Optimizing for Success: A New Score Function For Distantly Related Protein Sequence Comparison

Maricel Kann[†] and Richard A. Goldstein[†‡]
[†] Department of Chemistry,[‡]Biophysics Research Division,
University of Michigan, Ann Arbor, MI 48109.
mkann@umich.edu, richardg@umich.edu

Abstract

The exponential growth of the sequence data produced by the genome projects motivates the development of better ways of inferring structural and functional information about those newly sequenced proteins. Looking for homologies between these probe protein sequences and other protein sequences in the database has proved to be one of the most useful current techniques. This procedure, known as sequence comparison, relies on the use of an appropriate score function that discriminates homologs from non-homologs. Current score functions have difficulty identifying distantly-related homologs with low sequence similarity. As a result, there is an increased demand for a new score function that yields statistically-significant higher scores for all the pairs of homologous protein sequences including such distantly-related homologs. We present a new method for generating a score function by optimizing it for successful discrimination between homologous and unrelated proteins. The new score function (OPTIMA) outperforms other commonly used substitution matrices for the detection of distantly related protein sequences.

1 Introduction

Finding similarities between newly-determined protein sequences and existing sequences in the protein database provides us with access to an enormous amount of information. Since highly similar protein sequences generally share a common ancestor, sequence comparisons can shed light over the evolutionary history of these new protein sequences. These evolutionary related proteins or homologs share the same structure and may have the same function. By detecting these relationships, costly and time-consuming experimental techniques to determine structure and function of these proteins can be avoided. The advent of high performance computers and rapid search and sequence comparison algorithms during the last 20 years has made such searches a routine task, using programs such as FASTA [15], BLAST [2, 1], PSI-BLAST [3], and SSEARCH [15, 13]. Each of these algorithms furnish an alignment score representing the number of identical, similar, and dissimilar amino acids aligned as well as the number of gaps in the alignment. This score is used to identify the likelihood that the two sequences are evolutionary related. All of these methods rely on the choice of an appropriate score function.

The series of PAM (percent accepted mutations) matrices, based on the work of Dayhoff [4] are still among the most commonly used score functions, specially PAM250 (P250). Gonnet [6] et al and Jones [10] published matrices (GCB and JTT respectively) based in an enlarged sequence databases using methods similar to that of Dayhoff. Henikoff derived a series of matrices from a database of aligned sequence blocks, with BLOSUM 62 being the most commonly used of the series [9]. Overington and coworkers generated a matrix (STR) using a structure-based approach in combination with the cluster method developed by

RECOMB 2000 Tokyo Japan USA

Henikoff [14]. The improvements in these matrices made possible the detection of similarities among approximately half of the new genes discovered. There are many distant homologies that still cannot be identified with currently available scoring functions. The more sequence divergence, the more the choice of score function becomes critical. In this paper, we describe a new procedure to generate a score function optimized to detect distantly related pairs of protein sequences. We find that the new score function described in this paper give better results on distantly related pairs of homologs than other commonly-used score functions.

2 Methods

2.1 Theory

The score (S) for any alignment is given by the sum of the weights of the letters paired together, defined by:

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

where $n_{i,j}$ refers to the number of times that amino acid type i is aligned with amino acid type j, and $\gamma_{i,j}$ represents the score given for such a match or mismatch. The number of times that a gap occurs and the penalties for starting a gap is represented by $n_{\text{gap-I}}$ and $\gamma_{\text{gap-I}}$ respectively, while $n_{\text{gap-E}}$ and $\gamma_{\text{gap-E}}$ counts and penalizes the extension of those gaps. $\gamma_{i,j}$ is known as the score function, substitution matrix, or exchange residue matrix.

