ELSEVIER

Amino acid similarity matrix for homology modeling derived from structural alignment and optimized by the Monte Carlo method

Koji Ogata,* Masanori Ohya,† and Hideaki Umeyama*

**School of Pharmaceutical Sciences, Kitasato University, Tokyo, Japan*
†Faculty of Science & Technology, Science University of Tokyo, Chiba, Japan

In this paper, we obtained a similarity matrix for homology modeling based on the structure of proteins in a structural alignment. The alignment procedure was executed within dynamic programming generally used in alignment methods. An initial matrix derived from the structural alignment was optimized by the Markov chain Monte Carlo method at low temperature to fit its sequence alignment to the structural alignment. Structural alignment was performed on the basis of the superposition of Cα atoms for two protein structures. The objective function in the Monte Carlo procedure was defined by entropy in the information theory, allowing us to show that the amino acid similarity matrix aligned accurately. When compared with the structural alignment, the average number of incorrect amino acid residues in the sequence alignment was 22.6 for all residues and about 3.7 for residues in structurally conserved regions. The alignment with our matrix was more similar to structural alignment than to sequence alignments using other amino acid substitution matrices. © 1999 by Elsevier Science Inc.

Keywords: sequence alignment, structural alignment, amino acid similarity matrix, Monte Carlo method, probabilistic entropy, information theory.

INTRODUCTION

Many tertiary structures of proteins have been determined recently by newly developed technologies. Tertiary structural information has been used to explain protein functions, characterization, and other properties. Moreover, the physicochemical and statistical features of proteins have been derived from structural information, and these have been used to predict or analyze the tertiary structures.[1–4] In such predictions or analyses, alignment is a helpful tool with which to compare two protein sequences or structures.

Alignment algorithms have been applied in dynamic programming, which is widely used in sequence comparison.[5–9] They have also been improved for more accurate alignments.[10–13] These alignment algorithms have also been applied to structural alignment[14–19] and structurally based sequence alignment.[20,21] Moreover, other methods dealing with combinational optimization problems have been published.[22–24]

In sequence alignment, an amino acid similarity matrix defined between amino acid types has been used to determine molecular evolution or physicochemical amino acid properties.[25–27] However, for two proteins whose structures are known, sequence alignment with these matrices involves residues sometimes being given wrong assignments. In predicting protein structure by, e.g., homology modeling, fatal errors can be introduced and ruin a predicted structure if the residues in structurally conserved regions (SCRs) within a homologous family of proteins are shifted.

On the other hand, sequence–structural alignment (threading) has been performed with the potential defined between amino acid types and tertiary structures of proteins. It has been reported that these predictions could be used to guess the folding pattern of an amino acid sequence.[28–30] However, the potentials used do not provide the similarity between amino acid types. If we can adapt the structural information to the similarity matrix, this alignment can include structural information.

Various papers have shown the results of structural alignment.[31,32] However, sequences aligned with the obtained similarity matrix often differ from the answer sequences obtained from structural alignment, even though the matrix was derived

The Color Plate for this article is on page 254.

Address reprint requests to: K. Ogata, School of Pharmaceutical Sciences, Kitasato University, 5-9-1, Shirokane, Minato-ku, Tokyo 108-8641, Japan. Tel.: (+81)-3-3446-9553; FAX: (+81)-3-3446-9553; E-mail: ogatak@platinum.pharm.kitasato-u.ac.jp.

Journal of Molecular Graphics and Modelling 16, 178–189, 1998

655 Avenue of the Americas, New York, NY 10010

1093-3263/99/$–see front matter
PII S1093-3263(99)00008-X

from the answer sequences. The difference is caused by the fact that the structural alignment does not indicate the minimum score for the similarity matrix derived from it. Because the alignment is close to the answer sequences, it may be possible to achieve improvement by optimization of the matrix.

In this article, we have tried to obtain a structurally based amino acid similarity matrix for homology modeling. The matrix was derived from the structural alignment. An alignment procedure was used in dynamic programming to minimize the distance. The initial matrix obtained from the structural alignment based on the superposition of Cα atoms for protein pairs was optimized by the Markov chain Monte Carlo method at low simulation temperature, using an objective function based on the mutual entropy of information theory. Thus, the alignment was executed on the basis of the final matrix, and the average number of incorrect residues in aligned sequences was smaller than those obtained by other methods.

METHODS

Alignment procedure

The sequence and structural alignment procedure in this work is based on minimization.

Let us consider two amino acid sequences with a gap inserted at the top of each sequence. Here, a gap means the deletion of an amino acid residue and is represented by a dash (–). These two amino acid sequences are denoted by $\mathbf{A} = a_0\, a_1 \ldots a_m$ and $\mathbf{B} = b_0\, b_1 \ldots b_n$. We created a lattice graph and placed the amino acid residues on the x and y axes. The alignment procedure was then conducted to find the optimal path passing from the starting point $\mathbf{p}_0 = (0, 0)$ to the end point $\mathbf{p}_N = (m, n)$ on the lattice graph. Finding the optimal path is the same as obtaining the minimum distance or maximum similarity defined between two proteins. Here, a set of all paths is denoted as $p(\mathbf{p}_N)$, and the kth lattice point that passes through an optimal path is given by (i_k, j_k), where $0 \le i_k \le m$ and $0 \le j_k \le n$. A pair of amino acid residues corresponding to the kth path is defined by $\mathbf{x}_k = (a_{i_k}, b_{j_k})$. The distance between two sequences is defined by Eq. (1):

$$D(A,B) = \min\left\{\sum_{k=0}^{N} d(\mathbf{x}_k);\ \mathbf{p} = \{\mathbf{p}_0, \ldots, \mathbf{p}_N\} \in p(\mathbf{p}_N)\right\} \quad (1)$$

where function d means the distance or similarity defined between two amino acid residues. A path having the starting point $\mathbf{p}_0$ and the end point $\mathbf{p}_k$ is represented by $D(\mathbf{p}_k)$. Then, the optimal path is given by Eq. (2):

$$\begin{aligned} D(\mathbf{p}_k) &= \min\left\{\sum_{r=0}^{k-1} d(\mathbf{p}_r);\ \mathbf{p}_1, \ldots, \mathbf{p}_{k-1}\right\} + d(\mathbf{x}_k) \\ &= \min\{D(\mathbf{p}_{k-1});\ p(\mathbf{p}_{k-1})\} + d(\mathbf{x}_k) \\ &= \min\{D((i_k - 1, j_k)),\ D((i_k - 1, j_k - 1), \\ &\qquad D((i_k, j_k - 1))\} + d(\mathbf{x}_k) \end{aligned} \quad (2)$$

The optimal path passing from $\mathbf{p}_0$ to $\mathbf{p}_N$ can be found by iterating for all k ($1 \le k \le N$) in Eq. (2).

Structural alignment algorithm

First, the sequence alignment was performed for the proteins. The function d in Eq. (1) was defined by Eq. (3):

$$d(a,b) = \begin{cases} 0 & a = b \\ 1 & (a \ne b) \quad \text{and} \quad (a \ne - \quad \text{and} \quad b \ne -) \\ w & a = - \quad \text{or} \quad b = - \end{cases} \quad (3)$$

where a and b are amino acids and w is called the gap-penalty, which is set at 1.5 in this article. Next, the superposition of Cα atoms was performed for all pairs of amino acids except those that had gaps. The alignment procedure was performed for overlapped proteins using the function d, which is defined by Eq. (4):

$$d(a_i\, b_j) = \begin{cases} \|\mathbf{x}_i - \mathbf{y}_j\| & (a_i \ne -) \quad \text{and} \quad (b_j \ne -) \\ w_{\text{str}} & \text{else} \end{cases} \quad (4)$$

where $\|\bullet\|$ means norm, and $\mathbf{x}_i$ and $\mathbf{y}_j$ are the coordinates of Cα atoms for amino acid residues a_i and b_j, respectively. w_{str} is the gap penalty. Here, we set $w_{\text{str}} = 6.0$. From the result of this alignment, the r.m.s.d. (r_0) was calculated for the pairs of Cα atoms, excluding all pairs having gaps. Next, proteins were superimposed for paired residues having a distance of less than r_0 between two Cα atoms. Overlapped proteins were realigned using Eq. (4). This process was repeated until the r.m.s.d. for the Cα atoms converged or the number of overlapping steps was greater than 30, and a final solution of the structural alignment based on superposition of Cα atoms was obtained. The result of this alignment does not depend on the type of amino acid.

