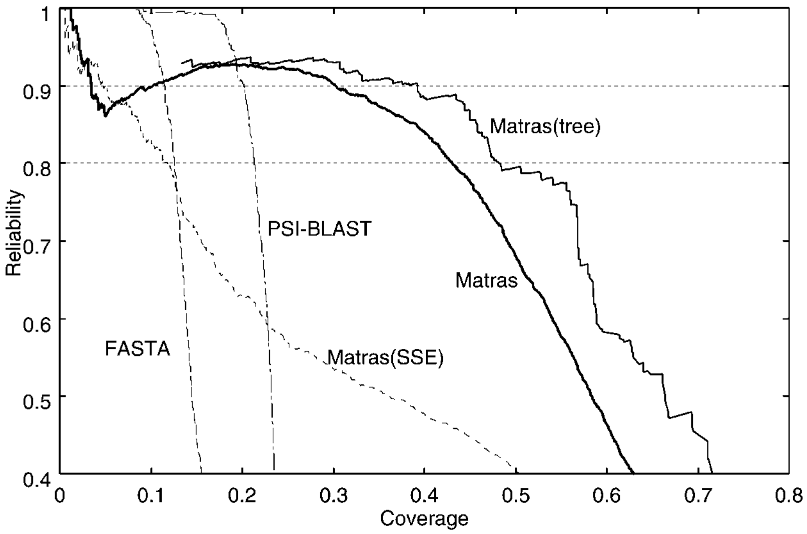


Fig. 6. Coverage-reliability plots for various methods. Profiles of PSI-BLAST were made using the nonredundant database composed of Swissprot, PIR, and DAD database (ftp.ddbj.nig.ac.jp/pub/database/dad) and the threshold *E*-value was set to 0.001, and the maximum number of iterations was set to 10.

First, the recognition test was performed for different evolutionary steps, $N = 1, 2, 3, 4$, and 5. The resulting coverage-reliability plots are shown in Figure 5. In this calculation, we used 744 representative domains that were randomly selected from the 1,487 lists to reduce the computational time. The results for $N = 1$ are apparently worse than those for the other cases. In the reliability range of 0.8–0.9, the coverage of $N = 3$ is slightly larger than the others. Hence, we used the evolutionary step $N = 3$ for the rest of our study.

Figure 6 shows the coverage-reliability plots of various methods applied to the 1,487 representative domains. The plot “FASTA” is the results of the popular pairwise sequence comparison method,[44] and the plot “PSI-BLAST” is the result of the profile-type sequence comparison method.[45] For these sequence comparison methods, an E-value is used for the similarity score. The plot “Matras” corresponds to our program Matras with the quadratic normalized Z-score. The plot “Matras (SSE)” is the result of the SSE comparison, corresponding to the first stage of our alignment strategy. Similarity was measured by the SSE score S_{sse}, which is normalized by the simple Z-score scheme. The plot “Matras (tree)” indicates the result of Matras using the tree-distance. The results in Figure 6 show that the “Matras” coverage is much larger than those of “Matras (SSE),” “FASTA,” and “PSI-BLAST” in most ranges of reliability. This means that the structure comparison program Matras is clearly better in its ability to detect homologous protein pairs than the other methods. In addition, “Matras (tree)” is even better than “Matras,” although the reason for this superiority is not clear. Although the shapes of the plots “FASTA” and “PSI-BLAST” are very similar, the “PSI-BLAST” coverage is much larger than that of “FASTA.” In the region of very high reliability (>0.95), the sequence comparison methods (FASTA, PSI-BLAST) are better than the structure comparison methods. It may be because analogous structures are more frequently observed than analogous sequences.

Comparison With FSSP Database

We compared our method with the structure database FSSP, which was automatically generated using the program DALI.[10–12] The program DALI is a famous structure comparison program, which is based on an analytically defined distance score as follows:

$$S_{\text{dali}}(D_{ix}, D_{jy}) = \left(0.2 - \frac{D_{ix} - D_{jy}}{\bar{D}}\right)\exp\left(-\left(\frac{\bar{D}}{20}\right)^{2}\right) \qquad (23)$$

where D_{ix} is the $C^{\alpha} - C^{\alpha}$ distance between the ith and xth residues, and D_{jy} is that between the jth and yth residues of another protein. $\bar{D}$ is an averaged distance defined as $(D_{ix} + D_{jy})/2$. An alignment of residues is performed using Monte Carlo methods. From the FTP site (ftp.embl-ebi.ac.uk/pub/databases/fssp/), we downloaded all of the FSSP database files created on June 16,1999, comprising 1,685 representative entries with sequence identities of 25% or less. It contains similar structure pairs with Z-score more than 2.0. After removing the nondescribed proteins in SCOP, as well as the multidomain proteins, the small proteins with less than 40 residues, and the proteins that belong to the SCOP class 6, we obtained the 755 PDB entries commonly compiled in FSSP and SCOP.

A mutual comparison between these 755 entries using Matras was performed, and the coverage-reliability plots for Matras and FSSP are shown in Figure 7. Tree-modified Z-scores were not used for this comparison. The coverage of Matras for the superfamily is much larger than that of FSSP (Fig. 7A), although the plots for the fold pairs (Fig. 7B) are similar. This means that Matras is particularly better in identifying homologous structure pairs than FSSP, probably because of the evolutionary model employed.

However, the situation is more complicated than we first expected. To clarify the difference between our score and the DALI score, we also calculated the plots of a combined procedure of Matras and DALI, called "M-Dali," in Figure 7. In the procedure M-Dali, an alignment is made by the Matras program with our score, but the final score is replaced by the DALI score S_{dali} defined by Eq. (23). Surprisingly, the plots of Matras and M-Dali are quite similar, although Matras is slightly better. Thus, the superiority of Matras over FSSP is not explained by only the difference in their scores. We suppose that both score normalization and alignment affect the results. To express the average score values, Matras and M-Dali use quadratic functions that depend on the number of compared residues (see "Score Normalization"), whereas the FSSP uses the polynomial that depends on protein size.[13] The difference in normalization is observed in Figure 8, where the dependence of the Z-score is plotted against the number of compared residues, N_{comp}. Apparently, the Z-score of FSSP correlates more positively with the number N_{comp}, than that of M-Dali. We suppose that this dependency negatively affects the coverage-reliability plot of FSSP, especially for small proteins. The alignments of FSSP and those of Matras were also compared. The correspondence between the two alignments was evaluated by the measure Q_{align}. If the alignments of FSSP and Matras are exactly identical, then the value of Q_{align} is 100%, whereas, if they do not overlap at all, it becomes 0%. Figure 9 shows the histogram for Q_{align} for 1,192 superfamily pairs whose alignments are available in FSSP. Many protein pairs have more than an 80% value of Q_{align}, but almost 14% of the pairs have no overlap (0%). Therefore, there are considerable differences between the FSSP and Matras alignments, which will be large enough to affect the recognition of a superfamily. It is difficult to decide which alignments are better, because the "correct" standard alignments are not available. We noticed that the viral coat protein (2.8.1) and the EF-hand like superfamily (1.37.1) frequently appeared in the nonoverlapping pairs. This is rather reasonable, because both folds contain repeated substructure, and they must have several candidates for the optimal structure alignment.