Imagine x is the score for the alignment of the target A_t with another possibly homologous sequence. We are interested in evaluating the statistical significance of this score, to accertain the appropriate confidence we should have that this represents a true homology. Consider that the scores for alignment of the target with non-homologous sequences is represented as S_r. We can use a Z-score to express the score of our possible match relative to the distribution of scores for alignments with non-related protein sequences as follows

$$Z = \frac{x - \langle S_r \rangle}{\sqrt{\langle S_r^2 \rangle - \langle S_r \rangle^2}} = \frac{\Delta}{\Gamma}$$

The distribution in the score function for ungapped alignments has been well studied [5, 11] and can be analytically represented by an extreme value distribution (EVD) [7]. For local alignments with gap allowance, it is possible to approximate the score distribution with an EVD. The probability that a random score S_r would equal or surpassing the observed score (x) by chance is given by [8]

$$\begin{aligned} p(S_r > x) &= 1 - \int_{-\infty}^{S_r} \rho(x)\, dx \\ &= 1 - \exp\left(-\exp\left(-1.282Z - 0.5772\right)\right) \end{aligned}$$

For a search of a database of size D, the average number of the resulting D random alignments with scores greater than x (false positives) is equal to $E = D\, p(S_r > x)$. Applying a Poisson distribution, the probability P of observing at least one alignment with score equal or greater than x is given by $P = 1 - \exp(-E)$.

The E-values are usually reported when using BLAST and PSI-BLAST searches and can range from 0 to D; P-values range from 0 to 1.

In this paper we define C as a measure of "success", to be maximized during the optimization. Our "success" in recognizing the homolog, represented by C, is based in the relative chance the match is a true match versus a random false positive. If we have one true homolog, our confidence in the match would be equal to the number of correct matches (N_C) divided by the number of incorrect ones (E). Assuming we are interested in evaluating whether we have found one true homolog, $N_C = 1$ and

$$C = \frac{N_C}{N_C + N_I} = \frac{1}{1 + E}$$

When a good score function is used to perform the alignments, the C values should ideally approach to 1.0, and the number of false negatives is minimal. Thus, we optimize the confidence $\langle C \rangle$ averaged over the data set.

2.2 Preparation of databases

We are interested in optimizing our score matrices for the identification of distant homologs that like in the "twilight zone", where standard score functions are inadequate. For this reason, we need an appropriate protein sequence database of known but distant homologs. We constructed a 234-pair training-set database using the Cluster of

Orthologs Genomes (COG) database assembled by Koonin and co-workers [12, 18] so that each pair shared less than 25% sequence identity. The COGS database allows us to assemble distant homologs liked by a network of more obvious homologies. Proteins from other COGs were chosen to represent random, non-homologous matches. A similar but disjoint 91-pair test set was also constructed in order to monitor the ability of the score function to detect homologies not contained in the training set. No more than one pair of proteins were taken form each COG.

2.3 Algorithms and programs

Our approach for the optimization of the score function is directed at maximizing the average confidence $\langle C \rangle$ in the set of correct homolog identifications, averaged over all of the matches in the training set. The optimization of score function requires an iterative method, since the alignments depend upon the score function which can only be optimized for a given alignment. Starting with the BLOSUM 62 matrix [9], we used the local dynamic programming approach with affine gaps [17] to align each of the target proteins in the training data set against a homologous and a set of non-homologous proteins, and calculated Z, P, E, and C for each pair of homologs, and thus $\langle C \rangle$. This was done for a large number of different gap penalties. The highest C values were obtained with gap penalties of -12 and -2 for the gap initialization ($\gamma_{\text{gap-I}}$) and extension ($\gamma_{\text{gap-E}}$) respectively. This scoring scheme (represented as BL62(12/2)) was our initial scoring function at the start of the optimization procedure, and was used to generate an initial set of alignments. The substitution matrix, including the gap penalties, was then adjusted with the current alignment to increase $\langle C \rangle$. As the matrix is symmetric (that is, the alignment of amino acid types i and j is counted the same as the alignment of types j and i) there are 212 adjustable parameters consisting of the 210 possible pairs of amino acid types and the two gap penalties. As multiplication of the score function by any constant does not change Z or any of the other statistics, we set one score function ($\gamma_{\text{gap-I}}$) equal to a constant, resulting in 211 adjustable parameters. After a few iterations of updating the matrix, the new matrix was used to realign the homologs and random pairs. In every cycle of the optimization, we used the steepest descent method To assure convergence to the maximal, Armijo and Goldstein criteria [16] was used to choose the step size in each cycle. Approximately 10 cycles of optimization and re-alignments were performed until the score function converged.