Answer sequences were made using the preceding structural alignment. In homology modeling, it is important to make residues in SCRs corresponding because residues shifted in SCRs cause fatal errors in the predicted structure. The preceding alignment method was able to provide what we call equivalent residues belonging to the SCRs. The results for equivalent residues in variable regions (VRs) were unreliable, because the distance between Cα atoms was too great to determine corresponding residues. In this article, we used the structural alignment mentioned above to make answer sequences for homology modeling.

Initial matrix

To obtain the initial matrix, first the probability of exact correspondence of each pair of amino acids was calculated from the results of structural alignment, as described below. The probability for the pair of amino acids m and n is denoted by p_{mn}. The initial matrix is then defined by Eq. (5):

$$M_0(m,n) = -\log(\mathbf{K}_p p_{mn}) \quad (5)$$

where K_p, which was set at 0.5, is a scaling constant value for each element.

Improvement of sequence alignment

The sequence alignment was performed using the initial matrix. Alignment by Eq. (2) tended to match the same amino acids and to scatter the gaps, because the increasing distance produced by inserted gaps is larger than that for amino acid pairs. We therefore redefined Eq. (2) by putting continuous

gaps in the target sequences. It was reported that alignment allowing for continuous gaps often gives more accurate results.[12]

Let us consider an optimal path passing through point $(i_k - 1, j_k)$. When the optimal path reaches $(i_k - 1, j_k)$ in the hth order, the amino acid pair corresponding to this point is given by $[a_{(i_k-1)_h}, b_{(j_k)_h}]$. The amino acid pair $[a_{(i_k)_l}, b_{(j_k-1)_l}]$ is given analogously. Equation (2) is then redefined as Eq. (6):

$$D(\mathbf{p}_k) = \min\left\{ \begin{array}{l} D((i_k - 1, j_k)) + E(b_{(j_k)_l})d(\mathbf{x}_k) \\ D((i_k - 1, j_k - 1)) + d(\mathbf{x}_k) \\ D((i_k, j_k - 1)) + E(a_{(i_k)_h})d(\mathbf{x}_k) \end{array} \right\} \quad (6)$$

where the function E is an operation to insert continuous gaps, which is defined by

$$E_g(a) = \left\{ \begin{array}{ll} 1 & a \neq - \\ K_g & a = - \end{array} \right. \quad (7)$$

Here, K_g is a constant value. If the value of K_g is determined to be less than 1.0, continuous gaps are inserted in the process of alignment. We varied the value from 0.3 to 1.0 in 0.1 increments and used the value having the best accuracy for the initial matrix.

Optimization of the similarity matrix by the Monte Carlo method

The amino acid similarity matrix was optimized by the Markov chain Monte Carlo method simulated at low temperature. Entropy in the information theory was applied to the objective function in the optimizing procedure. The definition of entropy is given in the next section.

The matrix was perturbed after repeating the optimizing procedure n times, which is denoted by M_n. For arbitrary i and j, the element $M_{n+1}(i, j)$ of the matrix was generated by Eq. (8):

$$M_{n+1}(i, j) = M_n(i, j) - \delta \quad (8)$$

where δ is a random value from -1.0 to 1.0. Sequence alignment was performed with M_{n+1}. The objective function, F_{n+1}, as described in the following section, was calculated on the basis of the results of sequence and structural alignment (Figure 1). The matrix was optimized by the Monte Carlo algorithm until the objective function F_{n+1} converged.

The solutions are different even if the Monte Carlo method is performed under the same conditions. Several optimizations were made, and the similarity matrix giving the small number of incorrect residues in comparison with the structural alignment was selected and determined.

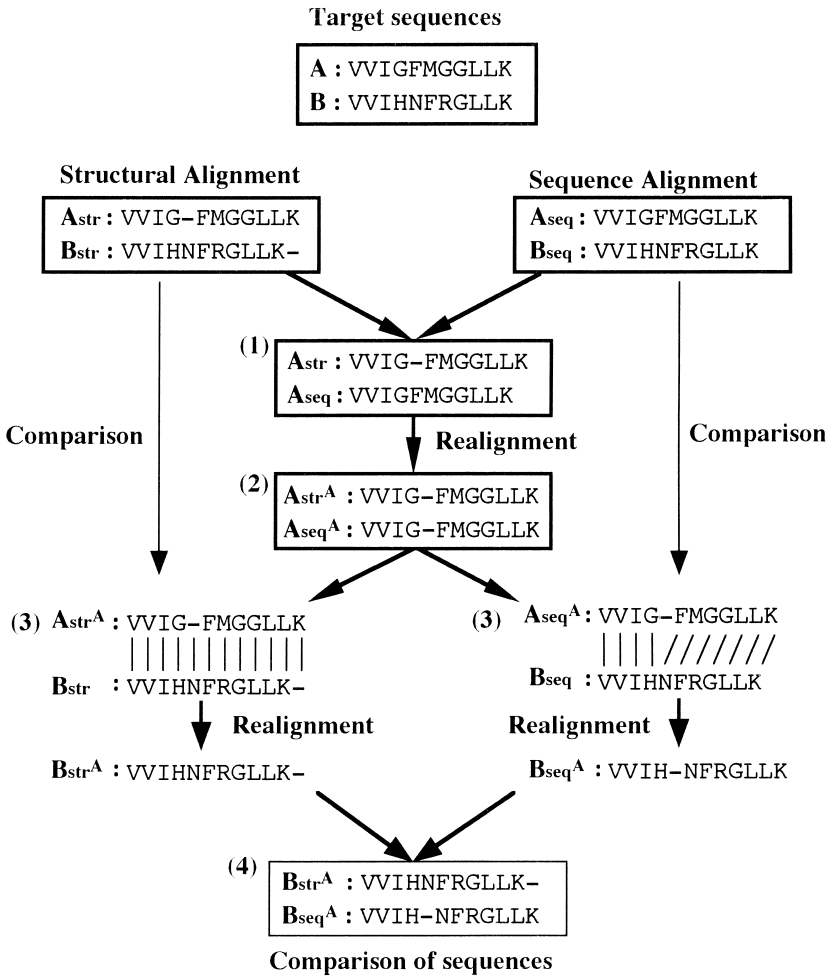

Figure 1. Comparison of aligned sequences. Structural and sequence alignments were performed for two target protein sequences, A and B. A_{str} and A_{seq} obtained from the structural and sequence alignments, respectively, were realigned to obtain the same amino acid pairs at equivalent positions (1). A gap was inserted between G and F residues in A_{seq}, and the results of the sequence pair are denoted by A^A_{str} and A^A_{seq} (2). The inserted gaps are represented by open dashes. For the sequences of A^A_{str} and A^A_{seq}, B_{str} was realigned for each corresponding amino acid residue of A^A_{str} and B_{str} to make it the same as A_{str} and B_{str}, and we obtained a sequence pair denoted by B^A_{str} and B^A_{seq}, in which a gap was inserted between the H and N residues (3). The final sequences of B^A_{str} and B^A_{seq} are compared (4). The comparison of A_{str} and A_{seq} is realigned analogously.

