NdPASA: A novel pairwise protein sequence alignment algorithm that incorporates neighbor-dependent amino acid propensities

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · FEBRUARY 2004

Impact Factor: 2.63 · DOI: 10.1002/prot.20359 · Source: PubMed

2 AUTHORS, INCLUDING:

Junwen Wang
The University of Hong Kong
73 PUBLICATIONS **1,829** CITATIONS

SEE PROFILE

PROTEINS: Structure, Function, and Bioinformatics 58:628–637 (2005)

NdPASA: A Novel Pairwise Protein Sequence Alignment Algorithm That Incorporates Neighbor-Dependent Amino Acid Propensities

Junwen Wang[1,2†] **and Jin-An Feng**[1,2*]
[1]*Department of Chemistry, Temple University, Philadelphia, Pennsylvania*
[2]*Center for Biotechnology, Temple University, Philadelphia, Pennsylvania*

ABSTRACT **Sequence alignment has become one of the essential bioinformatics tools in biomedical research. Existing sequence alignment methods can produce reliable alignments for homologous proteins sharing a high percentage of sequence identity. The performance of these methods deteriorates sharply for the sequence pairs sharing less than 25% sequence identity. We report here a new method, NdPASA, for pairwise sequence alignment. This method employs neighbor-dependent propensities of amino acids as a unique parameter for alignment. The values of neighbor-dependent propensity measure the preference of an amino acid pair adopting a particular secondary structure conformation. NdPASA optimizes alignment by evaluating the likelihood of a residue pair in the query sequence matching against a corresponding residue pair adopting a particular secondary structure in the template sequence. Using superpositions of homologous proteins derived from the PSI-BLAST analysis and the Structural Classification of Proteins (SCOP) classification of a nonredundant Protein Data Bank (PDB) database as a gold standard, we show that NdPASA has improved pairwise alignment. Statistical analyses of the performance of NdPASA indicate that the introduction of sequence patterns of secondary structure derived from neighbor-dependent sequence analysis clearly improves alignment performance for sequence pairs sharing less than 20% sequence identity. For sequence pairs sharing 13–21% sequence identity, NdPASA improves the accuracy of alignment over the conventional global alignment (GA) algorithm using the BLOSUM62 by an average of 8.6%. NdPASA is most effective for aligning query sequences with template sequences whose structure is known. NdPASA can be accessed online at http://astro.temple.edu/~feng/Servers/BioinformaticServers.htm. Proteins 2005;58:628–637.** © 2004 Wiley-Liss, Inc.

Key words: sequence alignment; propensity; protein structures; sequence pattern; secondary structure

INTRODUCTION

Protein sequence alignment has become an essential part of biomedical research. It is one of the standard approaches to explore potential functional activity of a newly discovered protein by identifying sequence homologues that may be evolutionarily related.[1–3] Structural and functional information of a new protein can often be inferred from the knowledge of well-characterized homologous proteins.[4–7] In general, closely related protein sequences are relatively easy to align using the existing sequence-based methods.[8] However the success rate of these methods in finding correct alignment is significantly reduced when the sequence identity between two aligned sequences is lower than 30%, a threshold often referred to as the twilight zone.[9]

The performance of a sequence alignment algorithm largely depends on its employed substitution matrix. PAM and BLOSUM are two of the most commonly used substitution matrices in sequence alignment algorithms. These matrices were mainly derived from the frequencies of amino acid substitutions in a series of compiled families of protein sequences.[10,11] Algorithms employing BLOSUM62 and PAM250 are most effective in identifying and aligning homologous proteins.[11] Efforts to develop improved sequence alignment of remote homologues have focused on incorporating additional information that improved the diversity of the substitution matrices. These included the use of position-specific substitution profiles derived from multiple sequence alignment of protein families.[3,12] In recent studies, structure-based substitution matrices were developed. Such matrices were derived from the frequency of amino acids occupying similar positions in a series of structurally aligned proteins. Algorithms employing structure-based substitution matrices appeared to have improved success in detecting and aligning remotely related protein sequences.[13–19] Another approach that has achieved promising improvement in pairwise sequence alignment, particularly for sequences with low homology, is the sequence–template alignment method.[16,20] This algorithm

†Current address: Center for Bioinformatics and Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104.

Grant sponsor: National Institutes of Health; Grant number: GM54630. Grant sponsor: American Cancer Society; Grant number: PRG9926301GMC. Grant sponsor: Commonwealth of Pennsylvania (appropriation).

*Correspondence to: Jin-An Feng, Department of Chemistry, Temple University, 1901 N. 13th Street, Philadelphia, PA 19122. E-mail: feng@astro.temple.edu

Received 16 March 2004; Accepted 19 August 2004

Published online 22 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20359

incorporated the structural knowledge of the template, as well as the amino acid propensities for secondary structures, into a substitution matrix for sequence alignment.

While methods relying on structure-based sequence profiles are effective in identifying specific functional motifs in proteins, the overall improvement of these methods over the pure sequence-based methods in sequence alignments is limited. One potential source of such limitation may be dependent on the structure and sequence diversity of the derived profile. If profile construction places too much emphasis on sequence diversity, it could result in profiles with limited signature value. On the other hand, profiles derived from conserved proteins are not sensitive in detecting distantly related homologues. Efforts to reach an optimal balance in selecting protein sequences to construct profiles have always been a challenge. A recent study has suggested that the most effective profiles can be derived from sequences sharing 30–50% homology.[21] Another key weakness of the structure-based sequence profiles is that they lack information for loop regions, since structural alignments often ignore this region of the proteins. For sequences with long and functional loop regions, structure-based profile alignment methods could be ineffective. A method incorporating both sequence- and structure-based profiles, the hybrid sequence profiles, has shown improved performance in aligning proteins with distant homologues.[22]