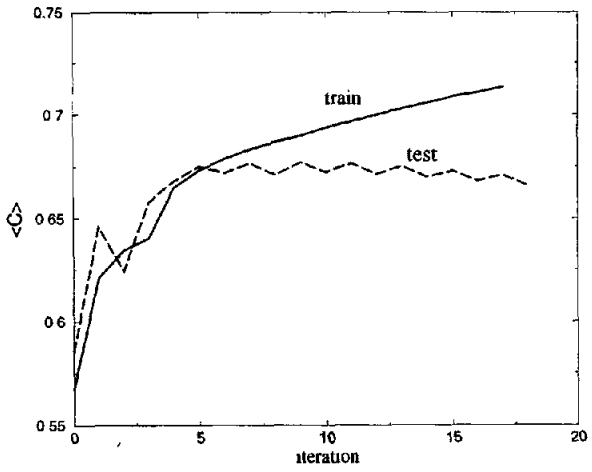

Figure 1: Average value of C for the training data set (234 pairs of protein sequences) and test data set (91 pairs of protein sequences). Increasing $\langle C \rangle$ during the optimization procedure represent an improvement in the confidence to find the true match among the non-related protein sequence pairs.

3 Results

The procedure described in this paper increases $\langle C \rangle$ as averaged over the training set, as shown in Figure 1. The main question is whether this increased discrimination between homologs and non-homologs also occurs for the training-set of proteins, those not included in the training set. Figure 1 shows that our ability to discriminate, as measured by $\langle C \rangle$, also is increased in the independent test set. We can conclude that we successfully optimized without overfitting to the training data. The new score function (OPTIMA) obtained after 10 iterations is shown in Table 1. Optimal gap penalties were -11.97 and -2.0, not far from the initial values, indicating that the better performance is due to optimal values of $\gamma_{i,j}$. The effect of changing the distribution of the scores during the opti-

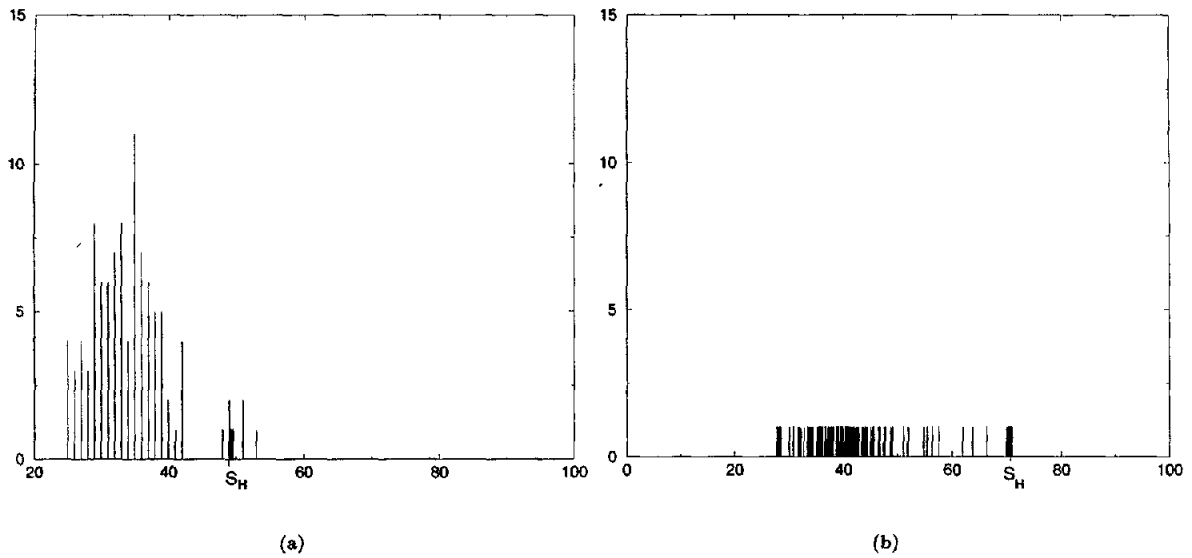

Figure 2: An example of the score distribution (a) using BL62(12,0.2) and (b) OPTIMA. The higher score for the homolog (S_H) obtained with OPTIMA, makes possible to discriminate it from all the of non-homologs pairs.