Probabilistic entropies for aligned sequences

Let Ω denote a set of twenty amino acids and a gap. Here, for the realigned kth pair of sequences Seq. X_k and Seq. Y_k, i.e., the pairs B^A_{str} and B^A_{seq} or A^B_{str} and A^B_{seq} in Figure 1, the occurrence probability for an amino acid residue ω ($\in \Omega$) in Seq. X_k is denoted by $p(\omega)$. Then the complete event system of X for the probability p is defined by Eq. (9):

$$\begin{pmatrix} X \\ p \end{pmatrix} = \begin{pmatrix} - & A & C & \cdots & Y \\ p(-) & p(A) & p(C) & \cdots & p(Y) \end{pmatrix} \quad (9)$$

Similarly, the complete event system of Y for the probability q, which is defined for the amino acid residues in Seq. Y_k, is defined by Eq. (10):

$$\begin{pmatrix} Y \\ q \end{pmatrix} = \begin{pmatrix} - & A & C & \cdots & Y \\ q(-) & q(A) & q(C) & \cdots & q(Y) \end{pmatrix} \quad (10)$$

The compound event system for Seq. X_k and Seq. Y_k is defined by Eq. (11):

$$\begin{pmatrix} X \times Y \\ r \end{pmatrix} = \begin{pmatrix} (-,-) & (-,A) & \cdots & (W,Y) & (Y,Y) \\ r(-,-) & r(-,A) & \cdots & r(W,Y) & r(Y,Y) \end{pmatrix} \quad (11)$$

Then, the Shannon entropy $S(X, Y)$ for the compound event system is given by Eq. (12):

$$S(X, Y) = -\sum_{\omega,\sigma\in\Omega} r(\omega, \sigma)\log r(\omega,\sigma) \quad (12)$$

and the mutual entropy $I(X, Y)$ is given by Eq. (13):

$$I(X, Y) = \sum_{\omega,\sigma\in\Omega} r(\omega, \sigma)\log \frac{r(\omega, \sigma)}{p(\omega)q(\sigma)} \quad (13)$$

Mutual entropy is the amount of information exchanged between sets X and Y.

Objective function

The objective function, F, was defined by entropy in the information theory,[33] and it was calculated for the pair of realigned sequences as shown in Figure 1. For these sequences, Shannon entropy, $S(X, Y)$, and mutual entropy, $I(X, Y)$, were calculated. The function of $f_k(X, Y)$ is defined by Eq. (14):

$$f_k(X, Y) = 1 - \frac{I(X, Y)}{S(X, Y)} \quad (14)$$

The second term is the transmitted ratio of information between X and Y (Figure 2). The smaller the value of function f_k, the more similar the sequences. The shape of function f_k for sequence homology is shown in Figure 3. The curve of f_k is flat for the range of low sequence homology, and is sharp for the range of high sequence homology.

Moreover, the objective function of F is defined as

$$F = \frac{1}{N}\sum_k f_k(X, Y) \quad (15)$$

where N is the number of the realigned sequence pairs for the data set. When the result of sequence alignment is different from that of the structural one, an optimal solution from various matrices in the neighborhood of the matrix may be searched. On the other hand, when the result of sequence alignment is close to that of the structural alignment, an optimal solution may easily be found along the sharp curvature.

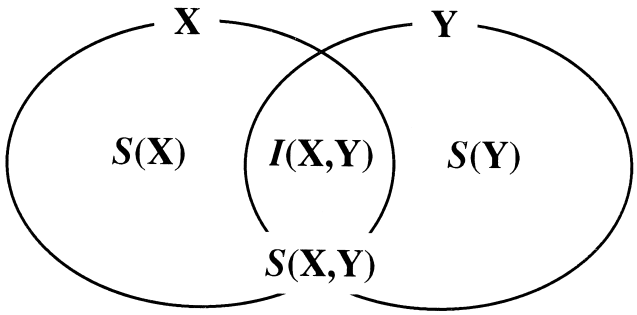

Figure 2. Conception of probabilistic based entropy, For event systems X *and* Y, S*(*X*) and* S*(*Y*) represent Shannon entropy in the areas of* X *and* Y*, respectively. Shannon entropy* S*(*X,Y*) for the compound events of* X *and* Y *is represented in the surrounding areas for* X *and* Y*. Mutual entropy* I*(*X,Y*) is represented in the jointed areas of* X *and* Y*.*

Sampling proteins

We used only the protein files derived from X-ray crystallographic studies in the Protein Data Bank[34] (PDB) released in January 1997. First, we excluded proteins having homol-

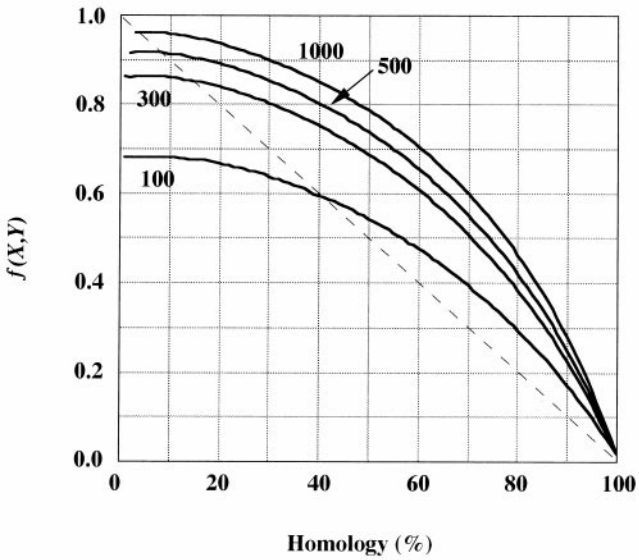

Figure 3. The relationship between entropy ratio and sequence homology. Four sets of 10 million random sequences pairs with sequence lengths of 100, 300, 500 and 1000 residues were prepared. For each set of sequences, the function f_k, which is defined by probabilistic entropy in information theory, was calculated for each sequence pair, and their values were averaged in the range of homology. The curves are plotted against the range of homology values.

ogy greater than 90%. The protein having the highest resolution of X-ray crystallographic studies among the remaining highly similar proteins was regarded as the standard protein, and the sequence of the standard protein was used as the representative of highly similar proteins in the following analysis. As a result, we obtained 1 338 standard proteins. Second, we performed structural alignment for standard proteins and chosen protein pairs having more than twenty SCR residues. Finally, we obtained 1 058 aligned sequence pairs generated from 350 proteins (see Appendix). These protein pairs were not biased for homology or the ratio of SCR residues against the length of aligned sequences. Moreover, there was no correlation between the homology and the ratio of SCR residues (Figure 4).

RESULTS AND DISCUSSION

Structurally based amino acid similarity matrix

The optimization of amino acid matrix was carried out by the following procedures. First, we obtained an initial matrix from structural alignment, and sequence alignment with initial matrix was performed to determine the K_g value. Next, the initial matrix was optimized by the Markov chain Monte Carlo method as described above.

Table 1 shows the initial matrix derived from the results of structural alignment. The values of the elements for (gap, amino acid) pairs and (amino acid, amino acid) pairs are not very different, and, moreover, some gap and amino acid pairs had lower values than amino acid pairs, e.g. the relationship between (A, N) and (A, H), although the value of gap penalty has usually been determined as the largest value in the matrix. The relative sizes of these values are why the structural alignment in this work was performed without distinguishing the type of amino acids. If the structural alignment distinguishes amino acid type, the distance between the same amino acids may be smaller than the distance between different amino acids. In particular, the distance between gap and amino acid is larger than that of initial values in Table 1. The number of gaps inserted in the alignment will be decreased. Therefore, when we aligned sequences having long insertions and deletions and having sequence identity greater than 30%, sequences in the part of insertions and deletions might be incorrect from answer sequences. In this work, we did not distinguish amino acid type in structural alignment to avoid such cases. As a result, the values of gap and amino acid pairs were relatively small when compared with those for amino acid pairs and gap penalties in other matrices, e.g., PAM,[25] BLOSUM,[26] etc.

