Optimization of a New Score Function for the Detection of Remote Homologs

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

ABSTRACT **The growth in protein sequence data has placed a premium on ways to infer structure and function of the newly sequenced proteins. One of the most effective ways is to identify a homologous relationship with a protein about which more is known. While close evolutionary relationships can be confidently determined with standard methods, the difficulty increases as the relationships become more distant. All of these methods rely on some score function to measure sequence similarity. The choice of score function is especially critical for these distant relationships. We describe a new method of determining a score function, optimizing the ability to discriminate between homologs and non-homologs. We find that this new score function performs better than standard score functions for the identification of distant homologies. Proteins 2000;41:498–503.** **© 2000 Wiley-Liss, Inc.**

Key words: **sequence alignment; dynamic programming; amino acid sequence; molecular evolution**

INTRODUCTION

The various genome projects have provided us with a plethora of new protein sequences with unknown structure and function. The first step in analyzing such a sequence is often a search for homologous proteins about which more is known, with the expectation that these proteins share a common structure and might have similar functions or mechanisms. In addition, identification of evolutionary relationships can be used to understand the evolutionary history of these sequences, as well as identify selective pressure on different parts of the protein chain. This can provide important clues about structure and function.

Sequence comparisons are now routine, taking advantage of sophisticated software such as FASTA,[1] BLAST,[2,3] and SSEARCH.[1,4] All of these methods rely on the choice of an appropriate method to compute an alignment score, generally of the form

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

where $n_{i,j}$ refers to the number of times that amino acid type i is aligned with amino acid type j, $n_{\text{gap-I}}$ is the total number of gaps in the alignment, $n_{\text{gap-E}}$ is the total number of residues in each gap beyond one, and $\gamma_{i,j}$, $\gamma_{\text{gap-I}}$, and $\gamma_{\text{gap-E}}$ represents 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, while $\gamma_{\text{gap-I}}$, and $\gamma_{\text{gap-E}}$ represent the gap penalties.

For any pair of proteins, the optimal alignment that maximizes the total score can be done quickly, using standard dynamic programming techniques.[5] The maximum possible score for a given pair of proteins is then used to determine whether the two proteins are homologous. This is often done by computing such quantities as $p(S_r > x)$, the probability that a random pair of proteins of the same length would have that score or higher, E, the expected number of random proteins in the database that would have at least that score, and P, the probability that there is at least one random pair with a higher score. Smaller values of $p(S_r > x)$, E, and P indicate a higher likelihood that the given pair is in fact homologous.

The first commonly used score matrices were the PAM (percent accepted mutations) series developed by Dayhoff and co-workers.[6] Others such as those developed by Gonnet et al. (GCB)[7] and Jones et al. (JTT)[8] have applied Dayhoff et al.'s method to larger sequence datasets. Henikoff and Henikoff used a dataset of aligned sequence blocks to construct their popular BLOSUM62 matrix.[9] Overington and coworkers used Henikoff and Henikoffs' cluster method to create a score matrix (STR) where the protein sequences were aligned based on their observed structures.[10]

While identifying closely related homologs is relatively easy with any of the commonly used score matrices, the choice of method becomes more important as the divergence increases.[11] Current matrices can detect homologies among approximately half of all newly discovered genes. It is likely that there are many more distant homologies that still cannot be detected with current score functions. Recently various iterative approaches such as PSI-BLAST[12] have been developed where sets of homologs are used to develop a statistical model that is then used to identify further homologs. Although these approaches are justifiably gaining in popularity, they still rely on the

Grant sponsor: NIH; Grant number LM0577; Grant sponsor: NSF; Grant number BIR9512955.

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

Received 28 February 2000; Accepted 4 August 2000

identification and alignment of an initial set of homologous proteins using standard methods. In addition, these iterative methods require a set of homologs that may not always be available. For these reasons, the development of score functions for pairs of sequences remains important.

