Optimization of a New Score Function for the Generation of Accurate Alignments

Bin Qian[1] **and Richard A. Goldstein**[1,2*]
[1]*Biophysics Research Division, University of Michigan, Ann Arbor, Michigan*
[2]*Department of Chemistry, University of Michigan, Ann Arbor, Michigan*

ABSTRACT **The accuracy of the alignments of protein sequences depends on the score matrix and gap penalties used in performing the alignment. Most score functions are designed to find homologs in the various databases rather than to generate accurate alignments between known homologs. We describe the optimization of a score function for the purpose of generating accurate alignments, as evaluated by using a coordinate root-mean-square deviation (RMSD)-based merit function. We show that the resulting score matrix, which we call STROMA, generates more accurate alignments than other commonly used score matrices, and this difference is not due to differences in the gap penalties. In fact, in contrast to most of the other matrices, the alignment accuracies with STROMA are relatively insensitive to the choice of gap penalty parameters. Proteins 2002;48:605–610.** © **2002 Wiley-Liss, Inc.**

Key words: sequence alignment; dynamic programming; score matrix; molecular evolution

INTRODUCTION

The various genome projects are providing us with a flood of new protein sequences. Unfortunately, little is known about the structure, function, or metabolic role of many of these proteins. One of the most powerful computational approaches to gain this type of information is through sequence comparison. In this method, the target sequence is aligned and compared with the protein sequences in the various protein databases. Proteins with high-sequence similarity are likely to be homologous, that is, to share a common ancestry. If the structure of the homologous protein is known, a model can be made of the target protein through homology modeling by using the structure of the identified homolog as a template.[1] This application will become increasingly common as the structures of more template proteins become available. Even in the absence of homologs with known structure, we can still derive much information through an analysis of the patterns of amino acid substitutions at different locations[2–6] and correlations between changes at various positions.[7,8] The identification of homologous sequences can also increase our ability to detect further homologs, either through the construction of a statistical model such as profiles[9] or Hidden Markov Models,[10,11] or through iterative refinement of the alignment matrix as in PSI-BLAST.[12] All of these approaches require the detection of a set of possibly distant homologies combined with the generation of accurate alignments between the target sequence and the identified homologs.

Most commonly used homology search engines such as FASTA[13] and BLAST[14,15] use approaches based on dynamic programming algorithms.[16,17] Given a pair of sequences and a score function, the algorithm computes the alignment that maximizes the alignment score, furnishing to the user both an alignment and a final score used to evaluate the evidence for homology. Homology search methods are designed to be good at detecting homologous relationships between protein sequences. However, they often fail to achieve correct alignments between the homologous sequences, especially when the sequences are distantly related. This is not surprising, because there is a significant difference between identifying a homologous protein from a database of protein sequences and identifying the correct alignment of a known set of homologs from the space of all possible alignments. Therefore, there is no reason to believe that the same score matrix or gap penalties should be optimal for both finding and aligning homologs. For example, Vogt and coworkers[18] derived different optimal gap penalties for the generation of correct alignments from those used for the detection of homologs. Because inaccurate alignments compromise the use of homologs in comparative modeling as well as other applications, it is important to create methods specifically designed for the alignment of previously identified homologs.

In a previous article,[19] we developed a new score matrix, OPTIMA, which can be used to improve the accuracy of homology searches, especially in the so-called "twilight zone" (homologous pairs of sequences with sequence identity less than 25%[20]). In contrast to the standard formulation of a score function as a log-odds matrix,[21,22] OPTIMA was optimized for the identification of distant homologies by using an iterative optimization technique. In this

Grant sponsor: College of Literature, Science, and the Arts; Grant sponsor: Program in Protein Structure and Design; Grant sponsor: Horace H. Rackham School of Graduate Studies; Grant sponsor: National Institutes of Health; Grant number: LM0577; Grant sponsor: National Science Foundation; Grant number: BIR9512955.