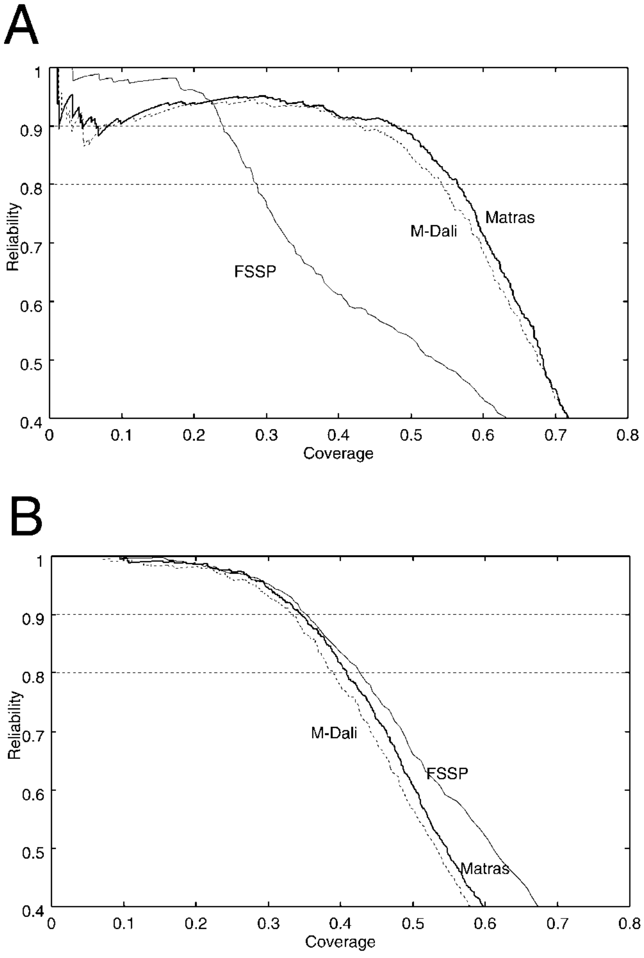


Fig. 7. Coverage-reliability plots of Matras and FSSP for distinguishing the superfamily pairs (**A**) and for the fold pairs (**B**). The curve "M-Dali" represents the method using the DALI score with the Matras alignment.

Missed Structure Similarity of Homologous Relationship

Our comparison method can recognize 48.4% of superfamily pairs with 80% reliability. This means that 51.6% of

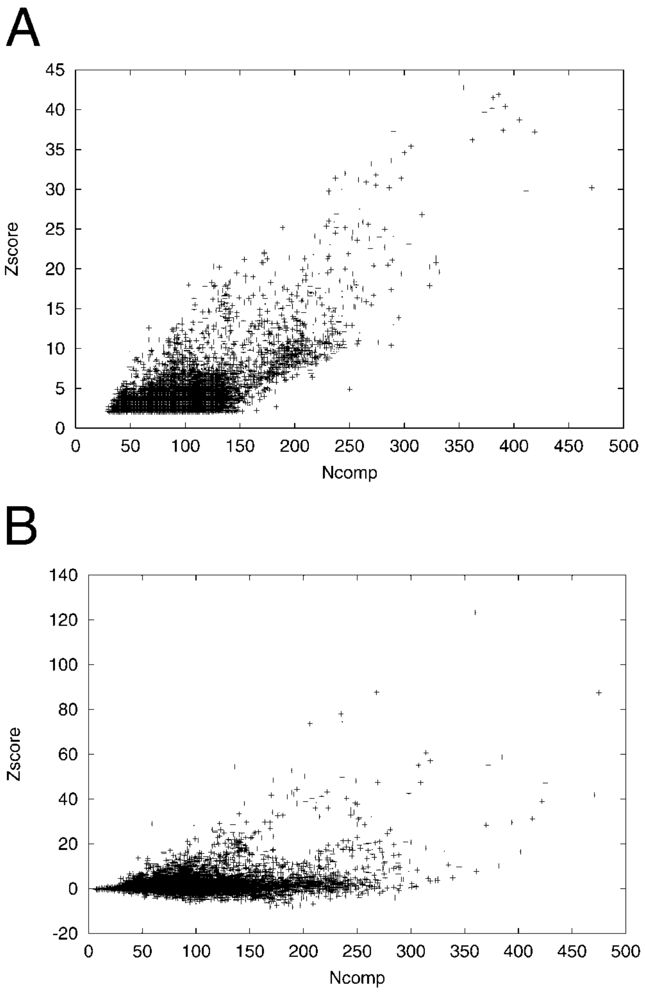


Fig. 8. Plots of *Z*-score versus number of compared residues N_{comp} for FSSP (**A**) and for M-Dali (**B**). The *Z*-score of FSSP correlates more with the number of compared residues than that of M-Dali.

the superfamily pairs lack a sufficiently high score, and 20% of the sufficiently high score pairs do not belong to common SCOP superfamilies. Therefore, it is interesting to know which superfamilies often have low similarity score, and which are misjudged by our comparison method. This survey is important for the following two points. First, we can find the defects of our program by checking its misjudgment. Second, missed homologous relationships will be examples of nonstructural factors, such as biological function, which show their homologous relationship, although they are not similar in structure. To find homologous relationships correctly, it is necessary to know what kinds of functional similarities suggest homologous relationship.