Current score functions have a number of limitations. Firstly, the standard alignment score functions assume that each location evolves in a manner independent of all other locations. In reality, many correlations can exist, both because of residues that interact structurally or functionally and through higher-order correlations across extended structures. A more serious limitation is that the score functions are computed in a manner inconsistent with how they are actually used. The score functions are generally derived from a database of properly aligned proteins, often only using sections of the protein that can be confidently aligned. Statistics are not gathered on alignments between random sequences. These score functions are used, however, to discriminate between optimally (but possibly incorrectly) aligned homologs from optimally aligned random sequences, where the optimal alignment is itself dependent upon the score function. There may be significant differences between optimal alignments and the correct alignments. In addition, the alignment is generally performed over significant parts of the proteins involving regions of varying similarity. A score function derived from less variable regions may not be appropriate for an alignment that includes more highly variable regions. Finally, the success of the score function is critically dependent upon the choice of a gap penalty. Unfortunately, it is difficult to calculate a priori what this penalty should be.

In this study, we describe a new procedure to generate a score function optimized to detect distantly related pairs of protein sequences. A training set of distant homologs was developed based on the Cluster of Orthologs Groups (COG) database of Koonin and co-workers.[13] We additionally create four independent test sets of homologous proteins, two representing distant homologs (percentage identity less than 25%) and two with closer homologs (percentage identity between 28 and 40%). We generated alignments between homologous and non-homologous proteins in the training set and maximized the ability of the score function to discriminate between homologous and non-homologous pairs. As the alignments are themselves a function of the score function, this process is iterated. Results with the independent test sets demonstrate the superiority of the resulting score function compared with other commonly used score functions for the detection of distant homologies.

METHODS

Theory

We sequentially align a target protein A_t with each of the proteins in a dataset of size D, achieving a distribution of scores $\{S_r\}$ as computed with Equation 1. The score for the alignment of the target protein and a putative homolog is x. We wish to characterize the significance of this score by calculating the likelihood that this score or higher would be obtained by a random match. We first compute the Z-score, defined as

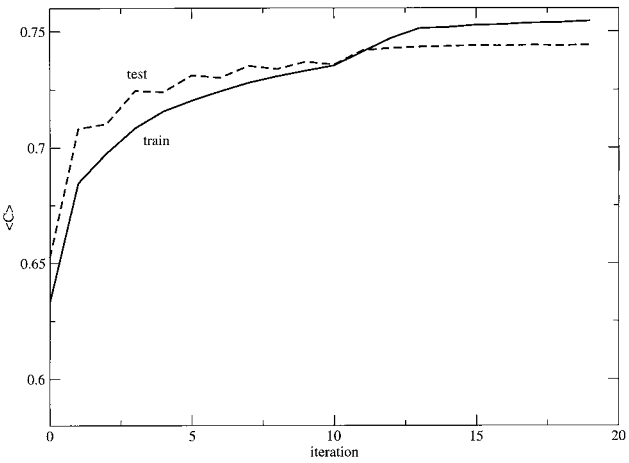

Fig. 1. Value of ⟨C⟩ averaged over the training dataset and test dataset of distant homologs during the optimization procedure.

$$Z = \frac{x - \langle S_r \rangle}{\sqrt{\langle S_r^2 \rangle - \langle S_r \rangle^2}} \qquad (2)$$

where the averages are over the alignments of the target protein with the ensemble of random non-homologous proteins in the dataset. By using the Z score, we automatically account for variations in the expected score with the length of the proteins. In addition, the value of Z will not be appreciably affected if a few accidental distant homologs are included in the set of random proteins.

We can represent the distribution of scores for ungapped[3,14,15] and gapped alignments[16] by an extreme value distribution (EVD).[17] In this case, the probability that a given random score S_r would be equal or greater than x is given by

$$p(S_r > x) = 1 - \int_{-\infty}^{S_r} \rho(x)dx$$

$$= 1 - \exp(-\exp(-\alpha Z - \beta)) \qquad (3)$$

$\alpha = 1.28$ and $\beta = 0.58$ for a perfect EVD,[18] although these parameters are generally adjusted based on the observed distribution. For a search of a database of size D, the expected number of scores between the target protein and random pairs is equal to $E = Dp(S_r > x)$. In this study, we use a value of $D = 100{,}000$. Assuming a Poisson distribution, the probability P of observing at least one alignment with score equal to or greater than x is given by

$$P = 1 - \exp(-E) \qquad (4)$$

Both the E-value and the P-value depend upon the size of the database being searched. E-values range from 0 to D, while P-values range from 0 to 1.