*Correspondence to: Richard A. Goldstein, Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055. E-mail: richardg@umich.edu

Received 2 October 2001; Accepted 12 February 2002

article, we use an analogous method to develop a new score matrix (STRucturally Optimized Matrix for sequence Alignment, STROMA) optimized to generate accurate alignments for previously identified distantly homologous sequences. This is performed by iterative optimization of the score function based on a coordinate-root-mean-square deviation (RMSD)-based merit function. Results with various test sets show our score function is superior to other commonly used score functions in generating accurate alignments between these distant but known homologs. This difference is not attributable solely to the choice of gap penalties.

MATERIALS AND METHODS

Theory

The various alignment methods generally find the alignment that maximizes an alignment score, computed by using an expression such as

$$S = \sum_{i,j} n_{i,j}\gamma_{i,j} + n_{\mathrm{gap-I}}\,\gamma_{\mathrm{gap-I}} + n_{\mathrm{gap-E}}\,\gamma_{\mathrm{gap-E}} \qquad (1)$$

where $n_{i,j}$ refers to the number of times that amino acid type i is aligned with amino acid type j, $n_{\mathrm{gap-I}}$ is the total number of gaps in the alignment, $n_{\mathrm{gap-E}}$ is the total number of residues in each gap beyond one, and $\gamma_{i,j}$, $\gamma_{\mathrm{gap-I}}$, and $\gamma_{\mathrm{gap-E}}$ represent the contribution to the score for any amino acid match or mismatch, initialization of a gap, and extension of a gap, respectively. $\gamma_{i,j}$ is known as the score function, substitution matrix, or exchange residue matrix, whereas $\gamma_{\mathrm{gap-I}}$, and $\gamma_{\mathrm{gap-E}}$ represent the gap penalties. The resulting alignment depends critically on the values of these parameters. We are interested in finding the best values for $\gamma_{i,j}$, $\gamma_{\mathrm{gap-I}}$, and $\gamma_{\mathrm{gap-E}}$ that produce the most accurate alignments of known homologs.

It has been observed that the three-dimensional structure of a protein is more conserved than its amino acid sequence during evolution.[23] Thus, when the sequence similarity is hard to detect, we can resort to structure similarity to guide us toward the correct alignment. We use the coordinate RMSD after optimal rigid body superposition[24] to quantitatively measure the structural similarity between structures A and B:

$$\mathfrak{D}(A,B) = \sqrt{\frac{1}{N_{\mathrm{a}}}\sum_{i=1}^{N_{\mathrm{a}}} (d_{ABi})^2} \qquad (2)$$

where N_{a} is the total number of aligned residues in the protein pair and d_{ABi} is the corresponding distance between the C_α atoms of the ith aligned residues from the two proteins A and B when oriented so as to minimize $\mathfrak{D}$. For any particular score function we can align all of the pairs of proteins in any particular data set and calculate $\langle\mathfrak{D}\rangle$, the RMSD averaged over all of the pairs of proteins. This is our merit function for optimizing and testing of the score function.

Database Preparation

We generated a training set of 700 protein pairs from the Distant Aligned Protein Structures (DAPS) database (December 99 release),[25] which is derived from the FSSP[26] and SCOP[27] databases. The DAPS database contains only sequence pairs with sequence identity $< 25\%$. An independent test set of 276 pairs was generated in the same way, with all pairs in the test set different from those in the training set. We generated an additional test set from the Combinatorial Extension (CE) database[28] by choosing one sequence pair with $< 25\%$ sequence identity from each of the structural families. All CE proteins with similarity (BLAST E-score < 0.1) to any of the proteins in the DAPS training and test sets were excluded, leaving 196 protein pairs in the final CE test set. All sequences shorter than 100 or longer than 600 residues were also excluded from all of the datasets.

Every residue in each protein was considered as a point, with the coordinate taken from its C_α atom. If there is more than one set of coordinate data for a certain protein sequence in its PDB file, we used the residue coordinate from the first set of data.