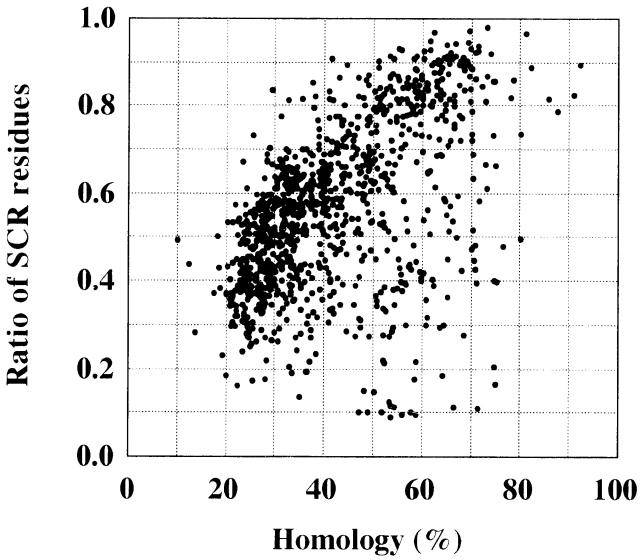

Figure 4. Correlation graph between homology value and the ratio of SCR residues. Homology was calculated for each structural alignment. The ratio of SCR residues was also calculated for the length of aligned sequences by structural alignment.

In order to determine K_g, sequence alignment with the initial matrix was performed for various K_g values from 0.3 to 1.0. The average number of incorrect amino acid residues in aligned sequences is shown in Table 2. As can be seen, the K_g value greatly influences sequence alignment. When we took a very small K_g ($K_g = 0.4$ and 0.3), the results of sequence alignment differed greatly from those of structural alignment. Since the best accuracy was obtained by $K_g = 0.6$, this value was used in the optimization process.

Figure 5a shows a graph of the values of the objective function versus the Monte Carlo steps. Each Monte Carlo step was the state in which all elements in the matrix were calculated. The objective function decreased for 10 steps and converged uniformly after that step. Figure 5b shows the r.m.s.d. value between the generated and initial matrix. The r.m.s.d. value increased for 10 steps and converged uniformly after that step. The final matrix was different from the initial matrix by about 0.8 in terms of the r.m.s.d. In higher steps, the slopes of the curves in both figures were close to zero, and we can regard the optimizing process as convergent.

Figure 5c and d shows the average number of incorrect residues in aligned sequences for all residues and SCRs, respectively, for each step. These graphs correlate with the changes of the objective function and r.m.s.d. in Figure 5a and b. The number of incorrect residues in the initial matrix was 26.4 and 4.8 for all residues and SCRs, respectively. In the final matrix, these results improved to 22.6 and 3.7. This shows that the matrix optimization operated effectively.

The upper and lower triangles in Table 3 show maximizing and minimizing algorithms of the final matrices, respectively. The final matrix maximizing algorithm $M_{max}(i, j)$ was calculated by the equation $M_{max}(i, j) = -M_{min}(i, j) + \alpha$, where $M_{min}(i, j)$ is the matrix minimizing algorithm and α is an arbitrary constant. We set $\alpha = 4.0$ based on the comparison of sequence and structural alignments. There was also a small difference between the values of the elements in the matrices when compared with those of PAM and BLOSUM. Almost none of the elements in the final matrix had significant changes from the initial matrix (Table 1), although the number of incorrect residues improved after alignment using the initial matrix. The value between methionine (M) and tryptophan (W) differed the most, at 2.33. The preceding results mean that even small changes in an element of the matrix yield great differences in the results of alignment. Therefore it should be noted that it is difficult to optimize the matrix using only an analytical approach.

Figure 6a and b shows the results of alignment using our method with the final matrix. The aligned sequences in Figure 6a have similar sequence identity between structural and se-

Table 1. Initial matrix derived from structural alignment based on superposition of Cα atoms

	A	R	N	D	C	Q	E	G	H	I	K	L	M	F	P	S	T	W	Y	V
A	4.23																			
R	7.26	5.00																		
N	6.88	7.26	4.91																	
D	6.87	7.74	6.31	4.46																
C	7.70	9.06	8.83	9.30	4.89															
Q	6.98	7.12	7.32	7.32	9.21	5.18														
E	6.69	7.47	7.15	6.26	9.60	6.55	4.79													
G	6.07	7.42	6.66	6.61	8.42	7.56	7.06	3.66												
H	8.04	7.97	7.48	8.22	9.89	7.88	8.10	8.07	5.41											
I	6.98	8.19	8.15	8.34	9.05	8.07	8.08	8.33	8.93	4.67										
K	6.65	7.56	7.65	8.02	8.61	7.43	7.81	7.70	8.22	5.70	4.04									
L	6.61	6.19	6.76	6.94	9.18	6.53	6.51	7.01	7.80	7.87	7.35	4.68								
M	7.63	8.81	8.78	9.40	9.42	8.13	8.72	8.67	9.38	7.23	6.49	8.39	6.13							
F	7.76	8.66	8.30	8.91	8.93	8.39	8.83	8.35	8.46	7.03	6.44	8.59	7.98	4.78						
P	6.72	7.85	7.72	7.52	9.39	7.79	7.56	7.33	8.52	8.12	7.70	7.32	9.50	8.67	4.43					
S	5.84	7.01	6.24	6.44	8.18	6.96	6.74	6.26	7.94	7.67	7.25	6.57	8.16	8.02	6.67	4.22				
T	6.26	7.16	6.67	6.97	8.16	7.06	6.95	7.00	8.20	7.06	6.94	6.53	8.00	8.03	7.14	5.56	4.37			
W	9.11	8.91	9.20	9.95	10.54	9.70	9.60	8.60	9.72	8.60	8.17	9.47	9.51	7.49	9.85	8.89	8.97	5.33		
Y	7.72	7.94	7.75	8.12	9.60	8.34	8.17	7.87	7.61	7.70	7.21	7.92	8.64	6.30	8.45	7.64	7.58	7.50	4.73	
V	6.04	7.57	7.64	7.79	8.14	7.50	7.43	7.50	8.31	5.38	5.68	7.18	7.41	7.10	7.77	6.98	6.36	8.84	7.41	4.12
–	6.36	7.07	6.76	6.48	8.62	7.17	6.64	6.14	7.97	7.51	7.00	6.56	8.40	7.71	6.58	6.18	6.67	8.67	7.33	7.11

quence alignments; and the result of sequence alignment is similar to that of the structural alignment. In particular, gaps were inserted at similar positions. In homology modeling, we can obtain an accurate model for such cases. On the other hand, the results of structural and sequence alignment in Figure 6b show that sequence identities are less than 30%. In these cases, the positions of inserted gaps were different between structural and sequence alignments. In particular, for sequence alignment, the positions of inserted gaps around C-terminal in 1VDR−A are different from in the structural alignment. In such cases, we cannot produce an accurate model using homology modeling. From these results, we believe that our alignment will yield an accurate alignment for pairs of sequences having greater than 30% sequence identity. However, it is difficult to achieve an accurate alignment for pairs of sequences having less than 30% sequence identity.

Table 2. Accuracy of alignment for K_g values

	Number of incorrect residues	
K_g	All	SCRs
0.3	203.9	121.7
0.4	132.4	67.0
0.5	30.5	6.3
0.6	26.3	4.8
0.7	26.5	4.9
0.8	27.3	5.2
0.9	28.7	5.8
1.0	30.1	6.5

Comparison with other matrices

To evaluate our amino acid similarity matrix, we tried other matrices using a general alignment algorithm and compared the results. Three matrices, PAM250,[25] BLOSUM62,[26] and GONNET,[27] were used, because they are popular in sequence analysis. Furthermore, the JOHNSON[32] method, which was obtained from tertiary structures of proteins, was also used. N/W alignment[5] and S/W alignment[7] were used as the alignment algorithms. The gap penalty, Gp, was defined by the equation $Gp = I + E(l - 1)$, where l is each gap length, I is the fixed gap penalty, and E is the gap penalty for extension. We determined the values as follows. For PAM250, BLOSUM62, and GONNET, I was varied from 5.0 to 18.0 by 1.0 increments, and E was also varied from 0.0 to 6.0 by 0.5 increments. The alignment was performed in 182 combinations of the fixed and extension penalties. For JOHNSON, I was varied from 15.0 to 30.0 by 1.0 increments, and E was varied from 0.0 to 6.0 by 0.5 increments. The alignment was performed in 208 combinations of the fixed and extension penalties. The values of I and E having the most accurate results were chosen for the PAM250, BLOSUM62, GONNET, and JOHNSON matrices.