We are interested in optimizing the ability of a score function to discriminate between homologous and non-

TABLE I. OPTIMA Score Matrix Achieved at the Tenth Iteration[a]

A	36																			
R	−9	56																		
N	−19	4	59																	
D	−20	−18	18	65																
C	6	−29	−30	−30	99															
Q	−3	12	2	2	−30	46														
E	−10	3	3	20	−39	19	40													
G	4	−18	7	−10	−29	−20	−23	67												
H	−19	3	12	−7	−29	3	2	−18	86											
I	−5	−28	−32	−34	−6	−30	−33	−41	−28	35										
L	−7	−20	−32	−43	−6	−23	−31	−42	−27	28	32									
K	−10	31	1	−4	−29	15	14	−18	−7	−31	−21	37								
M	−9	−10	−19	−30	−8	1	−21	−30	−19	12	24	−12	51							
F	−19	−30	−29	−33	−18	−29	−32	−32	−8	8	17	−29	2	57						
P	−5	−18	−17	−7	−30	−11	−7	−18	−18	−30	−33	−10	−21	−39	74					
S	12	−11	10	4	−10	0	−1	2	−9	−20	−22	3	−10	−19	−8	36				
T	0	−8	0	−10	−7	−7	−6	−17	−20	−8	−13	−8	−7	−18	−11	18	48			
W	−29	−29	−39	−40	−18	−19	−29	−19	−18	−28	−15	−30	−8	14	−38	−29	−19	110		
Y	−19	−15	−19	−20	−18	−9	−21	−29	20	−8	−2	−17	−9	37	−28	−19	−17	22	69	
V	6	−32	−31	−31	−6	−19	−28	−30	−29	35	18	−23	10	0	−18	−22	6	−28	−9	38
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

[a] The elements are multiplied by ten to increase precision; corresponding gap penalties are −120 and −20.

TABLE II. Comparison of the Various Score Matrices and Gap Penalties on PFAM and COG Test Sets of Distant Homologs (Percentage Identity Less Than 25%) and Closer Homologs (Percentage Identity Between 28% and 40%) as Evaluated With Average Confidence Value (⟨*C*⟩), Average Probability That a Random Score Would Be Higher Than The Known Homolog (⟨$p(S_r > x)$⟩), and Average Probability That at Least One of The Random Scores in a Dataset of 100,000 Proteins Would Be Higher Than The Known Homolog (⟨*P*⟩)[b]

	Gap penalties	COGs distant homologs			COGs close homologs			PFAM distant homologs			PFAM close homologs		
Score matrix	(Initiate/Extend)	$\langle C\rangle$	$\langle p(S_r > x)\rangle$	$\langle P\rangle$	$\langle C\rangle$	$\langle p(S_r > x)\rangle$	$\langle P\rangle$	$\langle C\rangle$	$\langle p(S_r > x)\rangle$	$\langle P\rangle$	$\langle C\rangle$	$\langle p(S_r > x)\rangle$	$\langle P\rangle$
OPTIMA	−11.97/−2.0	0.741	0.004	0.277	0.896	0.010	0.109	0.724	0.008	0.301	0.922	0.002	0.084
BLOSUM62[9]	−12/−2	0.652	0.009	0.372	0.899	0.010	0.107	0.645	0.016	0.376	0.928	0.003	0.078
BLOSUM62[9]	−8/−0.5[22]	0.248	0.024	0.800	0.854	0.020	0.154	0.312	0.033	0.737	0.783	0.001	0.237
PAM250[6]	−12/−2	0.480	0.017	0.549	0.863	0.021	0.144	0.669	0.012	0.359	0.879	0.001	0.133
PAM250[6]	−6/−1.3[22]	0.013	0.072	0.999	0.773	0.040	0.237	0.035	0.061	0.988	0.469	0.005	0.573
GCB[7]	−12/−2	0.647	0.007	0.377	0.874	0.021	0.132	0.703	0.022	0.324	0.884	0.001	0.128
GCB[7]	−7.5/−0.4[22]	0.023	0.061	0.997	0.789	0.036	0.221	0.030	0.042	0.987	0.473	0.006	0.568
STR[10]	−12/−2	0.515	0.035	0.509	0.903	0.010	0.103	0.575	0.033	0.450	0.925	0.007	0.082
STR[10]	−8/−0.5[22]	0.172	0.041	0.866	0.849	0.020	0.158	0.281	0.034	0.774	0.787	0.002	0.233
JTT[8]	−12/−2	0.517	0.009	0.516	0.862	0.023	0.146	0.642	0.022	0.392	0.863	0.001	0.149
JTT[8]	−10.5/−0.4[22]	0.076	0.035	0.958	0.816	0.031	0.191	0.127	0.036	0.916	0.650	0.003	0.375