Optimization of the Score Function

As described above, we are interested in minimizing the average coordinate RMSD $\langle\mathfrak{D}\rangle$ for the proteins in our training set. We perform this optimization by using a scheme similar to that previously used in the derivation of OPTIMA.[19] Starting with the GCB matrix,[29] we used global dynamic programming algorithm (end gaps not penalized) to align each of the structurally similar protein pairs in the DAPS training set and calculated $\langle\mathfrak{D}\rangle$. We then used a downhill simplex algorithm[30] to find the score function that minimizes $\langle\mathfrak{D}\rangle$, monitoring and evaluating the resulting performance with the DAPS and CE test sets. Downhill simplex method requires only function evaluations and can always find at least a local minimum. Multiple optimization iterations are used before we reached our final score function. Because multiplication of the score function by any constant does not change the resulting alignment, we fixed one entry (Cys-Cys score) of the score matrix, resulting in 211 adjustable parameters corresponding to the remaining 209 score matrix entries and the two gap penalties.

RESULTS

The values of $\langle\mathfrak{D}\rangle$ as averaged over the training set and DAPS test set during the optimization process are shown in Figure 1. The optimization is terminated when the $\langle\mathfrak{D}\rangle$ values converged. The convergence of the merit function for the test set in Figure 1 shows there is no overfitting before the optimization was terminated. The resultant score function is shown in Table I.

Table II lists the average values of $\langle\mathfrak{D}\rangle$ for the proteins in the two test sets. A comparison is made between the STROMA score function as well as other score functions in general use, including the revised PAM250 and PAM350 matrices of Dayhoff et al.,[21] the popular BLOSUM62 and BLOSUM30 matrices,[31] the matrix derived by Gonnet, Cohen, and Benner (GCB),[29] the matrix of Jones, Taylor, and Thornton (JTT),[32] the structure-based matrix of Overington and colleagues (STR),[33] the matrix recently intro-

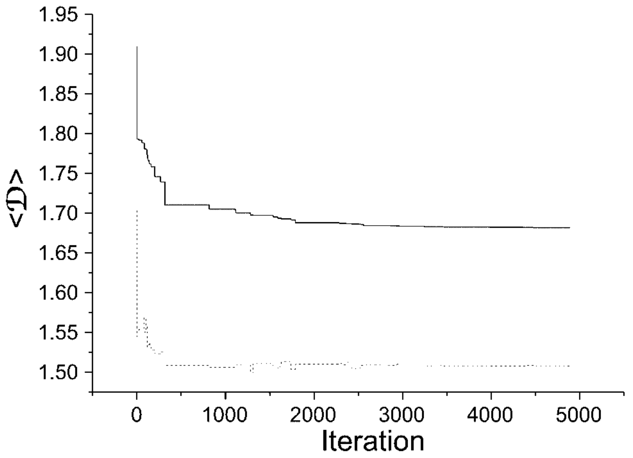

Fig. 1. Optimization of ⟨𝔇⟩ over the DAPS training set (—) and test set (- - - -).

duced by Blake and Cohen (BC0030),[34] and the matrix for distant homolog search (OPTIMA).[19] In addition, we used the BLOSUM approach[31] to construct a matrix, D-BL25, from the alignments in our DAPS training set. We used the commonly used −12/−1 gap penalty to compare the ⟨𝔇⟩ values. We also used the optimal gap penalties given by Vogt et al.[18] when available. For the score function constructed by Blake and Cohen,[34] we used the penalties suggested by the authors. In all cases, superior performance was observed with the optimized STROMA score matrix. Because the coordinate RMSD is calculated by considering only aligned residues, it is possible that STROMA is achieving higher accuracy because it is aligning fewer residues. As shown in Table II, this is not the case; STROMA actually aligns a higher fraction of residues than any of the other matrices.

It is possible that the poorer performance by other score matrices is due to the suboptimal choice of gap penalties. To investigate this possibility, we performed an exhaustive test of different combinations of gap penalties with all the alternative score matrices, identifying the gap penalties that maximized ⟨𝔇⟩ for the DAPS test set. The performance of the various score functions with these optimized gap penalties are also included in Table II. As shown, the increase in accuracy of the alignments generated by STROMA is maintained.