To study the superfamilies with low similarity scores, we employed the clustering score proposed by Przytycka et al.,[46] which qualifies the agreement between the SCOP grouping and the clustering tree obtained by automatic methods. Before calculating the score, the topological distance *td* must be evaluated for each homologous protein pair. To calculate the distance *td*,

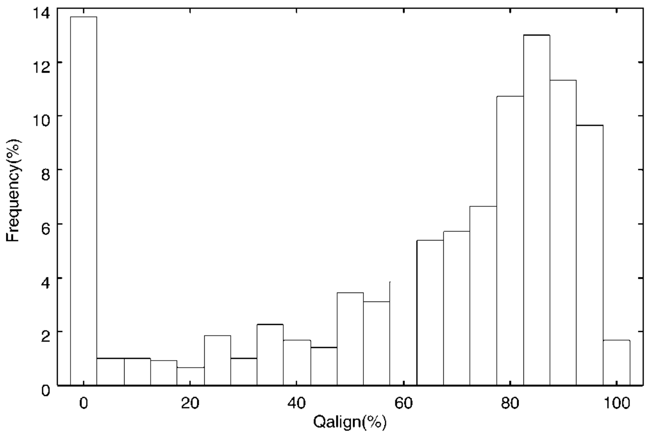


Fig. 9. Correspondence of alignment of FSSP and Matras for 1,192 protein pairs of the same superfamily. The horizontal axis is the measure Q_{align}, and the vertical axis is the observed frequency. Q_{align} is defined as $N_{iden}/\bar{N}_{comp}$, where N_{iden} is the number of pairs of residues where the alignment of FSSP is identical with that of Matras, and $\bar{N}_{comp}$ is the averaged number of compared residues by FSSP and Matras.

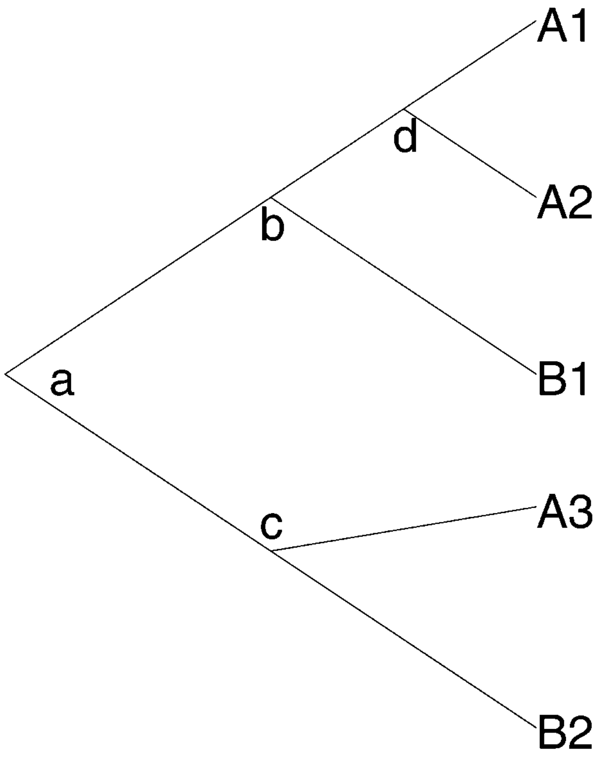


Fig. 10. Example of a calculation of the topological distance *td* and the clustering score. Proteins A1, A2, and A3 belong to the superfamily A, and proteins B1 and B2 belong to the superfamily B. The topological distance *td* between proteins A1 and A2 is 1.0, because their latest common ancestor is node “d” and all of the nodes under “d” belong to superfamily A. However, the distance *td* between proteins A1 and A3 is 0.6 (=3/5), because their latest common ancestor is node “a,” there are five nodes under node “a,” and three nodes among the five belong to the superfamily A. The clustering score of superfamily A is defined as an averaged topological distance: {*td* (A1, A2) + *td* (A1, A3) + *td* (A2, A3)}/3 = {1.0 + 0.6 + 0.6}/3 = 0.73.

TABLE I. Ten SCOP Superfamilies With the Lowest Clustering Score

		Fold/superfamily[a]	P^b	CS^c	Division of group[d]
1	3.29.1	P-loop nucleotide triphosphate hydrolases	23	0.29	9 7 5 1 × 2
2	1.107.1	Multiheme cytochromes	5	0.30	3 1 × 2
3	3.23.1	Biotin carboxylase N-terminal domain-like	5	0.42	3 2
4	4.3.1	Cysteine proteinases	7	0.48	5 1 × 2
5	4.12.1	FAD-linked reductases, C-terminal domain	6	0.51	4 1 × 2
6	2.29.4	OB-fold/nucleic acid-binding domain	18	0.53	7 4 3 1 × 4
7	3.47.3	Ribonuclease H-like motif/ribonuclease H-like	11	0.57	5 5 1
8	1.4.3	3-helical bundle/winged helix DNA binding	17	0.58	10 3 2 1 × 2
9	2.24.2	SH3-like barrel/SH3-domain	7	0.59	5 1 × 2
10	2.1.1	Ig-like/immunoglobulin	46	0.66	20 7 4 3 1 × 12

[a] Taxonomy number of SCOP and its name of fold and superfamily. If a fold contains only one superfamily, then the name of superfamily is omitted.
[b] Number of proteins that belong to a given superfamily.
[c] Clustering score
[d] These numbers represent how a superfamily is divided into monophyletic Matras groups. For example, "3 1 × 2" means five proteins are categorized as three Matras groups: a group with three proteins and two groups with one protein.