[b] For the purpose of these comparisons, we use both the standard default gap penalties as well as the gap penalties derived by Vogt and co-workers.[22]

homologous pairs of proteins. That is, we are interested in identifying a true homolog, and in having confidence in this identification. Our confidence in a putative match is equal to the number of correct matches divided by the number of matches, both correct and incorrect, with the same score or higher. Assuming that we have one true homolog in the dataset, the average confidence *C* can be quantified as

$$C = \frac{1}{1+E} = (1 + D(1 - e^{-\exp(-\alpha Z - \beta)}))^{-1} \qquad (5)$$

A *C* value close to 1 indicates a confident alignment, with *C* decreasing to $1/(1 + D)$ as the confidence of the alignment decreases. This represents our average relative chance that the match is to a true homolog. In this study, we optimize the score function by maximizing $\langle C\rangle$ averaged over the training set. By optimizing $\langle C\rangle$ we automatically focus on homologous pairs at the limit of detection, reducing the dependence of the score function on homologies that are either easily detectable ($E \ll 1$) or overly distant ($E \sim D$). In addition, optimization of $\langle C\rangle$ eliminates the contribution of falsely identified homologies in the training dataset, as these would presumably be at the overly distant limit.

Database Preparation

We are interested in optimizing our score functions for the detection of distant homologs, beyond the capability of current score functions. We, therefore, need a set of known homologs whose homology cannot be reliably determined with standard pairwise sequence comparisons. For this purpose, we took advantage of the Cluster of Orthologs Genomes (COG) database of Koonin and co-workers.[13] A 900-pair training set was constructed of pairs of proteins in the same COG but with less than 25% sequence identity.

It is always important to validate the results of such optimization procedures with independent test sets. For this study, we developed four such sets. A 177-pair test set was constructed in aI similar manner from the COG database, excluding all COGs that contributed to the training set, with each pair of proteins again having less than 25% sequence identity. In order to evaluate performance on a set of closer homologs, we similarly created an independent group of 900 protein pairs from COG with between 28 and 40% sequence identity. We also desired to develop test sets constructed independently of the method used to construct the training set. For this purpose, we took pairs of proteins identified as homologs in the PFAM database release 5.2.[19] In order to avoid overlap with the training and test sets derived from the COG database, we ran a BLAST search[2] (using BLOSUM62[9] with -12,-2 for the gap penalties) of all the sequences in the PFAM database against the 1,077 pairs from the COGs that we were using either as the training set or first test set, and excluded all PFAM families with any member with similarity to these proteins (i.e., $E < 10$). From this set of protein sequences, we selected 103 pairs that share less than 25% sequence identity as a third test set of distantly related sequences, and 362 pairs with between 28 and 40% sequence identity as a fourth test set. The proteins in the training and test sets are available as supplementary material.

Optimization of the Score Function

We are interested in maximizing the confidence value $\langle C \rangle$ averaged over proteins in the training set, where the calculation of C involves the distribution of scores for the optimal alignment of the target proteins with homologous and non-homologous proteins in the dataset. These optimal alignments are themselves dependent upon the value of the score function. Thus, an iterative scheme is required.

We started with the BLOSUM62 matrix[9] and used the local dynamic programming algorithm[5] to align each of the target proteins in the training dataset against a homolog and a set of non-homologous proteins with a large number of different gap penalties. We then calculated Z and C for each pair of homologs, and averaged over the pairs in the training set to yield $\langle C \rangle$. The highest C values were obtained with gap penalties of $\gamma_{\text{gap-I}} = -12.0$ and $\gamma_{\text{gap-E}} = -2.0$. This scoring scheme (BL62(12,2)) was then used to generate an initial set of alignments of the target proteins with homologous and non-homologous proteins. The observed distributions of the non-homologous proteins were used to adjust the values of $\alpha = 1.31$ and $\beta = 0.74$, similar to the values expected ($\alpha = 1.28$ and $\beta = 0.58$) for a perfect EVD.[18]