The distribution of 𝔇 values for protein pairs in the DAPS and CE test set are shown as a cumulative distribution in Figure 2. Comparison between STROMA and the other score matrices show a general improvement in the alignment accuracy over the entire test sets. The differences between the alignments is surprisingly large. Only 46% of the amino acid pairs are shared between alignments performed with use of the STROMA and BL62 matrices (with gap penalties equal to 3.4/3). Similarly, only 57% are shared with alignments generated with STROMA and D-BL25.

Figure 3 shows the dependence of ⟨𝔇⟩ on the gap penalties for some of the more popular score matrices. It is of interest that our score matrix is more robust when used

TABLE I. STROMA Score Matrix, Optimized for the Alignment of Known Distant Homologs[†]

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2.5																			
R	0.2	5.2																		
N	1.1	0.7	2.5																	
D	1	0.1	3.3	5.3																
C	1.2	−1.3	−1.9	−3.1	11.5															
Q	−0.1	2	1.9	1.1	−2.5	3.6														
E	1.2	1.9	2.3	3.2	−2.4	1.7	3.7													
G	1.4	−0.2	0.7	0.9	−1.3	−0.3	0.5	7.5												
H	−1.4	1.5	1.4	0.5	−1.7	1.4	0.3	−1.7	6.8											
I	0.3	−1.9	−2.4	−2.9	−3.2	−0.9	−3.1	−3.7	−1.8	4.5										
L	−0.2	−1.5	−2.4	−3.4	−1.6	−1.2	−1.5	−3.8	−2.4	3.4	5.2									
K	−0.2	3.4	1.6	1.4	−3	2.2	1.2	0.4	1.1	−1.5	−2	3.9								
M	−0.2	−1.4	−2.1	−2.8	−1.3	−0.6	−2	−3.8	−0.8	2.2	3.1	−0.5	5.4							
F	−1.6	−3.2	−2.5	−3.7	−0.8	−1.7	−13.7	−4.7	−0.9	2.2	3.7	−2.8	1.7	7						
P	0.7	−0.6	−0.1	−0.2	−3.6	1	0	−0.8	−2.1	−2.4	−1.4	0.2	−1.9	−4.1	8.1					
S	1.7	0.2	1.4	1.7	0.7	0.9	1.1	1.6	−0.1	−1.1	−0.8	1.4	−1.1	−2.5	2	2.8				
T	1.7	0.2	1.4	0.1	0.3	−0.1	1.6	−0.6	−0.2	0	0.3	1	−0.3	−0.8	1.1	2.6	0.4			
W	−3.3	−1.5	−4	−5.7	−0.5	−2.9	−4.7	−4.2	−1.2	−1.8	−1.2	−3	−0.6	3.7	−5	−2.8	−2.9	14.9		
Y	−1.8	−0.9	−0.8	−2.9	−0.3	−1.5	−2.2	−4.8	2.9	0.2	0.8	−1.5	0.5	5.2	−3.3	−0.9	−0.8	4.9	8.1	
V	1.9	−2.8	−0.9	−2.5	0.7	−1.5	−1.3	−1.4	−2.5	4.5	3.4	−1	1.7	0.9	−1.1	−3	1.5	−2.5	0.3	4.2

[†]Corresponding optimized gap penalties are −16.2 to initiate a gap and −1.1 to extend a gap.

TABLE 2. Performance Summary for Commonly Used Score Functions With Various Gap Penalties†