In this report, we describe a new pairwise protein sequence alignment method, NdPASA, a neighbor-dependent propensity-assisted sequence alignment. By utilizing the secondary structural information on the template sequence, this method showed significant improvements over the conventional global alignment (GA) in aligning sequence pairs with less than 20% sequence identity. Our algorithm implemented a structure-dependent gap opening and extension penalty scheme, and neighbor-dependent amino acid secondary structure propensities. A higher gap penalty was applied for gaps occurring within the regular secondary structures than for gaps occurring in the loops. The neighbor-dependent amino acid secondary structure propensities were derived from our recent studies on neighbor-dependent sequence analysis of proteins. This analysis calculated the effect of neighboring amino acid types on the propensity of residues for adopting α-helices, β-strands and loops in proteins.[23,24] The neighbor-dependent sequence analysis produced an enhanced statistical significance scale that allowed us to explore the positional preference of amino acids in different secondary structures. A number of unique sequence patterns were identified. The values of neighbor-dependent propensity reflected the likelihood of an amino acid pair adopting a particular secondary structure conformation. The rationale for the utilization of neighbor-dependent amino acid propensity in sequence alignment is easily recognized. Methods employing sequence-based substitution matrix often have limited success in aligning sequences sharing a low percentage of sequence identity. The incorporation of the neighbor-dependent amino acid propensities in NdPASA allowed us to estimate the probability of an amino acid pair to be aligned with a corresponding amino acid pair adopting a specific secondary structure in the template sequence. For example, an amino acid pair in the query having a low neighbor-dependent propensity for α-helical conformation would be less likely aligned with an amino acid pair in an α-helix of the template. NdPASA performs most effectively when the structural information of the template sequence is available.

Using a data set of homologous sequence pairs with similar structural folds derived from the Structural Classification of Proteins (SCOP) database, the performance of the NdPASA was compared with the PSI-BLAST, the standard GA using BLOSUM62, and the IPASA (individual propensity assisted sequence alignment), a global sequence alignment algorithm that incorporated individual amino acid propensity for protein secondary structures. Our comparison statistics measured the ability of the methods to produce accurate alignments. These results revealed that the information derived from neighbor-dependent sequence analysis of proteins was effective in improving the performance of the global pairwise sequence alignment algorithm.

METHODS

Data Set of Sequence Pairs

The data set of sequence pairs was constructed as follows (Fig. 1). Protein sequences sharing less than 90% sequence identity were first extracted from the nonredundant Cull-Protein Data Bank (PDB) database.[25] The resolution cutoff was at 2.5 Å. The homologous sequence pairs selected for this study must satisfy two criteria: (1) The pair must share a certain level of sequence homology; and (2) the protein pair must adopt the same structural fold. To find proteins homologous to the selected sequences in the Cull-PDB database, we performed PSI-BLAST searches running against a nonredundant nucleotide database, NCBI nr (October 2002 release of the nonredundant sequence database). Sequences returned with *e*-values within the range of 10^{-6}–1000 were retained. To ensure that the sequence pairs belonged to the same structural fold, we compared sequences with the SCOP classification database (November 2002 release, version 1.61).[26] Sequence pairs belonging to either the same protein family, protein super family, and fold in the SCOP database were selected for this study. The data set of sequence pairs contained a total of 5553 proteins pairs with sizes ranging from 15 to 1520 residues. Of these sequence pairs, 4061 sequence pairs had sequence identity of 13–25%. From this subset of protein sequences, we randomly selected 500 pairs of protein sequences as a training set, and 3561 pairs of protein sequences as a testing set. The training set was used to optimize the parameters in our alignment algorithm.

The Algorithm of Neighbor-Dependent Propensity-Assisted Sequence Alignment (NdPASA)

The NdPASA incorporated the information of secondary structure propensity into the Needleman–Wunsch global alignment algorithm with affined gap penalty.[27] The

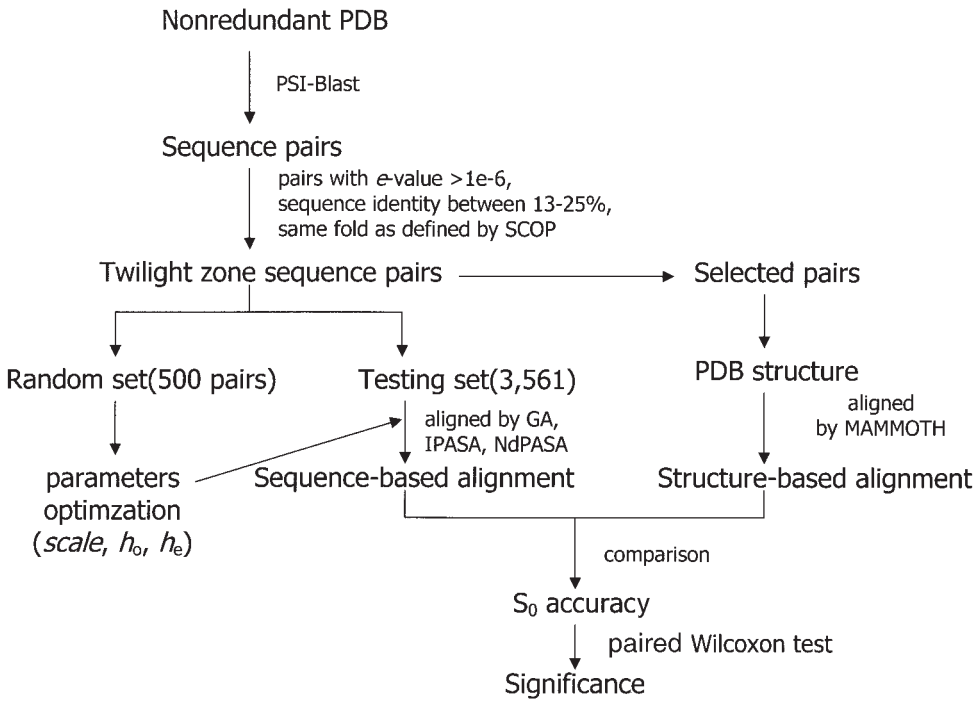

Fig. 1. A schematic workflow of the NdPASA algorithm.