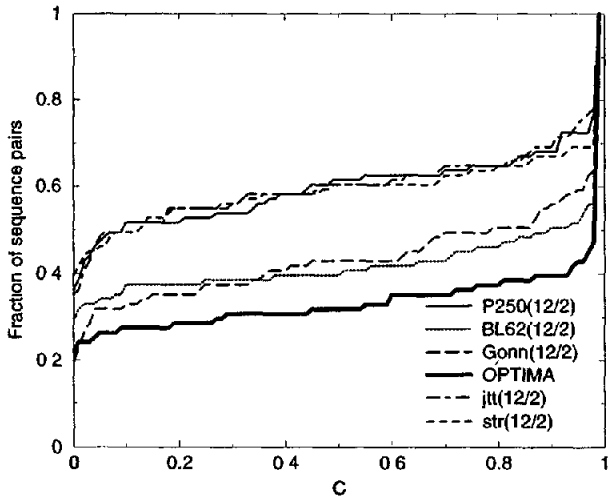

Figure 3: Cumulative plot of the distribution of C values for various score matrices, showing the fraction of all protein pairs in the test set recognized with under a given value of confidence. The optimized score function is displayed in a thicker line.

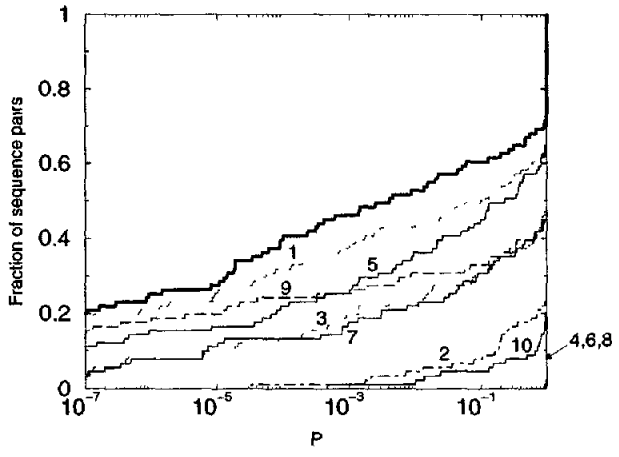

Figure 4: Cumulative plot of the distribution of P-values for the optimized score function (thicker line) and other commonly used score functions.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	A
		5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	R
A	31		6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	N
R	-9	53		6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-2	-3	D
N	-20	-1	56		9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	C
D	-20	-19	13	65		5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	Q
C	2	-30	-28	-29	90		5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	E
Q	-11	13	5	0	-29	44		6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	G
E	-14	0	0	26	-39	27	42		8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	H
G	4	-15	11	-15	-26	-22	-18	73		4	2	-3	1	0	-3	-2	-1	-3	-1	3	I
H	-21	7	10	-4	-30	6	0	-16	88		4	-2	2	0	-3	-2	-1	-2	-1	1	L
I	-9	-29	-33	-32	-5	-30	-34	-42	-30	37		5	-1	-3	-1	0	-1	-3	-2	-2	K
L	-9	-22	-30	-41	-2	-13	-31	-42	-30	30	31		5	0	-2	-1	-1	-1	-1	1	M
K	-4	30	-4	0	-30	17	9	-18	-9	-33	-19	43		6	-4	-2	-2	1	3	-1	F
M	-6	-9	-20	-31	-8	2	-18	-30	-18	12	23	-9	52		7	-1	-1	-4	-3	-2	P
F	-21	-30	-31	-32	-20	-33	-31	-29	-5	7	16	-30	0	59		4	1	-3	-2	-2	S
P	-14	-20	-18	-7	-29	-9	-5	-20	-17	-29	-29	-3	-19	-39	83		5	-2	-2	0	T
S	13	-8	9	1	-7	0	2	3	-10	-25	-24	-1	-11	-18	-9	43		11	2	-3	W
T	3	-12	4	-9	-7	-4	-5	-21	-20	-8	-15	-6	-7	-16	-9	14	44		7	-1	Y
W	-30	-29	-40	-39	-17	-18	-30	-22	-19	-27	-17	-25	-11	14	-41	-29	-18	109		4	V
Y	-19	-16	-18	-19	-19	-10	-20	-28	22	-11	0	-20	-7	37	-30	-20	-20	22	67		
V	5	-29	-32	-30	-4	-18	-26	-33	-29	33	19	-20	7	-6	-16	-21	-1	-28	-8	37	
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	

Table 1: BLOSUM 62 substitution matrix (upper side) and the optimized substitution matrix OPTIMA (lower side). In order to increase the precision, the elements in OPTIMA were multiplied by 10 and then rounded to the nearest integer.

mization is depicted in Figure 2 and explained below. As shown in Figures 2a, the initial score matrix (BL62(12/2)) used to score the pairwise alignments in this example from the training database does not allow discrimination between homologs and non-homologs, with many non-homologous pairs giving higher alignment scores than the pair of true homologs. On the other hand, as shown in Figure 2b, when using the optimized score function (OPTIMA), the scores for the homologs are clearly separated from the scores for the non-homologous pairs, making discrimination possible.