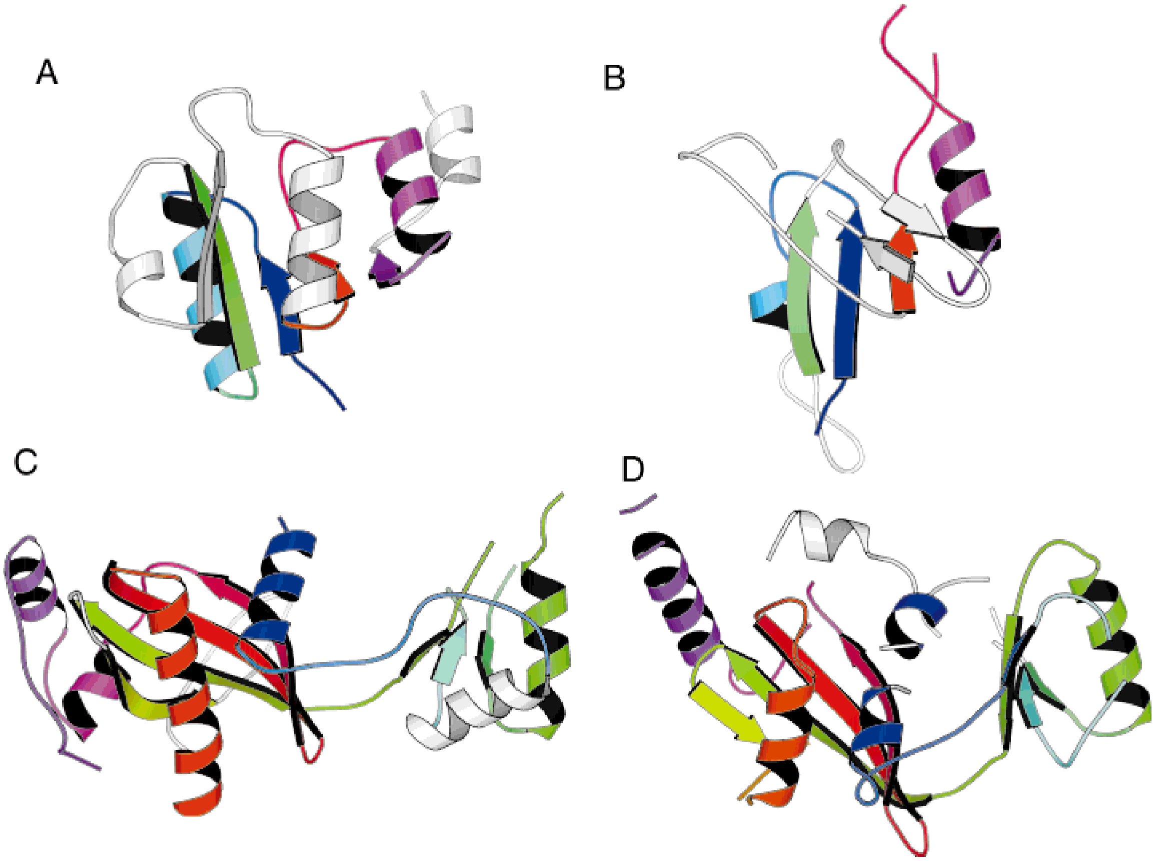


Fig. 11. Domains next to similar domains. (**A**) Biotin carboxylase N-terminal domain (1bncA1, 3.23.1). (**B**) Synapsin IA N-terminal domain (1auvA1, 3.23.1). (**C**) Biotin carboxylase central domain (1bncA2, 4.89.1). (**D**) Synapsin IA C-terminal domain (1auvA2, 4.89.1). The first domains (A and B) belong to the same superfamily (3.23.1), but their *Z*-score is not high enough for a homolgous relationship (*Z*-score = 1.7). However, the homology between them is supported by their neighboring domains (C and D), which have significant structural similarity for homology (*Z*-score = 28.1). All of the ribbon diagrams were drawn with Molscript.[54]

first, the latest common ancestor node between the two proteins should be found in the tree. Then, the distance *td* is evaluated as the number of the common superfamily proteins under the ancestor node divided by the number of proteins under the ancestor node (see Fig. 10). The clustering score of a superfamily is defined as the summed topological similarities between all pairs of members divided by the number of pairs. The score ranges from 0 to 1, and a high value means that the members of the superfamily belong to a nearly monophyletic group in the Matras tree. Figure 10 illustrates how to calculate this score.

To calculate a reliable clustering score, we only used 65 superfamilies in the dataset, each of which consisted of at least five proteins. Among them, nearly half of the superfamilies (31/65) are completely monophyletic (their clustering scores are 1.0). Table I summarizes the ten worst cases. These superfamilies are divided into several groups in the Matras tree. There might be two reasons to explain this disagreement. The first reason is a pure defect in the Matras program, and the second one is that the SCOP classified them based on local similarities that are functionally important, rather than global structure similarity. Most of the cases in Table I can be explained by the latter reason. The P-loop superfamily (3.29.1) has the worst clustering score, and is divided into three large, separated groups by Matras. This subdivision is reasonable, because this superfamily contains members with quite different numbers and topologies of β-strands,[47] whereas they commonly share the nucleotide binding local motif (called the "P-loop"). The second worst in Table I is the multiheme cytochrome superfamily (1.107.1), which is divided into one group with three members (1aqe-, 1wad-, and 1new-), and two one-member groups (1fgjA and 1prcC). Although all of the members of this family contain several hemes, bound to the "CxxCH" motif, the overall structures of 1fgjA and 1prcC are quite different from the others. A similar situation also occurred with the fourth superfamily, cysteine proteinases (4.3.1), which is subdivided into one large group (1aec-, 1cjl-, 1gcb-, 1mirA, and 1uch-) and two one-member groups (1avpA and 1fieA2). SCOP claims that all of these proteins have in common the catalytic triad Cys-His-Asn, however, the structures of the latter two proteins are quite different from those of the other five, because of drastic permutations and insertions of domains. Apparently, in the above three cases, SCOP classifies them in view of the local structure similarity that is important for biological function, but not of the global aspects of structures. However, the third superfamily in Table I, biotin carboxylase N-terminal domain-like (3.23.1) may arise for a different reason. The Matras program classified them into two separated groups: (1bncA1, 1jdbB1) and (1auvA1, 1glv-1, 1iow-1). The structures of these two groups are not very different, but are not similar enough for a homologous relationship. Judging from the comment of the SCOP database, evidence of homology exists in their neighboring domains. That is, the 3.23.1 domain is always followed by an ATP-grasp domain (4.89.1), which are very similar to each other. Figure 11 illustrates the situation, suggesting that SCOP employed the structural similarity of the neighboring domain as a homology criterion. Figure 12 shows the Matras clustering tree for the immunoglobulin superfamily (2.1.1). Classification of this large superfamily seems very difficult for automatic methods, because there are many confusing analogous structures, that belong to the immunoglobulin-like fold (2.1). The tree shows that the members of the V, I, and C1 sets are well clustered together, but those of the C2(2.1.1.3) and E(2.1.1.5) sets are scattered over wide ranges.