Needleman–Wunsch dynamic programming algorithm, a conventional global sequence alignment algorithm (GA), was effective in finding the optimal scoring alignment or a set of alignments. Incorporating structure propensity values, we formulated $M(i,j)$ as follows:

$$M(i,j) = \max\left\{\begin{array}{c} M(i-1,j-1) + s(x_i,y_j) + scale^{*}p(x_{i-1},x_i,x_{i+1}|ss_j) \\ I_x(i-1,j-1) + s(x_i,y_j) + scale^{*}p(x_{i-1},x_i,x_{i+1}|ss_j) \\ I_y(i-1,j-1) + s(x_i,y_j) + scale^{*}p(x_{i-1},x_i,x_{i+1}|ss_j) \end{array}\right\} \quad (1)$$

$$I_x(i,j) = \max\left\{\begin{array}{c} M(i-1,j) - d(ss_j) \\ I_x(i-1,j) - e(ss_j) \end{array}\right\} \quad (2)$$

$$I_y(i,j) = \max\left\{\begin{array}{c} M(i-1,j) - d(ss_j) \\ I_y(l-1,j) - e(ss_j) \end{array}\right\}, \quad (3)$$

where $M(i,j)$ was the best score up to (i,j) given that x_i was aligned to y_j; and $I_x(i,j)$ was the best score given that x_i was aligned to a gap; $I_y(i,j)$ was the best score given that y_j was an insertion with respect to x; and $s(x_i, y_j)$ was the amino acid substitution score of x_i and y_i given a substitution matrix (BLOSUM62 for this study; see Results and Discussion section).[11,28] *Scale* was the scaling factor, denoting a relative weight between propensity score and substitution score. $p(x_{i-1}, x_i, x_{i+1} \mid ss_j)$ was the neighbor-dependent secondary structure propensity score for adopting the secondary structure element (SSE) of the template sequence at position j. It depended on the amino acid at the ith position and the amino acid types of its neighboring positions both proceeding $(i - 1)$ and following $(i + 1)$ the position i:

$$p(x_{i-1},x_i,x_{i+1}|ss_j) = \frac{p(x_{i-1},x_i|ss_j) + p(x_i,x_{i+1}|ss_j)}{2}, \quad (4)$$

where $p(x_{i-1}, x_i \mid ss_j)$ and $p(x_{i+1}, x_i \mid ss_j)$ were the logarithm of the neighbor-dependent propensities of the residues pairs (x_{i-1}, x_i) and (x_i, x_{i+1}), respectively, for adopting the SSE in the template at position $j(ss_j)$. The overall propensity score was chosen as the average of them. If a gap was applied at positions either proceeding or following the residue x_i, its available propensity was chosen as the overall propensity. The neighbor-dependent propensities of residues adopting α-helix (H), β-strand (S), and loop (L) conformations were calculated as previously described.[23,24] NdPASA employed the logarithmic values of the neighbor-dependent propensities:

$$p(x_i,x_{i+1}|ss_j) = \log\left\{\frac{\text{Frequency of } x_i \text{ followed by } x_{i+1} \text{ in data set } ss_j}{\text{Frequency of } x_i \text{ followed by } x_{i+1} \text{ in whole protein data set}}\right\}. \quad (5)$$

Overall, three 20 × 20 neighbor-dependent propensity tables were generated, each specific for its respective secondary structure (i.e., H, S, and L). The tables were applied in the alignments according to Eq. (1). For alignments with the incorporation of individual amino acid propensity (i.e., the IPASA algorithm), the individual propensity score was the amino acid propensity at the ith position for adopting the SSE in the template at the jth position, without considering its neighbors[20]:

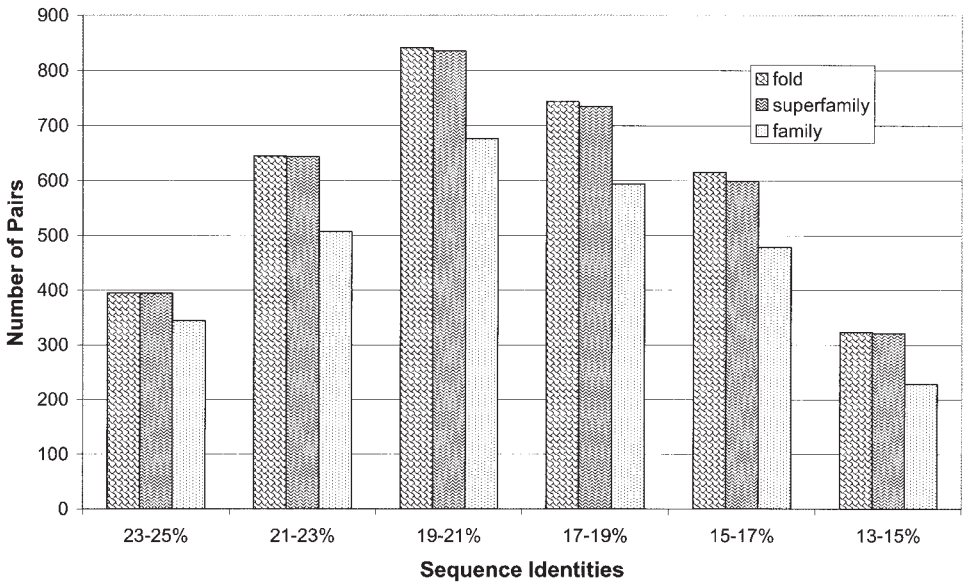

Fig. 2. A bar graph of the number of protein sequence pairs belonging to the same family, superfamily, and fold as defined by SCOP illustrating varied population distribution at different sequence identity ranges.

$$p(x_{i-1},x_i,x_{i+1}|ss_j) = p(x_i|ss_j) \quad (6)$$

where $p(x_i \mid ss_j)$ was calculated by

$$p(x_i|ss_j) = \log\left\{\frac{\text{Frequency of } x_i \text{ in } ss_j \text{ data set}}{\text{Frequency of } x_i \text{ in whole protein data set}}\right\}. \quad (7)$$

Again, three individual propensity tables were generated. The tables were applied in the alignments using Eq. (1), where $p(x_{i-1}, x_i, x_{i+1} \mid ss_j)$ was substituted by $p(x_i \mid ss_j)$. The gap opening and extension penalties were also depended on the template's secondary structure element:

$$d(ss_j) = \begin{pmatrix} h_o & ss_j = 'H'/'S' \\ l_o & ss_j = 'L' \end{pmatrix} \quad (8)$$

$$e(ss_j) = \begin{pmatrix} h_e & ss_j = 'H'/'S' \\ l_e & ss_j = 'L' \end{pmatrix}, \quad (9)$$

where parameter *scales*, h_o, h_e, l_o, and l_e, were estimated from optimizing the alignment accuracy of the training sequence pairs. Considering that regular secondary structures were often more conserved than loop regions, the gap opening penalties $d(ss_j)$ for the helices (H) and strands (S) were assigned higher values than those for the loops (L).