	Gap penalties	DAPS test set		CE test set	
Score matrix	(Initiate/Extend)	$\langle \mathfrak{D} \rangle$	$\langle p(\text{Aligned}) \rangle$	$\langle \mathfrak{D} \rangle$	$\langle p(\text{Aligned}) \rangle$
STROMA	$-16.2/-1.1^{£}$	**1.483**	**0.8582**	**1.461**	**0.9266**
D-BL25	$-12/-1^{§}$	1.704	0.5998	1.529	0.8674
D-BL25	$-2.8/-2.5^{£}$	1.511	0.8399	1.497	0.9141
BLOSUM62	$-12/-1^{§}$	2.021	0.2281	1.660	0.6979
BLOSUM62	$-8.4/-0.9^{¥}$	1.768	0.5891	1.620	0.8055
BLOSUM62	$-3.4/-3^{£}$	1.555	0.7586	1.569	0.8425
BLOSUM30	$-12/-1^{§}$	1.557	0.7870	1.561	0.8677
BLOSUM30	$-11.5/-1.5^{¥}$	1.538	0.8031	1.559	0.8716
BLOSUM30	$-10/-3^{£}$	1.521	0.8079	1.523	0.8917
PAM250	$-12/-1^{§}$	1.657	0.6763	1.567	0.8291
PAM250	$-11.5/-0.5^{¥}$	1.631	0.7082	1.581	0.8260
PAM250	$-4.4/-3^{£}$	1.553	0.7916	1.514	0.8723
PAM350	$-12/-1^{§}$	1.559	0.7811	1.531	0.8727
PAM350	$-7.8/-3^{£}$	1.527	0.8088	1.510	0.8953
PAM500	$-12/-1^{§}$	1.546	0.7990	1.532	0.8837
PAM500	$-6.6/-3^{£}$	1.529	0.8229	1.504	0.9004
GCB	$-12/-1^{§}$	1.625	0.6968	1.533	0.8715
GCB	$-14.2/-0.2^{¥}$	1.680	0.6861	1.602	0.8275
GCB	$-3/-2.9^{£}$	1.533	0.8017	1.524	0.8720
STR	$-12/-1^{§}$	2.046	0.1375	1.730	0.5943
STR	$-9.5/-0.5^{¥}$	1.906	0.4927	1.703	0.7407
STR	$-3.6/-2.5^{£}$	1.600	0.7231	1.571	0.8188
JTT	$-12/-1^{§}$	1.626	0.7189	1.548	0.8574
JTT	$-10.5/-1.5^{¥}$	1.606	0.7369	1.547	0.8615
JTT	$-5.6/-2.4^{£}$	1.537	0.7988	1.522	0.8817
BC0030	$-17/-2$	1.560	0.7868	1.538	0.8643
BC0030	$-20.2/-3^{£}$	1.541	0.8052	1.517	0.8887
OPTIMA	$-12/-1^{§}$	1.802	0.5196	1.612	0.8011
OPTIMA	$-12/-2$	1.943	0.3460	1.614	0.7601
OPTIMA	$-19.6/-3^{£}$	1.725	0.5587	1.764	0.6172

†The accuracy of the alignment is characterized with $\langle \mathfrak{D} \rangle$, the coordinate RMSD averaged over the two test sets, and $\langle P(\text{Aligned}) \rangle$, the average fraction of the residues in the pair of proteins that are aligned. The optimized score function (STROMA) is compared with BLOSUM62,[31] BLOSUM30,[31] PAM250,[21] PAM350,[21] PAM500,[21] GCB,[29] STR,[33] JTT,[32] BC0030,[34] OPTIMA,[19] and D-BL25. All matrices with the exception of STROMA and BC0030 are used with standard BLAST default gap penalties (§), with optimized gap penalties (if available) as given by Vogt *et al.* (¥),[18] and with the gap penalties that maximize $\langle \mathfrak{D} \rangle$ over the DAPS test set (£). BC0030 and OPTIMA are also used with the gap penalties suggested by the authors as well as optimal for the DAPS test set (£). The $\langle \mathfrak{D} \rangle$ values are given with three digits of precision in accordance with precision of the coordinates data in PDB database. All matrices used are nonscaled, as published by their authors. The best performance for each criterea (smallest $\langle \mathfrak{D} \rangle$ and largest $\langle P(\text{Aligned}) \rangle$ are highlighted in bold.

with different combination of gap penalties. As shown in the plot, our score matrix maintained its low RMSD score during most of the gap penalty combinations. On the other hand, the performance of other score matrices tended to vary significantly with changing gap penalties.