Color Plate 1A and B shows the variability of the number of incorrect residues against the combinations of fixed and extension penalties. The graphs of S/W and N/W alignments with the same matrix show similar curvature, but the results of N/W alignment have deeper wells than those of S/W alignment. The minimum number of incorrect residues can be seen in each graph. The curvature of the alignment with GONNET is relatively flat-shaped for the entire area. On the other hand, the alignment with JOHNSON shows a large number of incorrect residues at $E = 0.0$ but a flat curvature in the range of $0.5 \leq$

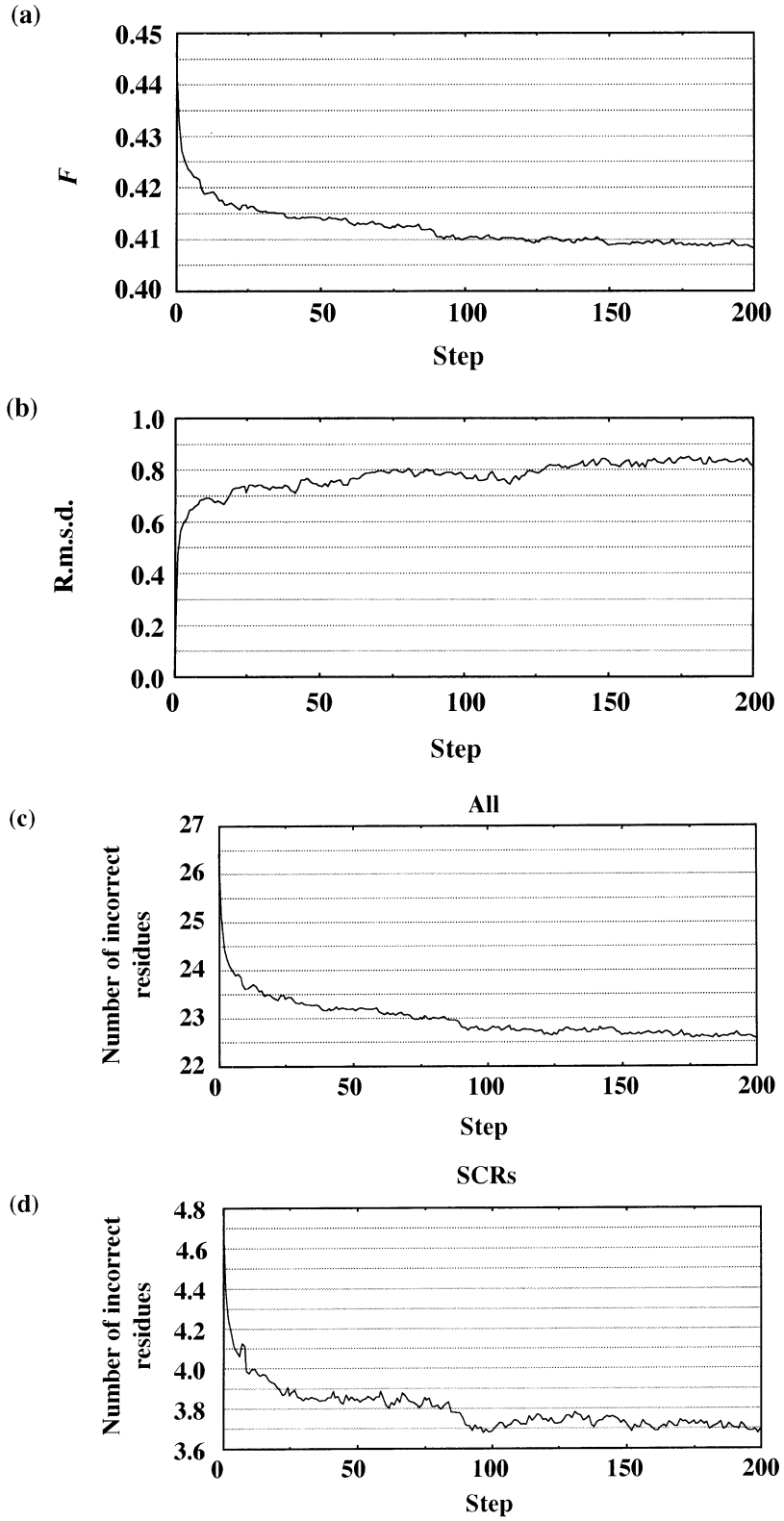

Figure 5. Perturbation and accuracy of each step. The value of the objective function F for each step is described in (a), and the r.m.s.d. for the generated matrices from an initial matrix is shown in (b). The number of incorrect residues for all residues and SCR residues is shown in (c) and (d), respectively.

$E \leq 6.0$, which, of all of the plotted methods, has the minimum number of incorrect residues. In particular, N/W alignment (Color Plate 1B) with JOHNSON has the largest area of accurate results, where the number of incorrect residues ranges from 24 to 24.5.

In Table 4, the most accurate result of the 182 alignment combinations for PAM250, BLOSUM62, and GONNET, and the 208 alignment combinations for JOHNSON, is shown for eight combinations of alignment algorithms and matrices. The accuracy of the S/W alignment in the three matrices was