Evaluation of the Alignment Accuracy

Structural alignments of proteins were often considered as the gold standard to measure the accuracy of alignments by a sequence alignment method. Aligned residues pairs of the structural alignments were derived from structural superposition. We chose MAMMOTH for structure alignments of sequence pairs in both training and testing data sets.[29] The aligned residue pairs produced by the program were within 4.0 Å of each other in the structural alignments and were called alignable pairs. The accuracy of the alignments was determined by comparing the aligned residue pairs determined by MAMMOTH with those produced by sequence alignment algorithms. Residue pairs were considered correctly aligned if they agreed with the MAMMOTH alignment without any positional shifts. The accuracy of the alignment algorithms was defined as the ratio between the correctly aligned residue pairs and the total number of structurally aligned residue pairs:

$$s_0 = \frac{\text{Correctly aligned pairs}}{\text{Total structure alignable pairs}}. \quad (13)$$

RESULTS AND DISCUSSION

The database of the testing sequence pairs contained a total of 3561 sequence pairs, with sequence identity ranging from 13% to 25%. Figure 2 shows the population distribution of the database. All sequence pairs belonged to the same fold based on the SCOP classification. A closer inspection of the sequence pairs found that almost 99% of the sequence pairs shared the same protein superfamily, while nearly 80% of these pairs shared the same protein family. The number of interfamily sequence pairs was significantly higher among sequence pairs sharing low percentages of sequence identity.[30] At sequence identity of 23–25%, the interfamily pairs constituted 12.7% of the total pairs in a superfamily, while at a sequence identity of 13–15%, the interfamily pairs were 29%. The testing set contained 326 sequence pairs with sequence identity at 13–15%. At sequence identity range of 15–23%, there were more than 600 pairs. The maximum number of pairs had sequence identity in the range of 19–21%, with 847 sequence pairs, of which 682 pairs belonged to the same family based on SCOP assignments.

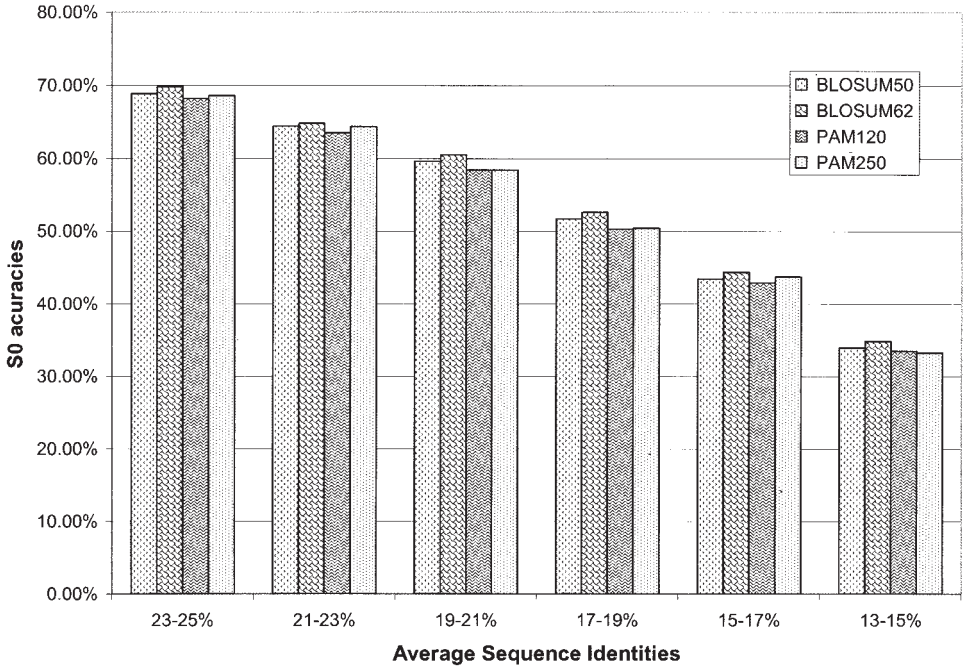

Fig. 3. Comparison of the average accuracies of NdPASA using 5 different substitution matrices: BLOSUM50, BLOSUM62, PAM120, and PAM250.

Selection of a proper substitution matrix could affect the performance of a sequence alignment algorithm. Some of the most widely used matrices have been BLOSUM50, BLOSUM62, PAM120, and PAM250. Of these substitution matrices, BLOSUM62 has been found to be most effective in local sequence alignment algorithms, such as PSI-BLAST and Smith–Waterman-based algorithms.[3,11,31,32] A more recent study by Elofsson focusing on optimizing parameters of various sequence alignment programs has also found BLOSUM62 to be the most effective substitution matrix for conventional global sequence (GA) algorithms,[30] although an earlier study found BLOSUM50 performed slightly better than BLOSUM62.[6] We evaluated the performance of NdPASA using the 4 substitution matrices. Figure 3 shows that, for sequence pairs of the various sequence homology ranges, the NdPASA performed consistently the best by using BLOSUM62. We also tested the performance of GA using the 5 substitution matrices. Our results also showed that BLOSUM62 was the best performing substitution matrix for the GA algorithm (data not shown). It should be noted that the pool of sequence pairs of this study is of low homologous proteins, with sequence identity in the range of 13–25%, while the earlier studies included mostly sequence pairs with higher sequence homology.[6]

We used the bootstrap approach to optimize parameters of the NdPASA that included *scale*, h_o, h_e, l_o, and l_e. The 500 randomly selected sequence pairs in the training set were used for the optimization procedure. The performance of NdPASA was evaluated by comparing alignment results with the structure alignments of sequence pairs in the training set using MAMMOTH (Fig. 1). Over 200 combinations of different gap and insertion penalties, as well as the *scale* values, were systematically tested for NdPASA algorithm. The best performance was obtained by using the following parameters: *scale* = 2.2; h_o = 19.0; h_e = 2.0; l_o = 11.0; and l_e = 1.0. These parameters were applied to perform alignments on the sequence pairs in the testing set.