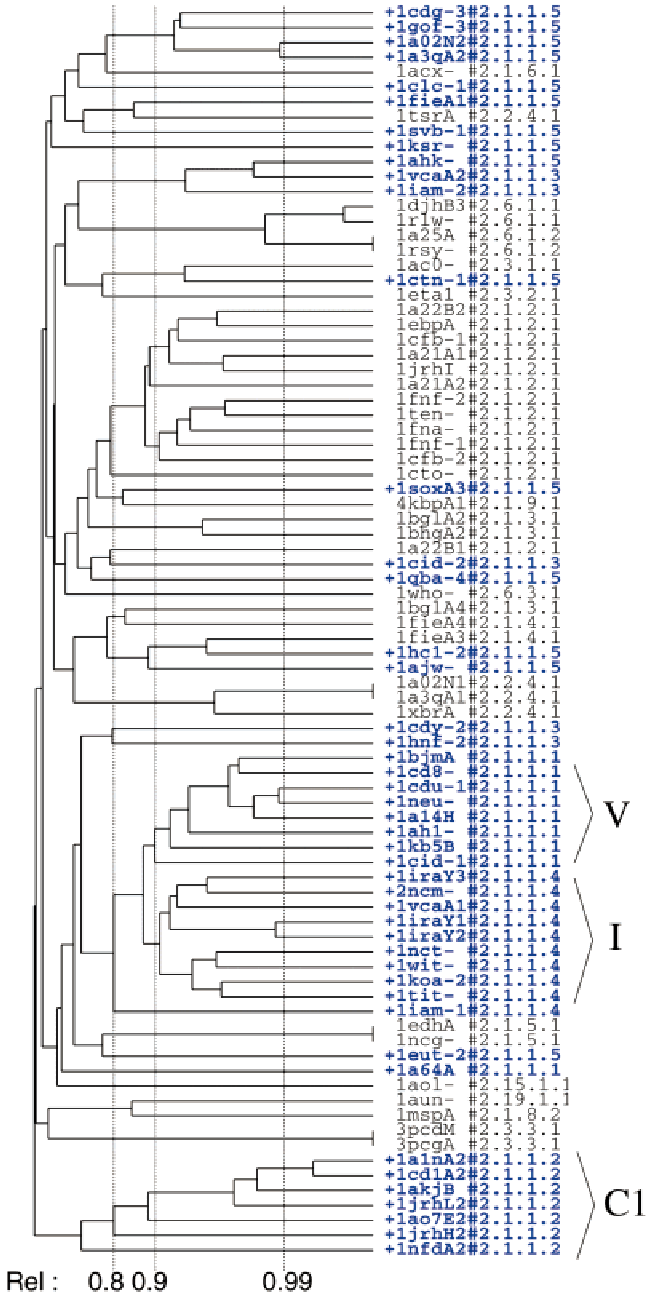


Fig. 12. The clustering tree for a Matras group including the immunoglobulin superfamily (2.1.1). The strings at each of the leaves show the PDB code name and the SCOP taxonomy number. The nodes of the immunoglobulin superfamily are colored blue. The clustering score of this superfamily is 0.66. Most of the proteins belong to the family V set (2.1.1.1), C1 set (2.1.1.2), and C1 set (2.1.1.3) are in each monophyletic group, but the proteins of the C2 (2.1.1.3) and E sets (2.1.1.5) appear in various positions in the tree. This tree includes other superfamilies of immunoglobulin-like folds (2.1.2, 2.1.3, 2.1.4, 2.1.5, 2.1.8) and other greek key β-sandwich folds, such as the common fold of diphtheria toxin/transcription factors/cytochrome f(2.2), the prealbumin-like fold (2.3), the C2 domain-like fold (2.6), the F-MuLV receptor-binding domain fold (2.15), and the thaumatin-like protein fold (2.19).

TABLE II. Ten Matras Groups With the Lowest Reliability Scores

	S[a]	F[b]	R[c]	PDBcode[d]			Comments

1	5	1	0.00	2bct-(1.91.1)
				1ft1A(1.91.6)
					1lrv-(1.91.2)
					1a17-(1.91.8)
						1sly-1(1.91.5)	α-α superhelix fold (1.91)

2	5	4	0.07	1bucA2(1.24.6)
				1e2aA(1.7.2)
					1aep-(1.53.1)
					1aosA(1.98.1)
						1aj3-(1.7.1)
						1fupA(1.98.1)
							Mixture of 3-5 helical bundle folds (1.7, 1.24, 1.53, 1.98)

3	6	3	0.11	1dim-(2.50.1)
				1gof-2(2.51.1)
				4aahA(2.52.1)
					1eur-(2.50.1)
					1mdaH(2.51.2)
					1aomA1(2.52.2)
						1inf-(2.50.1)
						1gotB(2.51.3)
							Mixture of 6,7,8-bladed β-propeller folds (2.50, 2.51, 2.52)

4	6	1	0.11	1le2-(1.24.1)
				1bbhA(1.24.3)
				1cgmE(1.24.5)
					1lih-(1.24.2)
					1a7vA(1.24.3)
					1fapB(1.24.7)
						256bA(1.24.3)
						1hmdA(1.24.4)
							Four-helical up-and-down bundle fold (1.24)

5	5	1	0.13	1hib-(2.31.1)
				1abrB1(2.31.2)
				1jlxA2(2.31.3)
				1hcd-(2.31.5)
					1ilr1(2.31.1)
					1abrB2(2.31.2)
					1af9-2(2.31.4)
						1afcA(2.31.1)
						1jlxA1(2.31.3)
						1ba7A(2.31.4)
							β-trefoil(2.31)

6	5	1	0.14	1cauA(2.64.1)
				1ipsA(2.64.3)
				1rgs-1(2.64.4)
					1cauB(2.64.1)
					2cgpC1(2.64.4)
					2aacA(2.64.5)
						1pmi-(2.64.2)
						1rgs-2(2.64.4)
							Double-stranded β-helix (2.64)

7	3	3	0.20	1dar-2(2.32.3)
				1bmfD1(2.37.1)
					1aipA2(2.32.3)
					1bmfA1(2.37.1)
						1aipA3(2.33.1)	Mixture of β-barrel folds.
							$S = 8$(2.37) and $S = 10$(2.32, 2.33).