DISCUSSION

Most score function construction methods rely on creating a data set of reliably aligned sequences or sequence fragments and gathering statistics on the relative number of times that each possible pairs of amino acid are aligned relative to the null model, that of random pairs of sequences. There are a number of problems with such an approach. For one thing, the statistical approach ignores correlations between various locations in the protein as well as variations in composition. More importantly, the statistics of random pairs of amino acids from nonhomologous protein sequences is an inappropriate null model for random alignments of known homologs. Our approach is to optimize the score matrix to perform the task in which we are interested, generating accurate alignments. We do this by determining a score function that minimizes the coordinate-RMSD-based merit function. By optimizing this merit function over a set of structurally similar protein pairs, we obtain a score matrix that can give us better alignment using pairwise alignment methods. The fact that STROMA performs better than D-BL25, which is built from alignments in our DAPS training set with method used to build BLOSUM matrices,[31] shows the optimization method used to construct STROMA captures more information than the traditional method. It is also interesting that the OPTIMA matrix optimized for homolog identification does not do particularly well in generating accurate alignments, again reinforcing the different statistical natures of these two tasks.

We did observe that the results of the optimization procedure depended on the initial starting matrix, indicat-

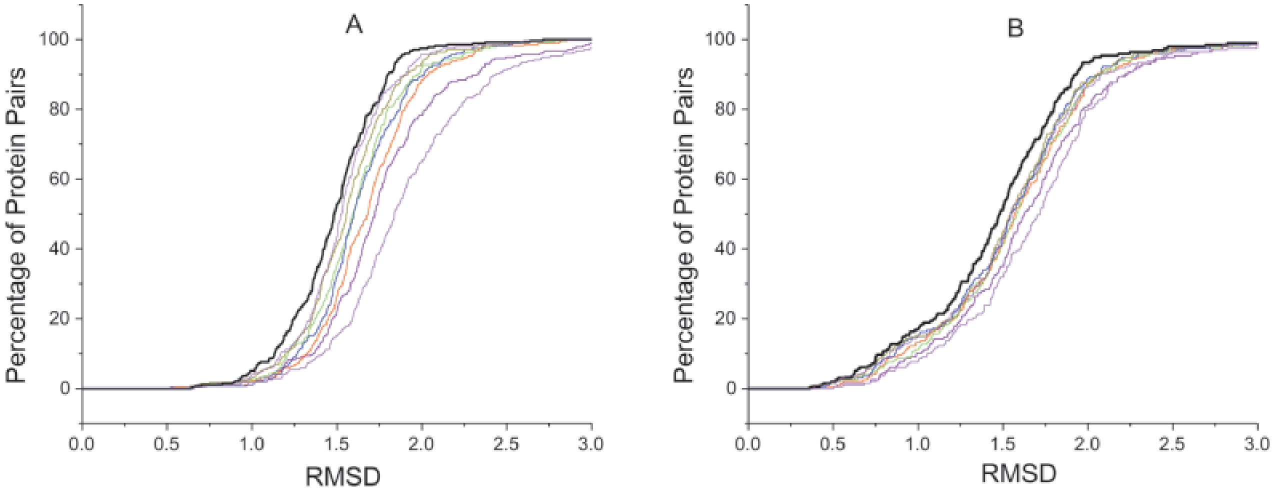

Fig. 2. Cumulative plots of the fraction of protein pairs with greater than a given value of $\mathfrak{D}$ for the DAPS test set (**A**) and CE test set (**B**) Various lines refer to different score matrix and gap penalties: STROMA: thick black; BLOSUM62[31]: thin black; BLOSUM30[31]: purple; GCB[29]: red; JTT[32]: green; PAM250[21]: blue; STR[33]: magenta; and BC0030[34]: dark yellow. Penalties for gap initation and extension are −16.2/−1.1 for STROMA and −17/−2 for BC0030; otherwise, they are the values optimized by Vogt et al.[18] listed in Table I.