As multiplication of the score function by any constant does not change Z or any of the other statistics, we fixed one score function ($\gamma_{\text{gap-I}}$) equal to -2.0, resulting in 211 adjustable parameters corresponding to the 210 possible pairs of amino acid types and the remaining gap penalty. Starting with the BL62(12,2) score scheme and the corresponding set of aligned sequences, we analytically calculated the approximate direction of a steepest descent for the adjustable parameters assuming the alignments remained unchanged. We then adjusted the parameters along that direction, realigning the sets of sequences at every point, until Armijo's rule was satisfied.[20] The next appropriate direction of steepest descent was then recalculated. Approximately 10 cycles of optimization and realignments were performed until the score function converged. Implementation of a Simplex optimization procedure[21] gave similar results. Performance was monitored by simultaneous calculations of $\langle C \rangle$ averaged over the proteins in the test set of distant homologs derived from the COG database. The statistics with the optimized score function indicated that the appropriate values of α and β did not appreciably shift.

RESULTS

The values of $\langle C \rangle$ as averaged over the training set and distant COG homolog test set during the optimization process are shown in Figure 1. The resulting score function (OPTIMA) obtained after 10 iterations is shown in Table I. The optimal value for $\gamma_{\text{gap-I}}$ was -11.97 with $\gamma_{\text{gap-E}}$ fixed at -2.0. The small change in the gap penalties indicates that most of the improvement comes from refinements of the values of $\gamma_{i,j}$. As shown in Table II, OPTIMA has a significantly improved average confidence ($\langle C \rangle$) value compared with other commonly used score matrices. This improvement is not confined only to values of C; both

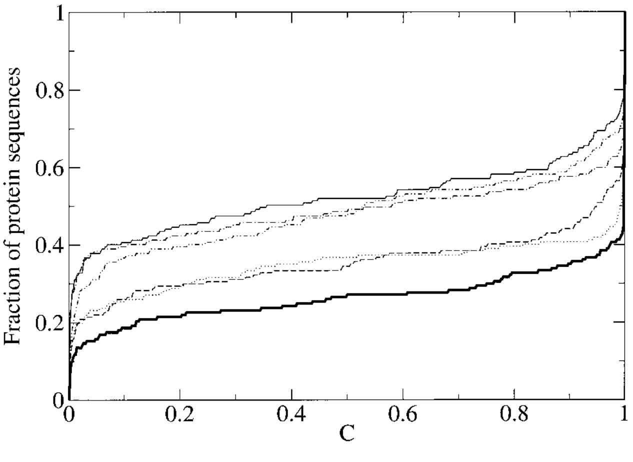

Fig. 2. Cumulative distribution of the C values for the various score matrices, showing the fraction of all protein pairs in the COGs test set of distant homologs with less than a given value of confidence. The various lines refer to the OPTIMA score matrix (—); BLOSUM62 (12/2)[9] (· · ·); GCB (12/2)[7] (---); STR (12/2)[10] (—·—·—·—); JTT (12/2)[8] (—··—··—); and PAM250 (12/2)[6] (—).