8	4	2	0.25	1air-(2.62.1)
				1rmg-(2.62.3)
				1thjA(2.63.1)
					1idjA(2.62.1)
					1lxa-(2.63.1)
					1xat-(2.63.1)
						1tsp-(2.62.2)
						1tdtA(2.63.1)
							Confusing right and left handed β-helix folds (2.62, 2.63)

9	3	2	0.30	1vhiA(4.34.9)
				1dar-4(4.34.13)
					2bopA(4.34.9)
					1tbd-(4.50.1)
						1dhmA(4.34.9)	Mixture of Ferredoxin-like fold (4.34) and similar fold (4.50)

10	4	1	0.31	1a9o-(3.58.1)
				1a2zA(3.58.3)
				1cbx-(3.58.4)
					1ecpA(3.58.1)
					1xjo-(3.58.4)
					1amp-(3.58.4)
						2pth-(3.58.2)
						1lam-(3.58.4)
						1obr-(3.58.4)
							Phosphorylase/hydrolase-like fold (3.58)

[a] Number of superfamilies in the Matras group.
[b] Number of folds in the Matras group.
[c] Reliability for each Matras group, defined as N_{tp}^{m}/N_{p}^{m}, where N_{p}^{m} is the number of all protein pairs within a Matras group m, and N_{tp}^{m} is the number of protein pairs that belong to the same superfamily in the Matras group m.
[d] PDB code and SCOP taxonomy number in parentheses.

Overestimated Structure Similarity of Nonhomologous Relationship

Next, we focus on the opposite cases, where nonhomologous protein pairs were assigned as homologs by Matras. To analyze this disagreement, first, the similar structural groups are constructed using the Matras program. The threshold value of the tree-normalized Z-score is set to 6.11, which corresponds to 80% reliability. We examined 77 Matras groups with at least five members, according to the reliability score. Table II summarizes the ten Matras groups with the lowest reliability scores. Each group contains members of different SCOP superfamilies, although in half of the ten groups, all of the members belong to a single SCOP fold. In the list, we noticed some tendency toward discrepancy. First of all, some of these discrepancies were caused by partial similarity. For example, Matras group 2 consists of six members with helical bundle folds, having various numbers of α-helices. Figure 13 shows a typical example of the partial similarity between the acyl-coA dehydrogenase C-domain(1bucA2:1.24.6) and arginosuccinate lyase (1aosA:1.98.1). The latter includes the entire four-helix bundle folds that are similar to the former structure, in addition to many other α-helices. The Matras group 9 contains the ferredoxin-like fold (4.34) and the replication origin DNA-binding domain of SV40 T-antigen (4.50). The latter fold (4.50), whose β-strand topology is "41325," includes the "4132" topology that corresponds to the ferredoxin-like fold (4.34).

In addition, repetitive structures, such as α-α superhelix fold (1.19; Matras group 1) and β-propeller folds (2.50, 2.51, 2,52; Matras group 3), also tend to have partial similarity in another sense. The members of the α-α superhelix fold have in common several α-helix pairs, although the number of pairs is different from one another. β-propeller folds are composed of several repeats of "blades," consisting of antiparallel β-sheets. Examples of 7- and 8-bladed propeller folds are shown in Figure 14, where two structures, the guanine nucleotide-binding protein β-subunit (1gotB:2.51.3) and the nitrite reductase domain (1aomA1:2.52.2), are similar except for an extra blade

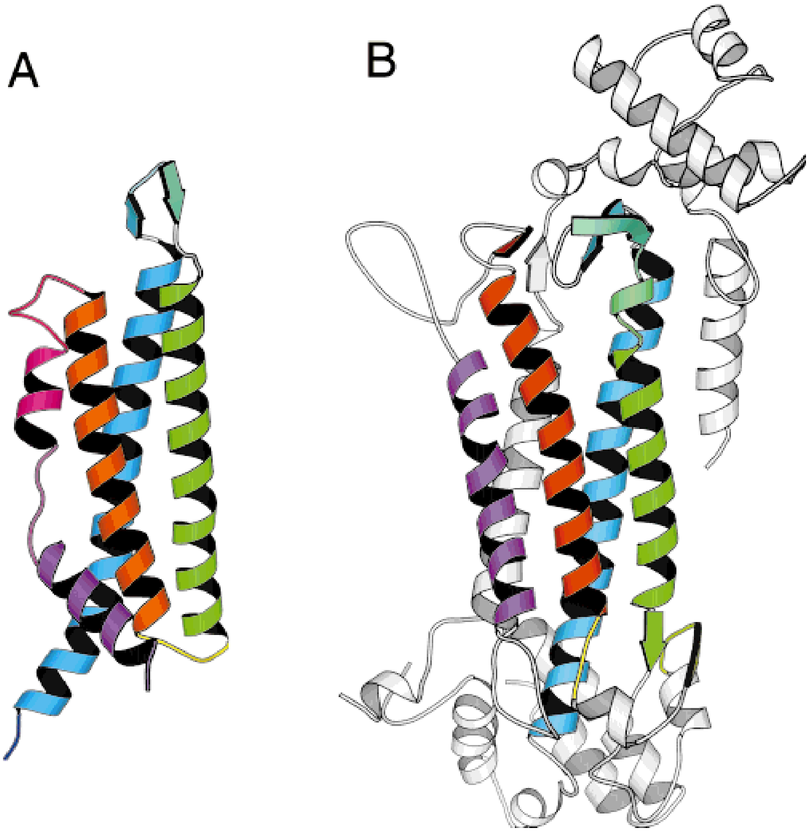


Fig. 13. Disagreement between SCOP and Matras in helical bundle folds. (**A**) acyl-coA dehydrogenase N-terminal domain (1bucA2:1.24.6). (**B**) Argininosuccinate lyase (1aosA:1.98.1). They have significant structural similarity in Matras for a homologous relationship, but SCOP classified them as different folds.