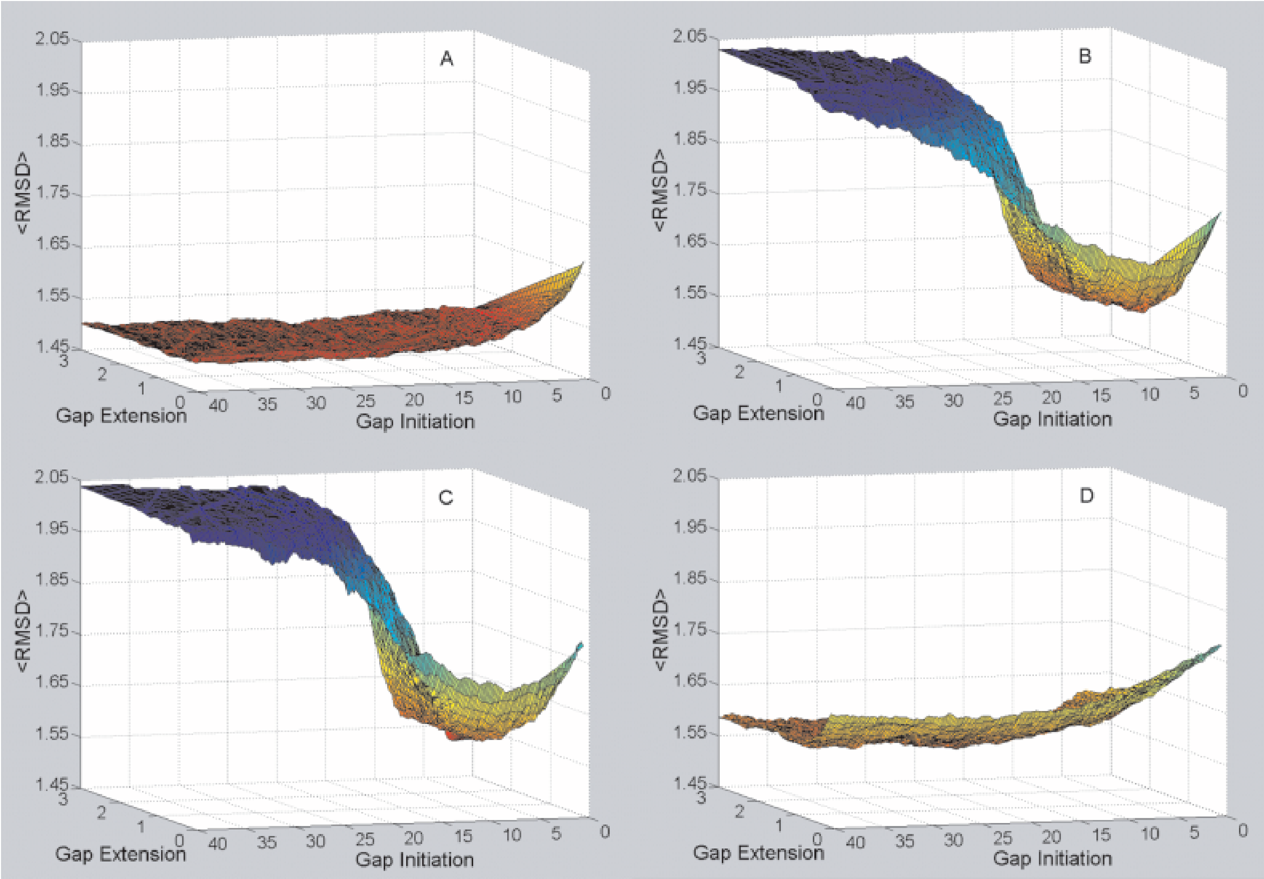

Fig. 3. $\langle \mathfrak{D} \rangle$ for the DAPS test set as a function of the gap penalties for different score matrices. **A:** STROMA; **B:** GCB[29]; **C:** PAM250[21]; **D:** BC0030.[34]

ing that our optimization procedure is likely finding a local minimum in $\langle \mathfrak{D} \rangle$. This is not surprising, because the optimization is performed in a large, "terraced" space where small changes in the score function can produce large changes in the resulting alignments. Although the matrix we provide performs better than other standard matrices, we still might be able to find a better matrix through more complete optimization.