Table 3. Final matrix

	A	R	N	D	C	Q	E	G	H	I	K	L	M	P	F	S	T	W	Y	V	–	
	−0.84	−3.00	−3.14	−2.92	−3.66	−3.01	−2.68	−1.81	−3.73	−3.08	−2.70	−2.71	−3.96	−3.89	−2.76	−1.61	−2.20	−5.04	−3.53	−1.91	−2.47	A
		−2.23	−3.30	−3.30	−4.34	−3.47	−3.53	−3.49	−4.08	−3.66	−3.92	−2.61	−3.83	−4.13	−3.93	−2.89	−2.82	−5.40	−4.40	−3.40	−3.81	R
A	4.84		−0.75	−2.25	−4.59	−3.11	−2.52	−2.76	−2.84	−3.60	−3.66	−2.85	−3.42	−4.18	−3.99	−2.45	−2.38	−4.63	−3.01	−3.22	−2.74	N
R	7.00	6.23		−1.02	−4.30	−3.27	−2.42	−2.35	−4.06	−4.34	−3.11	−2.94	−4.19	−5.53	−3.30	−1.68	−2.47	−5.75	−3.17	−3.12	−2.45	D
N	7.14	7.30	4.75		0.81	−4.31	−6.21	−3.53	−5.25	−5.02	−3.68	−4.54	−4.74	−4.88	−4.25	−4.24	−3.48	−7.97	−6.61	−4.11	−5.45	C
D	6.92	7.30	6.25	5.02		−2.47	−2.77	−3.63	−3.73	−3.86	−3.41	−2.65	−4.54	−4.59	−4.08	−2.95	−3.04	−4.30	−3.33	−3.46	−4.54	Q
C	7.66	8.34	8.59	8.30	3.19		−1.42	−3.44	−3.81	−4.28	−3.42	−2.68	−4.09	−4.22	−3.82	−2.61	−2.82	−3.92	−3.52	−3.32	−3.47	E
Q	7.01	7.47	7.11	7.27	8.31	6.47		−0.20	−3.73	−4.03	−4.04	−2.75	−4.19	−4.14	−3.17	−2.30	−2.73	−3.15	−3.50	−3.29	−2.14	G
E	6.68	7.53	6.52	6.42	10.21	6.77	5.42		−2.06	−4.22	−4.08	−3.57	−5.67	−4.15	−3.49	−3.02	−3.99	−5.67	−2.76	−3.53	−4.74	H
G	5.81	7.49	6.76	6.35	7.53	7.63	7.44	4.20		−1.53	−1.47	−4.07	−3.39	−3.22	−3.24	−3.22	−3.09	−3.71	−3.38	−1.68	−4.20	I
H	7.73	8.08	6.84	8.06	9.25	7.73	7.81	7.73	6.06		0.00	−3.22	−2.99	−2.64	−2.64	−2.42	−2.21	−2.84	−2.63	−1.76	−3.05	K
I	7.08	7.66	7.60	8.34	9.02	7.86	8.28	8.03	8.22	5.53		−1.49	−4.22	−4.84	−3.48	−2.62	−2.34	−5.48	−3.34	−2.67	−2.69	L
K	6.70	7.92	7.66	7.11	7.68	7.41	7.42	8.04	8.08	5.47	4.00		−2.65	−4.46	−4.80	−3.46	−3.31	−3.18	−3.86	−3.17	−5.78	M
L	6.71	6.61	6.85	6.94	8.54	6.65	6.68	6.75	7.57	8.07	7.22	5.49		−1.69	−4.80	−4.29	−3.73	−2.91	−1.96	−3.12	−5.80	P
M	7.96	7.83	7.42	8.19	8.74	8.54	8.09	8.19	9.67	7.39	6.99	8.22	6.65		−1.65	−2.78	−2.65	−6.36	−4.10	−3.54	−2.57	F
P	7.89	8.13	8.18	9.53	8.88	8.59	8.22	8.14	8.15	7.22	6.64	8.84	8.46	5.69		−0.60	−1.50	−3.58	−3.33	−2.82	−2.27	S
F	6.76	7.93	7.99	7.30	8.25	8.08	7.82	7.17	7.49	7.24	6.64	7.48	8.80	8.80	5.65		−0.95	−3.56	−2.81	−2.16	−2.70	T
S	5.61	6.89	6.45	5.68	8.24	6.95	6.61	6.30	7.02	7.22	6.42	6.62	7.46	8.29	6.78	4.60		−0.97	−2.55	−3.79	−4.70	W
T	6.20	6.82	6.38	6.47	7.48	7.04	6.82	6.73	7.99	7.09	6.21	6.34	7.31	7.73	6.65	5.50	4.95		−0.68	−3.24	−2.98	Y
W	9.04	9.40	8.63	9.75	11.97	8.30	7.92	7.15	9.67	7.71	6.84	9.48	7.18	6.91	10.36	7.58	7.56	4.97		−0.76	−3.09	V
Y	7.53	8.40	7.01	7.17	10.61	7.33	7.52	7.50	6.76	7.38	6.63	7.34	7.86	5.96	8.10	7.33	6.81	6.55	4.68		—	–
V	5.91	7.40	7.22	7.12	8.11	7.46	7.32	7.29	7.53	5.68	5.76	6.67	7.17	7.12	7.54	6.82	6.16	7.79	7.24	4.76		
–	6.47	7.81	6.74	6.45	9.45	8.54	7.47	6.14	8.74	8.20	7.05	6.69	9.78	9.80	6.57	6.27	6.70	8.70	6.98	7.09	—	
	A	R	N	D	C	Q	E	G	H	I	K	L	M	P	F	S	T	W	Y	Z	–	

Maximizing algorithm

Minimizing algorithm

Structural Alignment **A**

```
1FKB    : -GVQVETISPGDGRTFPKRGQTCVVHYTGMLEDGKKFDSSRDR-------NKPFKFMLGKQEVIRGWEEG
                  ::    ::: :       ::: : ::   ::               ::   :   :    ::::: :
1PBK    : PKYTKSVLKKGDKTNFPKKGDVVHCWYTGTLQDGTVFDTNIQTSAKKKKNAKPLSFKVGVGKVIRGWDEA

1FKB    : VAQMSVGQRAKLTISPDYAYGATGHPG-IIPPHATLVFDVELLKLE
             ::  :   : : : :   :::   : :    ::: : : : :::
1PBK    : LLTMSKGEKARLEIEPEWAYGKKGQPDAKIPPNAKLTFEVELVDID    (homology = 38.8%)
```

Sequence Alignment

```
1FKB    : G-VQVETISPGDGRTFPKRGQTCVVHYTGMLEDGKKFD------SSRDRN-KPFKFMLGKQEVIRGWEEG
                  ::    ::: :       ::: : ::   ::             : ::   :   :    ::::: :
1PBK    : PKYTKSVLKKGDKTNFPKKGDVVHCWYTGTLQDGTVFDTNIQTSAKKKKNAKPLSFKVGVGKVIRGWDEA

1FKB    : VAQMSVGQRAKLTISPDYAYGATGHPGI-IPPHATLVFDVELLKLE
             ::  :   : : : :   :::   : :    ::: : : : :::
1PBK    : LLTMSKGEKARLEIEPEWAYGKKGQPDAKIPPNAKLTFEVELVDID    (homology = 39.7%)
```

Structural Alignment **B**

```
1VDR_A : ELVSVAALAENRVIGRDGELPWPSIPADKKQYRSRIADDPVVLGRTTFESMRDDLP-GSAQIVMSRSERS
               :    :  :: :: :::    : :     :          : :: : ::              :     :
3DFR   : -TAFLWAQNRNGLIGKDGHLPW-HLPDDLHYFRAQTVGKIMVVGRRTYESFPKRPLPERTNVVLTHQEDY

1VDR_A : FSVDTAHRAASVEEAVDIAASLDAETAYVIGGAAIYALFQPHLDRMVLSRVPGEYEGDTYYPEWDAAEWE
                           :              ::: :    :      :      :   :  ::::
3DFR   : QAQGAVVVHDV-AAVFAYAKQHLDQELVIAGGAQIFTAFKDDVDTLLVTRLAGSFEGDTKMIPLNWDDFT

1VDR_A : LDAETDHE------GFTLQEWVRS-
                :          :    :
3DFR   : KVSSRTVEDTNPALTHTYEVWQKKA      (homology = 21.8%)
```

Sequence Alignment

```
1VDR_A : ELVSVAALAENRVIGRDGELPWPSIPADKKQYRSRIADDPVVLGRTTFESM-RDDLPGSAQIVMSRSERS
               :    :  :: :: :::    : :     :          : :: : ::      ::      :     :
3DFR   : TAFLWA-QNRNGLIGKDGHLPW-HLPDDLHYFRAQTVGKIMVVGRRTYESFPKRPLPERTNVVLTHQE-D

1VDR_A : FSVDTAHRAASVEEAVDIA-ASLDAETAYVIGGAAIYALFQPHLDRMVLSRVPGEYEGDT-----YYPEW
               :      :       :    :: :      ::: :    :      :      :   :  ::::
3DFR   : YQAQGAVVVHDVAAVFAYAKQHLDQELV-IAGGAQIFTAFKDDVDTLLVTRLAGSFEGDTKMIPLNWDDF

1VDR_A : DAAEWELDAETDHE-GFTLQEWVR-S
                  :          :    :
3DFR   : TKVSSRTVEDTNPALTHTYEVWQKKA      (homology = 25.9%)
```

Figure 6. Alignment using our method. The results of structural and sequence alignments for 1FKB (FK506 binding protein) and 1PBK (FKBP12 homologous domain of HFKBP25), and for 1VDR−A (dihydrofolate reductase) and 3DFR (dihydrofolate reductase, E.C.1.5.1.3) are shown. The sequence alignment was performed using our alignment algorithm with the optimized matrix.

slightly worse than that of the N/W alignment. Such a tendency was already reported by Vogt et al.[12] The best among the eight combination alignments was the N/W alignment with JOHNSON at $I = 24.0$ and $E = 2.0$. The average number of incorrect amino acid residues in aligned sequences was 24.1 and 4.3 for all residues and SCR residues, respectively. Finally, the accuracy of our alignment was slightly better than the best of the other methods.