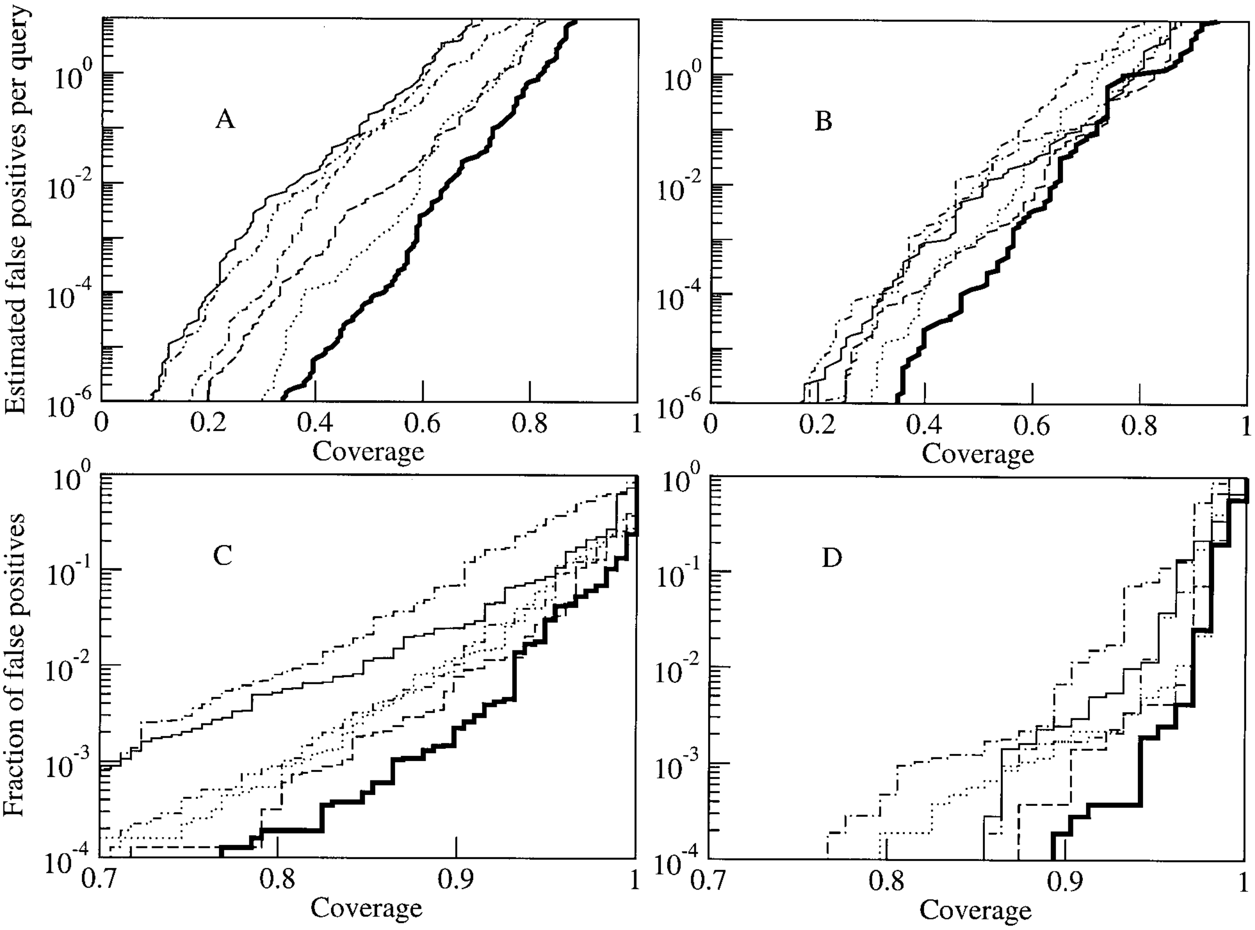

Fig. 3. **A:** Expected number of false positives of different score matrices as a function of the number for protein sequences pairs (Coverage) for the COG test dataset of distant homologs. The plots show the fraction of the homologous pairs with fewer than a given number of false positives of higher score (E) expected for D = 100,000. **B:** Similar plot for the PFAM test dataset of distant homologs. **C:** Coverge vs. number of false positives for the COG test dataset of distant homologs. For this plot, we calculated the E values for each homologous and non-homologous pair of proteins in the test set, and then ranked these values from lower to higher. We then considered all matches with a value of E lower than a given cut-off value, and calculated the fraction of the true positives included in this set (Coverage) as well as the fraction of non-homologous pairs also included (Fraction of false positives). These two values were then plotted as a parametric plot. This approach, applied to homology detection by Brenner et al.,[11] is related to the Receiver Operating Characteristic (ROC) measure.[23,24] **D:** Similar plot for the PFAM test dataset of distant homologs. All of the curves are designated as in Figure 2.

$\langle p(S_r > x)\rangle$, the average probability that any random score would be higher than the homolog, as well as $\langle P\rangle$, the average probability that at least one random score is higher than the homolog, are both substantially decreased compared with other matrices.

Figure 2 shows the cumulative distribution of C values for COG distant-homolog test set with the different score functions. As shown, the greater discriminatory power of the OPTIMA score function is represented by the larger fraction of the protein sequence pairs having larger values of C. That implies a substantial improvement in our ability of making confident predictions compared with other standard score functions.

Figure 3A and B show the coverage or fraction of true positives vs. the estimated number of false positives per query for the distant homolog COGs and PFAM test sets, respectively, where the estimated number of false positives per query represents the expected number of random sequences with a score greater than the pair of homologous sequences. The better performance of OPTIMA can be seen from the large number of homologous pairs with a lower estimated number of false positives.