The gap opening and extension, as well as the end-gap, penalties are essential parameters in optimizing the performance of NdPASA. We observed that the effect of the end-gap penalty was more significant on NdPASA performance when aligning sequence pairs with low homology. For sequence pairs with 17–25% identity, the S_0 of the NdPASA with the end-gap penalties was 1–12% higher than that of NdPASA without the end-gap penalty. For sequence pairs sharing less than 17% sequence identity, the NdPASA with the end-gap penalties performed significantly better with S_0 values 20–46% higher than that of the NdPASA without the end-gap penalty. For conventional GA algorithm, the end-gap penalty had a similar effect on its performance, although to a much lesser extent. For sequence pairs with 17–25% sequence identity, there was no significant difference in performance between GA with and without the end-gap penalty. On the other hand, for remotely related sequence pairs with sequence identity below 17%, the GA with the end-gap penalties scored 1–10% higher than that of the GA without the end-gap penalty. An earlier study found that the GA algorithm performed better without the end-gap penalty when aligning sequence pairs that shared sequence identity higher than 20%. It appeared that the end-gap penalty was more effective for aligning sequences with remote homology. A similar pattern was also found for the IPASA algorithm. There were no significant differences between the end-gap

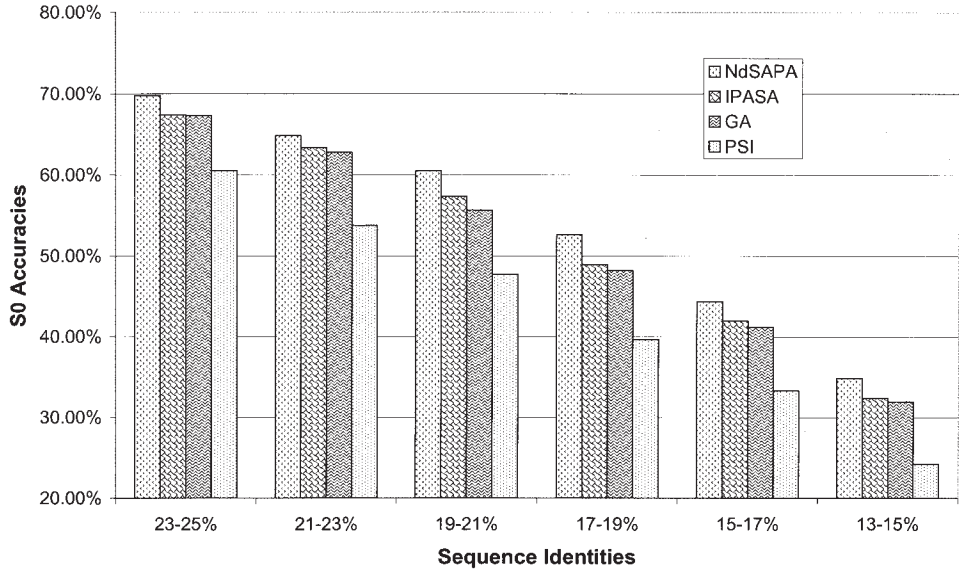

Fig. 4. Comparison of the average accuracies of 4 alignment methods at different sequence identity ranges.

alignment and no-end-gap alignments by IPASA at sequence identity greater than 15%. However, at sequence identity between 13% and 15%, the IPASA algorithm with end-gap penalty performed better than the algorithm with no-end-gap penalty by about 10% (data not shown). Due to the employment of end-gap penalty in NdPASA, the algorithm was less effective in aligning sequence pairs with significant difference in size. Specifically, NdPASA had less significant improvement over the conventional GA algorithm in aligning sequences between a single-domain protein and a protein with multiple domains than that of sequence pairs with comparable size.

In order to evaluate the performance of NdPASA in alignment accuracy, we carried out sequence alignments of the sequence pairs in the test set using 4 different algorithms, including PSI-BLAST, pairwise GA, NdPASA, and IPASA. The latter 3 were all implemented with end-gap penalty. The results of these alignments are illustrated in Figure 4. The range of sequence identity was 13–25% for the tested sequence pairs. The accuracy of the alignments was measured by their S_0 values. By comparison, NdPASA performed well in all subsets of the percentage of sequence identity, with an average of 2–43% improvement over other algorithms. The improvements of NdPASA algorithm over other algorithms were in the range of 3–9% for GA algorithm, 2–8% for IPASA algorithm, and 15–44% for the PSI-BLAST. The significantly large improvement over PSI-BLAST could be due in part to the fact that PSI-BLAST only aligned sequence pairs locally.[3,31] For sequence pairs sharing more than 30% identity, the NdPASA algorithm showed no significant improvement over the conventional GA algorithms (data not shown).

The performance of the NdPASA was significantly better in aligning sequence pairs with low sequence homology (Fig. 4). For sequence pairs with 21–25% identity, NdPASA had an average accuracy of 2.8% and 3.4% better than that of IPASA and GA algorithms, respectively, while for sequence identity in the 13–19% range, its average accuracy was 6.9% and 8.6% better than that of IPASA and GA algorithms, respectively. These results suggested that the sequence patterns derived from the neighbor-dependent sequence analysis of protein structures had a more significant contribution to sequence alignments for sequence pairs that were remotely related. For protein pairs with sequence identity above 21%, 38% (397) of the 1039 alignments produced by NdPASA were better than those produced by IPASA, as shown in Table I. In contrast, the percentage of alignments generated by the NdPASA that were better than IPASA increased to 49% for protein pairs sharing less than 21% sequence identity. For protein pairs with sequence identity above 21%, 49% (510) of the 1039 alignments produced by NdPASA were better than those produced by GA. For sequence pairs sharing less than 21% identity, the number of alignments in which NdPASA performed better increased to 56% (1412) of the 2522 alignments, while the number of alignments in which GA was better decreased to 36%.