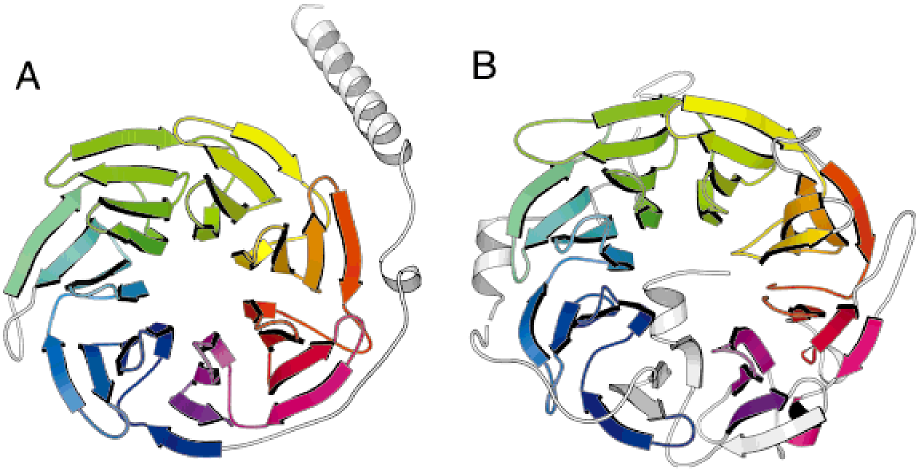


Fig. 14. Disagreement between SCOP and Matras in the β-propeller fold. (**A**) Guanine nucleotide-binding protein beta subunit (1gotB), which belongs to the seven-bladed β-propeller fold (2.51). (**B**) Nitrite reductase (1aomA, domain 1), which belongs to the eight-bladed β-propeller fold (2.52). These proteins belong to different folds, but their *Z*-score in Matras is high enough ($Z = 20.0$) for homology.

domain was in the latter. Although these two proteins have no functional similarity, the homologous relationship was suggested from a PSI-BLAST search,[45] which detected weak sequence similarity between them.[48]

Another kind of discrepancy is seen in the shear number (S) of β-barrel structures,[49] which seems to be one of the important criteria in the SCOP classification.[50,51] Matras assigned β-barrels of different shear numbers together into one group (Matras group 7 in Table II). On the other hand, SCOP classified them into three different folds (2.32, 2.33, and 2.37). Figure 15 shows examples: the shear number is $S = 8$ for 1bmfA1 (2.37), and $S = 10$ for 1aipA2 (2.32) and 1aipA3 (2.33). The latter two structures, corresponding to different domains of elongation factor TU, are classified into the separate SCOP folds because of a difference in the topology of one strand (Fig. 15). The CATH database[52] assigned these two domains as homologous because the two domains are encoded in tandem on the gene, suggesting a distant gene duplication event.[53]

The confusion seen in group 8 is a pure fault of Matras. Right- (2.62) and left- (2.63) handed β-helices, or chirality in general, cannot be recognized by our distance score. This defect could be corrected by considering another measure, such as rmsd values. The rest of the disagree-

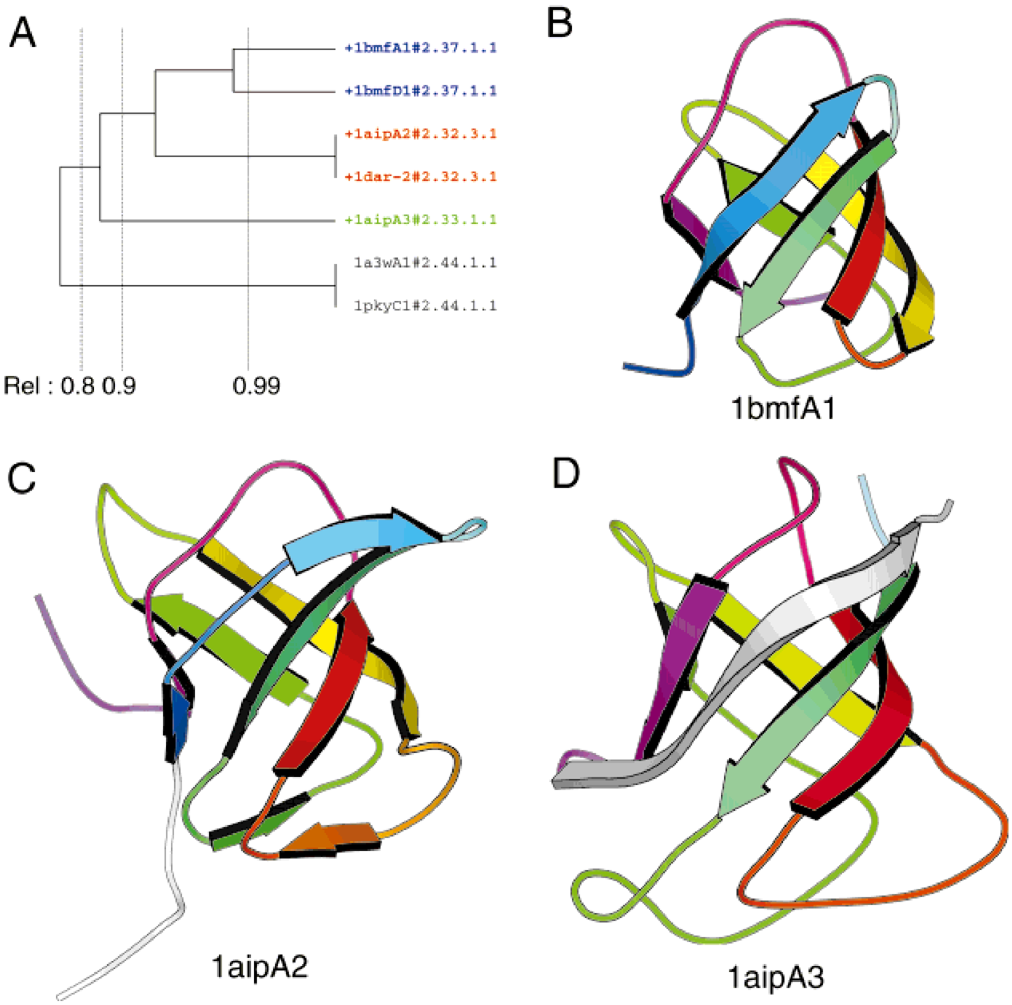


Fig. 15. β-barrel folds of different shear numbers (S = 8 and S = 10). (**A**) A dendrogram that contains all of the domains of the SCOP folds 2.33 and 2.37, and a part of the fold 2.32. (**B**) ATP synthetase α chain domain 1 (1bmfA), with shear number 8. (**C**) Elongation factor TU (1aipA) domain 2, which belongs to the fold 2.32, with shear number 10. (**D**) Elongation factor TU (1aipA) domain 3, which belongs to the fold 2.33, with shear number 10. The topology of the gray colored strand is different from those of the corresponding strands (blue color) of the other two structures.

ments in groups 4, 5, and 10 are not serious. These groups consist of various superfamilies of the same fold: four-helical up-and-down bundle folds (1.24), β-trefoil folds (2.31), and phosphorylate/hydrolase-like folds (3.58). They have common structural property that are significant enough to suggest a homologous relationship, but no common function or sequence features to support their homology.