As a further test, we constructed a parametric plot where we calculated the fraction of true positive homologs identified (coverage) and the fraction of non-homologous pairs identified incorrectly as homologies (fraction of false positives) as a function of the cut-off value of E. The results are shown for the distant homolog COGs and PFAM test sets in Figure 3C and D, respectively. While this parametric plot may be compromised by the presence of true homologs incorrectly labeled as non-homologs (false false-positives) in the test sets, the qualitative agreement with the previous plots further supports the performance of OPTIMA compared with the other score functions.

Although OPTIMA was optimized for the detection of distant homologs, the resulting matrix was still among the top score matrices in terms of $\langle C \rangle$, $\langle p(S_{r} > x) \rangle$, and $\langle P \rangle$ when applied to the test sets of closer homologs (percentage identity between 28 and 40%), as shown in Table II.

DISCUSSION

Most methods for constructing a score function rely on creating a dataset of reliably aligned sequences or sequence fragments and gathering statistics on the relative number of times that each possible pair of amino acids are aligned. In practice, however, we are interested in distinguishing optimally (but possibly incorrectly) aligned homologs from optimal alignments of non-homologs. Our approach towards generating a score matrix is to optimize the ability of this matrix to do what we want to do: discriminate between homologs and non-homologs. In order to do this, we derived a measure of merit of the score function, the average confidence of the homolog identification, and maximized this measure over a set of homologous and non-homologous pairs of proteins. Different measures of merit can be handled in a similar way. Our score function still represents statistics derived from real homologous protein sequences, and can be analyzed in terms of evolutionary substitutions and the physical-chemical properties of the amino acids. In contrast to standard derivations, the gap penalties can be treated as additional parameters to be optimized. In tests with two disjoint set of test proteins, we are able to demonstrate that this score function achieves greater success at discriminating between homologs and non-homologs compared with standard score matrices.

ACKNOWLEDGMENTS

We thank Stephen Altschul, Ting-Lan Chiu, William Pearson, Romesh Saigal, Stephen Bryant, and John Spouge for helpful discussions, and Todd Raeker and Michael Kitson for computational assistance.

REFERENCES

1. Pearson WR, Lipman DJ. Improved tools for biological sequence analysis. Proc Natl Acad Sci USA 1988;85:2444–2448.
2. Altschul SF, Gish W, Miller W, Myers E, Lipman D. Basic local alignment tool. J Mol Biol 1990;215:403–410.
3. Altschul SF, Gish W. Local alignment statistics. Methods Enzymol 1996;215:460–480.
4. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science 1985;227:1435–1441.
5. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
6. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure, vol. 5. Silver Springs; National Biomedical Research Foundation; 1978. p. 345–352.
7. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein database. Science 1992;256:1443–1445.
8. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. CABIOS 1992;8: 275–282.
9. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
10. Overington J, Donnelly D, Johnson MS, Šali A, Blundell TL. Environment-specific amino-acid substitution tables: tertiary templates and prediction of protein folds. Prot Sci 1992;1:216–226.
11. Brenner SE, Chothia C, Hubbard TJP. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 1998;95:6073–6078.
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. Tatusov RL, Galperin MY, Koonin EV. The COG database: a tool for genome-scale analysis of proteins functions and evolution. Nucleic Acids Res 2000;28:33–36.
14. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 1990;87:2264–2268.
15. Dembo A, Karlin S, Zeitouni O. Limit distribution of maximal non-aligned two-sequence segmental score. Ann Prob 1994;22: 2022.
16. Pearson WR. Empirical statistical estimates for sequence similarity searches. J Mol Biol 1998;276:71–84.
17. Gumbel EJ. Statistics of Extremes. New York: Columbia University Press; 1958.
18. Gumbel EJ. Statistics theory of extreme values and some practical applications. Washington, DC: U.S. Government Printing Office: National Bureau of Standards Applied Mathematics Series 33; 1954.
19. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer ELL. Pfam 3.1: 1313 multiple alignments match the majority of proteins. Nucleic Acids Res 1999;27:260–262.
20. Dennis JE Jr, Schnabel RB. Numerical methods for unconstrained optimization and nonlinear equations. New York: Prentice-Hall; 1983.
21. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in C. Cambridge: Cambridge University Press; 1992.
22. 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.
23. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986;21:720–733.
24. Swets JA. Measuring the accuracy of diagnostic systems. Science 1988;240:1285–1293.