To evaluate the statistical significance of the improvements that NdPASA had over other algorithms, we performed an one-tailed nonparametric Wilcoxon signed rank test.[33] Alignments with $P < 0.05$ from the test were considered significant improvement. For sequence pairs of different identity ranges, the alignment accuracy scores (S_0) against the MAMMOTH structural alignment standard were computed for GA, IPASA, PSI-BLAST, and NdPASA algorithms. The S_0 scores from each of the two compared algorithms were used as two input sets to test if they were significantly different. Table I compares alignment scores S_0 between NdPASA and GA algorithms at various identity ranges. Overall, the scores of the NdPASA alignments were higher for 54% of sequence pairs in the

TABLE I. Statistical Comparison of Alignment Accuracy from NdPASA, IPASA, GA, and PSI-BLAST

Sequence identity range	NdPASA vs IPASA				NdPASA vs GA				NdPASA vs PSI-BLAST			
	No. of pairs				No. of pairs				No. of pairs			
	Better	Same	Worse	*p*-value	Better	Same	Worse	*p*-value	Better	Same	Worse	*p*-value
23–25%	155	126	114	1.6e-5	197	33	165	6.0e-5	288	35	72	<2e-16
21–23%	242	212	190	7.4e-6	313	62	269	1.6e-4	510	32	102	<2e-16
19–21%	397	228	216	<2e-16	483	54	304	1.4e-15	661	34	146	<2e-16
17–19%	371	184	188	<2e-16	421	61	261	1.6e-15	580	41	122	<2e-16
15–17%	308	145	162	2.8e-11	320	59	236	2.7e-7	475	47	93	<2e-16
13–15%	165	59	99	8.0e-6	188	20	115	2.2e-5	242	28	53	<2e-16
All ranges	1638	954	969	<2e-16	1922	289	1350	<2e-16	2756	217	588	<2e-16

testing set than those of the GA alignments. Eight percent of sequence pairs had the same alignment scores by both algorithms. The GA algorithm scored higher aligning 38% of the sequence pairs in the testing set. The proportion of NdPASA alignments that had higher S_0 values over GA alignments varied for sequence pair groups with a different degree of homology. For sequence pairs sharing 21–23% identity, NdPASA performed better than GA in aligning 49% of sequence pairs, while for sequence pairs sharing 13–15% identity, NdPASA had better performance for 58% of the sequence pairs. Wilcoxon estimation showed the *p*-values for all sequence identity ranges were less than 0.001, indicating the better performance of NdPASA over GA was statistically significant. Similar comparisons were made for NdPASA versus IPASA, as well as for NdPASA versus PSI-BLAST. The results are listed in Table I.

Comparative protein structure modeling is one of the most widely used structure prediction techniques. Its success relies heavily on the accurate sequence alignment of the query sequence with a template sequence whose three-dimensional (3D) structure is known.[34] The structure assignment of a residue in the query sequence is dependent on its alignment with the template sequence. When modeling a query sequence that shares less than 25% sequence identity with that of the template, structural assignments of greater regions of the protein sequence are uncertain due to poor quality of sequence alignment. NdPASA, with improved performance in aligning sequence pairs of low homology, could be a valuable tool for comparative protein structure modeling. Such application is illustrated in the following example on structural modeling of cytochrome c′.

Cytochrome c′ is one of the key components in promoting bacterial photosynthesis, denitrification, nitrogen fixation, and sulfur oxidation.[35] Cytochrome c′ exists in both dimeric and monomeric forms depending on its origin of species. While the exact biological functional role of the cytochrome c′ is still unclear, it has been shown to bind NO, CO, cyanide, and alkylisocyanide molecules. Structural studies of the proteins revealed key residues that are essential to both its heme-coordinated architecture and perhaps its ligand-binding activity.[35] The *Rhodopseudomonas palustris* cytochrome c′ (RPCP; PDB code: 1A7V) and the *Rhodospirillum molischianum* cytochrome c′ (RMCP; PDB code: 2CCY) shared 19% sequence identity. Both proteins adopt similar 4-helical bundle structural fold with a coordinating heme group. The superimposition of the two structures by MAMMOTH yielded an overall standard deviation of 2.4 Å.

The *R. palustris* cytochrome c′ has 4 antiparallel α-helices packed against each other to adopt a familiar fold of helical bundle.[35] A hydrophobic core encircled by the helices stabilizes the structural fold. The essential heme group is partially embedded in the hydrophobic core. Unlike cytochrome c, the heme-iron of the cytochrome c′ has an unoccupied coordination site at the axial position, which is protected by a hydrophobic pocket. Residues that line the hydrophobic pocket are Leu12, Met15, Leu85, Phe55, Phe78, and Phe82. They function as a gate modulating the access of potential ligands to the heme-binding site.

Figure 5 shows the sequence alignments of two cytochrome c′ proteins using PSI-BLAST, GA, IPASA, and NdPASA. The residue coverage for this alignment by PSI-BLAST was 98%. Using results of structural alignment by MAMMOTH as a standard, the NdPASA algorithm matched 83% of all the residues correctly, while the PSI-BLAST and the IPASA correctly matched 48% and 37% of all the residues, respectively. All 3 algorithms properly aligned the heme-coordinating histidine residues of RPCP (His117) and RMCP (His121). On the other hand, only NdPASA correctly aligned all 6 functionally important hydrophobic residues of both proteins. PSI-BLAST aligned Phe55 of RPCP with Glu68 of RMCP, Phe78 of RPCP with Glu91, and Leu81 of RPCP with Lys94 of RMCP, which would introduce unfavorable charges into the hydrophobic pocket. Both GA and IPASA mismatched 3 out of 6 of these functional hydrophobic residues, including Phe55 of RPCP aligned with Pro65 of RMCP by IPASA (Glu68 of RMCP by GA); Phe78 of RPCP aligned with Leu88 of RMCP by IPASA (Glu91 of RMCP by GA); and Leu81 of RPCP aligned with Glu92 of RMCP by IPASA (a gap of RMCP by GA). The errors of these alignments were caused by an introduction of frame shifts by both PSI-BLAST and IPASA. These shifts would likely hinder the effectiveness of interpreting the structure–function of the RMCP based on sequence alignment.