In conclusion, the discrepancies listed in Tables I and II are rather rational, and each has its own reasoning behind it. We know that the classification of superfamily in the SCOP database has been deliberately made, not automatically but sometimes manually with intuition, and taking not only the structure itself but also the functional aspects of proteins into account. Thus, it is impossible to produce identical results in an automatic way. Matras is better at distinguishing homologous relationships than FSSP, although the superiority of our novel score itself is not confirmed. As a whole, we can say that Matras produces reasonable results (except for the few faults already mentioned) in a fully automatic way, and within a short computational time. The method may be suitable to prepare a large number of objective structural similarity data for general use.

ACKNOWLEDGMENTS

A postdoctoral fellowship was provided to TK by the ACT-JST Program, Japan Science and Technology Corporation.

REFERENCES

1. Murzin AG. Structural classification of proteins: new superfamilies. Curr Opin Struct Biol 1996;6:386–394.
2. Murzin AG. How far divergent evolution goes in proteins. Curr Opin Struct Biol 1998;8:380–387.
3. Chothia C, Finkelstein V. The classification and origins of protein folding patterns. Annu Rev Biochem 1990;59:1007–1039.
4. Hwang KY, Chung JH, Kim SH, Han YS, Cho Y. Structure-based identification of a novel NTPase from *Methanococcus jannaschii.* Nature Struct Biol 1999;6:691–696.

5. Moult J, Hubbard T, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round III. Proteins (suppl.) 1999;3:2–6.
6. Holm L, Sander C. Searching protein structure databases has come of age. Proteins 1994;19:165–173.
7. Orengo CA. Classification of protein folds. Curr Opin Struct Biol 1994;4:429–440.
8. Taylor WR, Orengo CA. Protein structure alignment. J Mol Biol 1989;208:1–22.
9. Orengo CA, Brown NP, Taylor WR. Fast structure alignment for protein databank searching. Proteins 1992;14:139–167.
10. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
11. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–602.
12. Holm L, Sander C. Alignment of three-dimensional protein structures. Meth Enzymol 1996;266:653–662.
13. Holm L, Sander C. Dictionary of recurrent domains in protein structures. Proteins 1998;33:88–96.
14. Rossmann MG, Argos P. Exploring structural homology of proteins. J Mol Biol 1976;105:75–95.
15. Russel RB, Barton GJ. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 1992;14:309–329.
16. Gerstein M, Levitt M. Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. Prot Sci 1998;7:445–456.
17. Rose J, Eisenmenger F. A fast unbiased comparison of protein structures by means of the Needleman-Wunsch algorithm. J Mol Evol 1991;32:340–354.
18. Zu-Kang F, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. Folding Design 1996;1:123–132.
19. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extention (CE) of the optimal path. Prot Engin 1998;11:739–747.
20. Taylor WR. Protein structure comparison using iterated double dynamic programming. Prot Sci 1999;8:654–665.
21. Mitchell EM, Artymiuk PJ, Rice DW, Wilett P. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. J Mol Biol 1989;212:151–166.
22. Mizuguchi K, Go N. Comparison of spatial arrangements of secondary structure elements in proteins. Prot Engin 1995;8:353–362.
23. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. Proteins 1995;23:356–369.
24. Alexandrov NN, Fischer D. Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. Proteins 1996;25:354–365.
25. Lathrop RH. The protein threading problem with sequence amino acid interaction preferences is NP-complete. Prot Engin 1994;7: 1059–1068.
26. Dayhoff MO, Shwartz RM, Orcutt BC. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure, Volume 5, Supplement 3. Washington DC: National Biomedical Research Foundation; 1978. p 345–352.
27. Henikoff S, Henikoff AG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
28. Fukami-Kobayashi K, Tateno Y, Nishikawa K. Domain dislocation: a change of core structure in periplasmic binding proteins in their evolutionary history. J Mol Biol 1999;286:279–290.
29. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of protein database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
30. Russel RB, Barton GJ. Structural features can be unconserved in proteins with similar folds: an analysis of side-chain to side-chain contacts, secondary structure and accessibility. J Mol Biol 1994; 244:332–350.
31. Russel RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJE. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 1997;269:423–439.
32. Holm L, Sander C. Decision support system for the evolutionary classification of protein structures. ISMB 1997;5:140–146.
33. Matsuo Y, Bryant SH. Identification of homologous core structures. Proteins 1999;35:70–79.
34. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. CABIOS 1992;8: 275–282.
35. Sippl MJ. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
36. Kabsh W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
37. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
38. Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol 1982;162:705–708.
39. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci USA 1998;95:5913–5920.
40. Brenner SE, Chothia C, Hubbard TJP. Assessing sequence comparison methods with reliable structurally identified evolutionary relationship. Proc Natl Acad Sci USA 1998;95:6073–6078.
41. Park J, Karplus K, Barret C, Hughney R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998;284:1201–1210.
42. Rice DW, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
43. Russel RB, Sasieni PD, Sternberg MJE. Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol 1998;282:903–918.
44. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–2448.
45. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 1997;25: 3389–3402.
46. Przytycka T, Aurora R, Rose GD. A protein taxonomy based on secondary structure. Nature Struct Biol 1999;6:672–682.
47. Swindells MB. Classification of double round nucleotide binding topologies using automated loop searches. Prot Sci 1993;2:2146–2153.
48. Wolf YI, Brenner SE, Bash PA, Koonin EV. Distribution of protein folds in the three superkingdoms of life. Genome Res 1999;9:17–26.
49. MacLachlan AD. Gene duplication in the structural evolution of chymotrypsin. J Mol Biol 1979;128:49–79.
50. Murzin AG, Lesk AM, Chothia C. Principle determining the structure of β-sheet barrels in proteins. I. A theoretical analysis. J Mol Biol 1994;236:1369–1381.
51. Murzin AG, Lesk AM, Chothia C. Principle determining the structure of β-sheet barrels in proteins. II. The observed structures. J Mol Biol 1994;236:1382–1400.
52. Orengo CA, Michie AD, Jones S, Jones DT, Swindells M, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;51–22.
53. Hadley C, Jones DT. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure 1999;7:1099–1112.
54. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Cryst 1991;24:946–950.