The gap penalty is extremely important in protein sequence alignment, as shown by the big performance fluctuation when using the same score function with different gap penalties. The traditional way of choosing gap penalty is to obtain the score matrix first and then optimize the gap penalty for that score matrix. This empirical treatment of gap penalty is not comparable with its importance. In contrast, we treat the gap initiation and extension penalties as two parameters during the optimization procedure, which can give us more accurate gap penalty values compatible with the 210 score matrix entries. The score matrix generated in this way is more robust and tolerate to gap penalty changes.

ACKNOWLEDGMENTS

We thank Maricel Kann for helpful discussions, and Todd Raeker and Michael Kitson for computational assistance.

REFERENCES

1. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sǎli A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
2. Goldman N, Thorne JL, Jones DT. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J Mol Evol 1996;263:196–208.
3. Thompson MJ, Goldstein RA. Constructing amino-acid residue substitution classes maximally indicative of local protein structure. Proteins 1996;25:28–37.
4. Koshi JM, Goldstein RA. Models of natural mutations including site heterogeneity. Proteins 1998;32:289–295.
5. Lio P, Goldman N, Thorne JL, Jones DT. PASSML: combining evolutionary inference and protein secondary structure prediction. Bioinformatics 1998;14:726–733.
6. Soyer O, Dimmic MW, Neubig RR, Goldstein RA. Using evolutionary methods to study G-protein coupled receptors. In: Altman RB, Dunker AK, Hunter L, Klein TE, editors. Pacific Symposium on Biocomputing '02. Singapore: World Scientific; 2002; 625–636.
7. Benner SA. Predicting de novo the folded structure of proteins. Curr Opin Struct Biol 1992;2:402–412.
8. Shindyalov I, Kolchanov N, Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 1994;7:349–358.
9. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84: 4355–4358.
10. Eddy SR. Hidden Markov Models. Curr Opin Struct Biol 1996;6: 361–365.
11. Karplus K, Sjölander K, Barrett C, Cline M, Haussler D, Hughey R, Holm L, Sander C. Fold recognition using predicted secondary structure, sequences, and Hidden Markov Models of protein folds. Proteins 1997;Suppl 1:123–128.
12. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DL. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
13. Pearson WR, Lipman DJ. Improved tools for biological sequence analysis. Proc Natl Acad Sci USA 1988;85:2444–2448.
14. Altschul SF, Gish W, Miller W, Myers E, Lipman D. Basic local alignment tool. J Mol Biol 1990;215:403–410.
15. Altschul SF, Gish W. Local alignment statistics. Methods Enzymol 1996;215:460–480.
16. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
17. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J Mol Biol 1970;48:443–453.
18. Vogt G, Etzold T, Argos P. An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. J Mol Biol 1995;249:816–831.
19. Kann M, Qian B, Goldstein RA. Optimization of a new score function for the detection of remote homologs. Proteins 2000;41: 498–503.
20. Doolittle RF. Of URFs and ORFs: a primer on how to analyze derived amino acid sequences. Mill Valley: University Science Books. 1986.
21. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. In: Atlas of Protein Sequence and Structure. Vol. 5, suppl. 3. Dayhoff MO, editor. Washington (DC): National Biomedical Research Foundation; 1978. Vol. 5, suppl 3, p. 345.
22. Altschul SF. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol 1991;219:555–565.
23. Lesk AM, Chothia C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol 1980;136:225–270.
24. Rao ST, Rossmann MR. Comparison of super-secondary structures in proteins. J Mol Biol 1973;76:241–256.
25. Mallick P, Rice D, Eisenberg D. DAPS: database of distant aligned protein structures. http://www.doe-mbi.ucla.edu/~parag/DAPS/, 1999.
26. Holm LL, Sander C. Mapping the protein universe. Science 1996;273:595–602.
27. Murzin AG, Brenner SE, Hubbard TJP, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
28. Shindyalov I, Bourne P. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
29. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein database. Science 1992;256:1443–1445.
30. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in C. Cambridge: Cambridge University Press; 1992.
31. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
32. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. CABIOS 1992;8: 275–282.
33. Overington J, Donnelly D, Johnson MS, Sǎli A, Blundell TL. Environment-specific amino-acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci 1992;1:216–226.
34. Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. J Mol Biol 2001;307:721–735.