The introduction of secondary structure information in protein sequence alignment has been reported in a number of methods. Some of these methods employed structure-derived profiles to detect remotely related proteins by using dynamic programming to align the

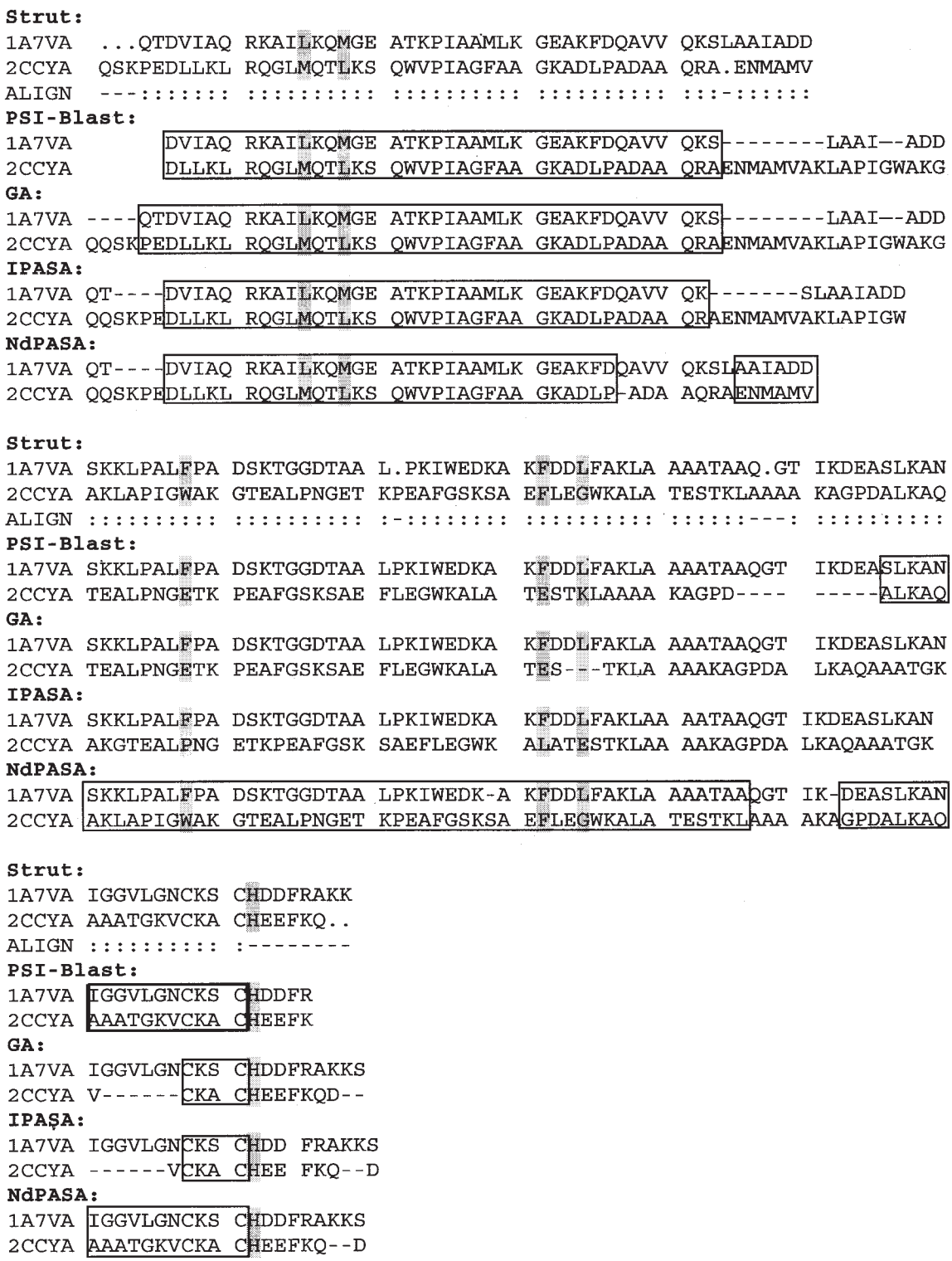

Fig. 5. Alignments of cytochrome c′ from *R. palustris* (PDB code: 1A7V) and *R. molischianum* (PDB code: 2CCY) by structural alignment (MAMMOTH), GA, PSI-BLAST, IPASA, and NdPASA. Residues aligned by the MAMMOTH are indicated by colons. Correctly aligned regions are marked by brackets. Six functional residues of the cytochrome c′ are shaded with gray. Sequence identity: Structure alignment, 19%; PSI-BLAST, 22%; GA, 23%; IPASA, 23%; NdPASA, 21%. Alignment score (S_0): PSI-BLAST, 48%; GA, 39%; IPASA, 37%; NdPASA, 83%.

sequence profiles of both query and template.[19,36–38] Others incorporated the secondary structure propensity of amino acids to improve pairwise sequence alignment.[20,22] While the utilization of structure information varied in different techniques, the performance of these methods showed significant improvement over methods

that relied on sequence information alone. Statistical analyses of the performance of NdPASA indicated that the introduction of sequence patterns of secondary structure derived from neighbor-dependent sequence analysis clearly improved alignment performance for sequence pairs sharing less than 20% sequence identity. In comparison with the performance of IPASA, the residue pair propensity of the NdPASA was more effective than individual residue propensity of the IPASA in assisting the alignment of remotely related sequences. Unlike profile-based alignment methods, NdPASA optimized alignment by evaluating the likelihood of a residue pair in the query sequence matching against a corresponding residue pair adopting a particular secondary structure in the template sequence. Since the residue pair propensity was derived from a data set not restricted to a particular protein fold, these sequence patterns represented their intrinsic structural preference in proteins. NdPASA essentially utilized the sequence patterns as an additional independent parameter for alignment optimization. For sequence pairs with low sequence homology, where the substitution matrix was often less effective for alignment, NdPASA optimized alignment by matching sequence patterns with a preference for the corresponding secondary structure of the template sequence.