However, it should be noted that our alignment was optimized for our data set while the others were developed

Table 4. Comparison with other methods for the small data set

Matrix	Alignment method[a]	I[b]	E[c]	The average number of incorrect residues	
				All	SCRs
PAM250	S/W	12.0	2.5	27.3	5.6
BLOSUM62	S/W	13.0	1.5	26.6	5.1
GONNET	S/W	12.0	1.0	26.1	5.0
JOHNSON	S/W	22.0	2.0	26.4	5.1
PAM250	N/W	15.0	2.0	25.4	5.1
BLOSUM62	N/W	15.0	1.0	24.6	4.6
GONNET	N/W	12.0	1.0	24.2	4.3
JOHNSON	N/W	24.0	2.0	24.1	4.3
This work					
Initial matrix				26.4	4.8
Optimized matrix				22.6	3.7

[a] S/W, Smith and Waterman algorithm; N/W, Needleman and Wunsch algorithm.
[b] The fixed gap penalty.
[c] The extension gap penalty. The values of I and E were optimized for the small data set.

independent of that set. Accordingly, the preceding conclusion would be supported if we are able to obtain similar results by using the larger data set instead of the data set mentioned in Methods, which we call the small data set. The larger data set was made in the same way as described in Methods from the PDB set released in October 1997. We obtained 3 518 aligned sequence pairs, which was about three times as large as the small data set. Aligned sequence pairs in the small data set were involved in the large data set.

Table 5 shows the average number of incorrect amino acid residues for the larger data set using the gap penalties of I and E from the small data set. The results of the large data set using the matrices in Tables 1 and 3 are described in the last two rows of Table 5. Minimum average values of incorrect residues for the optimized gap penalties of I and E are also shown in Table 5. The values of the gap penalties given from the small data set are different from those of the large data set. This means that the values of gap penalties depend on the size of the data set. Therefore, we should obtain the gap penalties for a data set as large as possible. The average number of incorrect residues in our method was the smallest compared with other alignment methods.

CONCLUSION

We obtained a structurally based amino acid similarity matrix and optimized it by a Monte Carlo method. The most significant difference from other reported matrices was the treatment of the gap penalty, as gaps were regarded as amino acid types. The average number of incorrect residues in our alignment was 22.6 and 3.7 for all residues and SCR residues, respectively. These values were slightly better than those obtained from other reported matrices. Therefore, we believe our method will become a helpful tool for amino acid alignment in homology modeling.

ACKNOWLEDGMENTS

This work was supported by a grant-in-aid for special project research from the Ministry of Education, Science, Sports and Culture of Japan.

Table 5. Comparison with other methods for the large data set

Matrix	Alignment method[c]	Gap penalties for small data set[a]				Minimum[b]			
		I[d]	E[e]	Average number of incorrect residues		I[d]	E[e]	Average number of incorrect residues	
				All	SCRs			All	SCRs
PAM250	S/W	12.0	2.5	22.5	3.7	15.0	1.0	22.3	3.7
BLOSUM62	S/W	13.0	1.5	21.9	3.4	14.0	1.5	21.9	3.4
GONNET	S/W	12.0	1.0	21.7	3.3	12.0	1.0	21.7	3.3
JOHNSON	S/W	22.0	2.0	21.7	3.4	23.0	2.5	21.7	3.4
PAM250	N/W	15.0	2.0	21.0	3.3	17.0	1.0	20.9	3.3
BLOSUM62	N/W	15.0	1.0	20.7	3.1	15.0	1.0	20.7	3.1
GONNET	N/W	12.0	1.0	20.5	2.9	15.0	0.5	20.4	3.0
JOHNSON	N/W	24.0	2.0	20.3	3.0	24.0	2.5	20.3	2.9
This work									
Initial matrix				22.9	3.6				
Optimized matrix				20.0	2.8				

[a] The results using gap penalties, I and E, in Table 4.
[b] The results show the minimum average number of incorrect residues for all residues, and the values of I and E were optimized for the large data set.
[c] S/W, Smith and Waterman algorithm; N/W, Needleman and Wunsch algorithm.
[d] The fixed gap penalty.
[e] The extension gap penalty.

REFERENCES

1 Greer, J. Comparative modeling methods: Application to the family of the mammalian serine proteases. *Proteins,* 1990, **7,** 317–334.
2 Matsuo, Y., and Nishikawa, K. Protein structural similarities predicted by a sequence-structure compatibility method. *Protein Sci.* 1994, **3,** 2055–2063.
3 Yoneda, T., Komooka, H., and Umeyama, H. A computer modeling study of the interaction between tissue factor pathway inhibitor and blood coagulation factor Xa. *J. Protein Chem.* 1997, **16,** 597–605.
4 Ogata, K., and Umeyama, H. The role played by environmental residues on side-chain torsional angles within homologous families of proteins: A new method of side-chain modeling. *Proteins* 1998, **31,** 355–369.
5 Needleman, S.B., and Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 1970, **48,** 443–453.
6 Sellers, P.H. On the theory and computation of evolutionary distance. *SIAM J. Appl. Math.* 1974, **26,** 787–793.
7 Smith, T.F., and Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* 1981, **147,** 195–197.
8 Gotoh, O. An improved algorithm for matching biological sequences. *J. Mol. Biol.* 1982, **162,** 705–708.
9 Ohya, M., and Uesaka, Y. Amino acid sequences and DP matching. *Inform. Sci.* 1992, **63,** 139–151.
10 Miyazawa, S. A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng.* 1994, **8,** 999–1009.
11 Taylor, W.R. An investigation of conservation-biased gap-penalties for multiple protein sequence alignment. *Gene* 1995, **165,** GC27–GC35.
12 Vogt, G., Ttzold, T., and Argos, P. An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. *J. Mol. Biol.* 1995, **249,** 816–831.
13 Gotoh, O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structure alignment. *J. Mol. Biol.* 1996, **264,** 823–838.
14 Taylor, W.R., and Orengo, C.A. Protein structure alignment. *J. Mol. Biol.* 1989, **208,** 1–22.
15 Sali, A., and Blundell, T.L. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* 1990, **212,** 403–442.
16 Russell, R.B., and Barton, G.J. Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins* 1992, **14,** 309–323.
17 Gibrat, J.F., Madej, T., and Bryant, S.H. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* 1996, **6,** 377–385.
18 Holm, L., and Sander, C. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* 1998, **26,** 316–319.
19 Toh, H. Introduction of a distance cut-off into structural alignment by the double dynamic programming algorithm. *Comput. Appl. Biosci.* 1997, **13,** 387–396.
20 Kanaoka, M., Kishimoto, F., Ueki, Y., and Umeyama, H. Alignment of protein sequences using the hydrophobic core scores. *Protein Eng.* 1989, **2,** 347–351.
21 Swindells, M.B. A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci.* 1995, **4,** 93–102.
22 Ohya, M., Miyazaki, S., and Ogata, K. On multiple alignment of genome sequences., *IEICE Trans. Commun.* 1992, **E75-B,** 453–457.
23 Ishikawa, M., Toya, T., Hoshida, M., Nitta, K., Ogiwara, A., and Kanehisa, M. Multiple sequence alignment by parallel simulated annealing. *Comput. Appl. Biosci.* 1993, **9,** 267–274.
24 Notredame, C., and Higgins, D. SAGA: Sequence alignment by genetic algorithm. *Nucleic Acids Res.* 1996, **24,** 1515–1524.
25 Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. A model of evolutionary change in proteins. In: *Atlas of Protein Sequence and Structure* (Dayhoff, M.O., ed.), Nat. Biomed. Res. Found., London, Vol. 5, Suppl. 3. 1978, pp. 345–352.
26 Henikoff, S., and Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* 1992, **89,** 10915–10919.
27 Gonnet, G.H., Cohen, M.A., and Benner, S.A. Exhaustive matching of the entire protein sequence database. *Science* 1992, **256,** 1443–1445.
28 Madej, T., Gibrat, J.F., and Bryant, S.H. Threading a database of protein cores. *Proteins* 1995, **23,** 356–369.
29 Ota, M., and Nishikawa, K. Assessment of pseudo-energy potentials by the best-five test: A new use of the three-dimensional profiles of proteins. *Protein Eng.* 1997, **10,** 339–351.
30 Chiu, T.L., and Goldstein, R.A. Optimizing energy potentials for success in protein tertiary structure prediction. *Folding Design* 1998, **3,** 223–228.
31 Risler, J.L., Delorme, M.O., Delacroix, H., and Henaut, A. Amino acid substitutions in structurally related proteins. A pattern recognition approach. *J. Mol. Biol.* 1988, **204,** 1019–1029.
32 Johnson, M.S., and Overington, J.P. A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J. Mol. Biol.* 1993, **233,** 716–738.
33 Umegaki, H., and Ohya, M. *Entropies in probabilistic systems.* Kyoritsu Publishing Company, 1983.
34 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Jr., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. The Protein Data Bank: A computer based archival file for micromolecular structures. *J. Mol. Biol.* 1977, **112,** 535–542.