The results presented in this work suggest that NdPASA offers significant improvement over traditional GA algorithm on pairwise sequence alignment, particularly for sequence pairs with low sequence homology. Due to its reliance on the structure knowledge of the template sequence, NdPASA is most effective in aligning query sequence with templates whose 3D structure is known. In the case where no structural information is available, one could acquire the secondary structure information of the template from secondary structure prediction methods before applying NdPASA for sequence alignment. Improved performance in aligning proteins of remote homology could have significant impact on functional knowledge of the query protein. Accurate pairwise alignment of enzymes could help identify residues in the active sites. NdPASA could also be used to complement PSI-BLAST database searches for homologues. Developed as a local alignment method, PSI-BLAST is effective in identifying segments of strong sequence conservation that may define certain motifs in proteins. However, proteins sharing segments of motif sequences may not always be related homologues. This could be one of main factors that PSI-BLAST often generates multiple positive results with comparable scores in database searches. Using NdPASA to perform subsequent alignments of sequence pairs generated by PSI-BLAST could help select the best positives in the search results.

CONCLUSIONS

NdPASA is an effective method to align protein sequences where the structural knowledge of the template sequence is known. It optimizes alignment by evaluating the likelihood of a residue pair in the query sequence matching against a corresponding residue pair adopting a particular secondary structure in the template sequence. Statistical analysis of our results has shown that the method achieved higher accuracy in aligning proteins with less than 20% sequence identity than that of PSI-BLAST, conventional GA algorithm using BLOSUM62, and IPASA. Consistent with previous findings where use of structural information in alignment algorithms increased accuracy,[19,20,22] the incorporation of the sequence patterns of secondary structures derived from the neighbor-dependent sequence analysis of proteins made significant contributions to improving the performance of NdPASA. NdPASA is maintained as a web server, and it is available for public access online at http://astro.temple.edu/~feng/Servers/BioinformaticServers.htm.

ACKNOWLEDGMENTS

Our thanks to members of the Feng laboratory for helpful discussions. We would also like to acknowledge one of the referees, whose helpful suggestions have improved the overall presentation of this article.

REFERENCES

1. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–2448.
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
3. 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. Nucleic Acids Res 1997;25:3389–3402.
4. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
5. Scharf M, Schneider R, Casari G, Bork P, Valencia A, Ouzounis C, Sander C. GeneQuiz: a workbench for sequence analysis. Proc Int Cont Intell System Mol Biol 1994;2:348–353.
6. Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol 1997;273:355–368.
7. Teichmann SA, Chothia C, Gerstein M. Advances in structural genomics. Curr Opin Struct Biol 1999;9:390–399.
8. Feng DF, Johnson MS, Doolittle RF. Aligning amino acid sequences: comparison of commonly used methods. J Mol Evol 1985;212:112–125.
9. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
10. Dayhoff M, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. In: Dayhoff M, Editor. Atlas of protein sequence and structure. National Silver Springs, MD: Biomedical Research Foundation; 1978. p 345–352.
11. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
12. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84:4355–4358.
13. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence–structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001;310:243–257.
14. Ogata K, Ohya M, Umeyama H. Amino acid similarity matrix for homology modeling derived from structural alignment and optimized by the Monte Carlo method. J Mol Graph Model 1998;16:178–189.
15. Johnson MS, Overington JP. A structural basis for sequence comparisons: an evaluation of scoring methodologies. J Mol Biol 1993;233:716–738.
16. Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 1997;269:423–439.
17. May AC, Johnson MS. Improved genetic algorithm-based protein

structure comparisons: pairwise and multiple superpositions. Protein Eng 1995;8:873–882.
18. Prlic A, Domingues FS, Sippl MJ. Structure-derived substitution matrices for alignment of distantly related sequences. Protein Eng 2000;13:545–550.
19. Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. J Mol Biol 2001;307:721–735.
20. Yang AS. Structure-dependent sequence alignment for remotely related proteins. Bioinformatics 2002;18:1658–1665.
21. Panchenko AR, Bryant SH. A comparison of position-specific score matrices based on sequence and structure alignments. Protein Sci 2002;11:361–370.
22. Tang CL, Xie L, Koh IYY, Posy S, Alexov E, Honig B. On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. J Mol Biol 2003;334:1043–1062.
23. Crasto CJ, Feng JA. Sequence codes for extended conformation: a neighbor-dependent sequence analysis of loops in proteins. Proteins 2001;42:399–413.
24. Wang J, Feng JA. Exploring the sequence patterns in the alpha-helices of proteins. Protein Eng 2003;16:799–807.
25. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
26. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
27. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
28. Durbin R, Eddy SR, Krogh A, Mitchison GJ. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge, UK: Cambridge University Press; 1998.
29. Ortiz AR, Strauss CE, Olmea O. MAMMOTH: matching molecular models obtained from theory: an automated method for model comparison. Protein Sci 2002;11:2606–2621.
30. Elofsson A. A study on protein sequence alignment quality. Proteins 2002;46:330–339.
31. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
32. Shpaer EG, Robinson M, Yee D, Candlin JD, Mines R, Hunkapiller T. Sensitivity and selectivity in protein similarity searches: a comparison of Smith–Waterman in hardware to BLAST and FASTA. Genomics 1996;38:179–191.
33. Zar JH. Biostatistical analysis. 3rd edition. Upper Saddle River, NJ: Prentice-Hall; 1996.
34. Sali A, Kuriyan J. Challenges at the frontiers of structural biology. Trends Cell Biol 1999;9:M20–M24.
35. Shibata N, Iba S, Misaki S, Meyer TE, Bartsch RG, Cusanovich MA, Morimoto Y, Higuchi Y, Yasuoka N. Basis for monomer stabilization in *Rhodopseudomonas palustris* cytochrome c′ derived from the crystal structure. J Mol Biol 1998;284:751–760.
36. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
37. Jaroszewski L, Rychlewski L, Zhang B, Godzik A. Fold prediction by a hierarchy of sequence threading and modeling methods. Protein Sci 1998;7:1431–1440.
38. Ginalski K, Pas J, Wyrwicz LS, von Grotthuss M, Bujnicki JM, Rychlewski L. ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res 2003;31:3804–3807.