APPENDIX

Answer sequences were generated by the following proteins, whose ID codes are 1LZ1, 1LMQ, 135L, 1ALC, 1HCB, 1DMX−A, 1COT, 1CCR, 1C2R−A, 1CXC, 5CYT−R, 1YCC, 3C2C, 1CRY, 2IMN, 1JHL−L, 1LVD, 1FGV−L, 1IVL−A, 1VFA−A, 1DVF−C, 2RHE, 5PTI, 1KNT, 1AAP−A, 1BUN−B, 1ARS, 2CST−A, 1DIF−A, 1FMB, 2RSP−A, 1RDS, 1FUS, 1ABR−A, 1MRG, 7RSA, 1AGI, 1ONC, 1AHT−H, 1PFX−C, 4PTP, 1LMW−B, 1DST, 1HYL−A, 1HCG−A, 1MCT−A, 1TRY, 1FON−A, 3RP2−A, 1GCT−A, 2TBS, 1SGT, 1ELT, 1PPF−E, 1TON, 1ELG, 1FUJ−A, 1ABO−A, 1SEM−A, 1SHF−A, 1SHG, 1ABM−A, 3MDS−A, 1IDS−A, 3SDP−A, 1ADS, 1RAL, 2CTX, 1NTN, 1NXB, 1CDT−A, 1KBA−A, 1TGX−A, 1FAS, 1CSE−I, 1YPC−I, 8ATC−A, 2AT2−A, 1FLR−L, 1OPG−L, 8FAB−A, 1CLY−L, 1EAP−A, 2FB4−L,

1BAF−H, 1HIL−B, 6FAB−H, 2FGW−H, 1GIG−H, 1EAP−B, 1VGE−H, 1GGB−H, 1MRD−H, 8FAB−B, 1GAF−H, 1OPG−H, 2FBJ−H, 1IND−H, 1FLR−H, 1TET−H, 7FAB−H, 2ACH−A, 1OVA−A, 1ANT−L, 1HLE−A, 2ACT, 1PPO, 1ADL, 1PMP−A, 1CRB, 1FTP−A, 1IFC, 1HMR, 1OPA−A, 1CBS, 3ADK, 1UKZ, 2MSB−A, 1HUP, 1GSE−A, 1GLQ−A, 1AGX, 3ECA−A, 1CYD−A, 1HDC−A, 2RAN, 1ANN, 1AK2, 1AKY, 2AK3−A, 1ALL−A, 1ALL−B, 1CPC−A, 2ALP, 1SGP−E, 1APA, 1PAF−A, 1PPL−E, 1HRN−A, 1HTR−B, 1EPM−E, 1SMR−A, 2APR, 1MPP, 3PSG, 3CMS, 1PCA, 1CPB, 2ASR, 2LIG−A, 1POD, 1PPA, 1POA, 4BP2, 1PP2−R, 1PSJ, 1BUN−A, 1BAB−A, 1HDS−B, 1HBH−A, 1BAB−B, 1HBH−B, 1IVD, 1NNC, 1NMB−N, 1NSC−A, 1EPT−C, 1MTN−C, 1BBT−3, 1COV−3, 2MEV−3, 1TME−3, 1R1A−3, 1PVC−3, 1R09−3, 2BBK−H, 2MAD−H, 1TYS, 1TSY, 1XNB, 1XYN, 1BET, 1BND−B, 1BKF, 1PBK, 2BLM−A, 1BTL, 2BPA−1, 1GFF−1, 1BTM−A, 7TIM−A, 1TPH−1, 1TPF−A, 1TRE−A, 1LHS, 1MYT, 1CAA, 6RXN, 1RDG, 8RXN−A, 1CBG, 1PBG−A, 1CGO, 1RCP−A, 1CGT, 1CIU, 4CPV, 1RTP−1, 1PVA−A, 5PAL, 1RRO, 2CYR, 1CZJ, 1CEA−A, 2HPP−P, 1PML−A, 1CER−O, 1GAD−O, 1GD1−O, 1HDG−O, 1GYP−A, 4GPD−1, 3GPD−R, 1HFC, 1MMQ, 1TGS−I, 1SGP−I, 1CHR−A, 1MUC−A, 1CLX−A, 2EXO, 1XYZ−A, 2VAA−A, 1MHC−A, 1SXA−A, 1XSO−A, 1SRD−A, 1COV−1, 1R1A−1, 1R09−1, 2RHN−1, 1COV−2, 1R09−2, 2MEV−2, 1POV−0, 1R1A−2, 1PVC−2, 1TME−2, 1THE−A, 1CSB−B, 2CRO, 1R69, 1CSE−E, 2PKC, 1ST3, 1THM, 2DRI, 1GCA, 4DFR−A, 1DYR, 8DFR, 3DFR, 2DGC−A, 1FOS−E, 4XIA−A, 2GYI−A, 1DOI, 1FXA−A, 4FXC, 1FXI−A, 1FRD, 1OVT, 1LCF, 1VFA−B, 1FGV−H, 1NMB−H, 1DVF−D, 1JHL−H, 1EBD−A, 1LVL, 2TPR−A, 3LAD−A, 1NDA−A, 1GES−A, 1EBH−A, 1PDZ, 1IGP, 2PRD, 1NPC, 8TLN−E, 1F3G, 1GPR, 1FCA, 1FXD, 1FXR−A, 2FCR, 1OFV, 1RCD, 1HRS, 1LAT−A, 1HCQ−A, 1GTA, 2GST−A, 3SDH−A, 1SCT−B, 1POH, 2HPR, 1PCH, 1HGX−A, 1HMP−A, 2RIG, 1RFB−A, 1HLB, 1HLM, 1HLP−A, 1LLD−A, 1HYH−A, 1LDN−A, 5LDH, 1LLC, 6LDH, 2HNT−C, 2PKA−A, 1IDO, 1LFA−A, 3IL8, 1NAP−A, 1PLF−A, 2LTN−A, 1SBA, 1LEC, 1PEL−A, 1LBI−A, 1TLF−A, 1MIN−B, 1MIO−B, 1NHK−R, 1NPK, 1PCR−L, 1PRC−L, 1PRC−M, 1PLC, 7PCY, 1QPG, 1PHP, 4SBV−A, 1SMV−A, 1TGL, 1TIA, 1TIB.