Figure 3 shows the cumulative distribution of C values for the test-set proteins with the different score functions. As shown, the greater discriminatory power of the OPTIMA score function is represented by the larger fraction of the proteins sequences pairs having greater values of C when using OPTIMA to perform the alignments. That implies a substantial improvement in our ability of making confident predictions compare with other standard score function. Figure 4 shows a corresponding cumulative distribution of the P-values, which represent the probability of having one or more random sequences with a score greater than a pair of homologous sequences. The better performance of OPTIMA can be seen from the large number of homologous pairs with lower P-values.

4 Conclusion

Most methods for constructing a score function 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 pair of amino acids are aligned. In practice, however, we are interested in distinguishing more distant homologs that may be impossible to accurately align from optimal alignments of non-homologs. It is assumed that statistics based on the regions of homologs that can be confidently aligned can be extrapolated to possibly incorrect alignments between more variable regions. In contrast our approach is based on considering the alignments made by the score function in generating the statistics. As the alignments are dependent on the score function, this requires an iterative procedure as described above. Another drawback of previously-existing matrices is the difficulties in choosing the gap penalty. In this paper we included the optimal gap penalty it in the optimization as one more pairwise scoring function element.

Our method combines the use of a new database with the optimization of $\langle C \rangle$-values, the inclusion of gaps and the possibility of changing the alignments between iterations. These are some of the features in the procedure presented that make it

ideal for obtaining the optimal score scheme for sequence comparison, and the reason for the outstanding performance at detecting distantly related protein sequences of OPTIMA compared with other commonly-used scoring schemes.

5 Acknowledgments

We thank R. Saigal and T-L. Chiu for for their insigth in the optimization procedure, and S. Altschul and S.Bryant for valuable discussions. Financial support was provided by Horace H. Rackham School of Graduate Studies, NIH Grant LM0577 and NSF equipment grant BIR9512955.

References

[1] S.F. Altschul and W. Gish. Local alignment statistics. *Methods Enzymol.*, 266:460–480, 1996.

[2] S.F. Altschul, W. Gish, W.Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[3] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W.Miller, and D.J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.

[4] M. O. Dayhoff, R. M Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, suppl. 3, page 345. National Biomedical Research Foundation, Washington, D.C., 1978.

[5] A. Dembo, S. Karlin, and O. Zeitouni. Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Prob.*, 22:2022, 1994.

[6] G. H. Gonnet, M. A. Cohen, and S. A. Benner. Exhaustive matching of the entire protein database. *Science*, 256:1443–1445, 1992.

[7] E. J. Gumbel. *Statistics of Extremes.* Columbia University Press, New York, 1958.

[8] E.J. Gumbel. *Statistics Theory of Extreme Values and Some Practical Applications.* National Bureau of Standards Applied Mathematics Series 33. Washington: U.S. Government Printing Office.

[9] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Nat. Acad. Sci., U.S.A.*, 89:10915–10919, 1992.

[10] D. T. Jones, W. R. Taylor, and J. M Thornton. The rapid generation of mutation data matrices from protein sequences. *CABIOS*, 8:275–282, 1992.

[11] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Nat. Acad. Sci., U.S.A.*, 87:2264–2268, 1990.

[12] E.V. Koonin, R.L. Tatusov, and M.Y. Galperin. Beyond complete genomes: from sequence to structure and function. *Curr. Op. Struc. Bio.*, 3:355,363, 1998.

[13] D. J. Lipman and W. R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227:1435–1441, 1985.

[14] J. Overington, D. Donnelly, M. S. Johnson, Andrej Šali, and T. L. Blundell. Environment-specific amino-acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci.*, 1:216–226, 1992.

[15] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence analysis. *Proc. Nat. Acad. Sci., U.S.A.*, 85:2444–2448, 1988.

[16] J.E. Dennis Jr.and R.B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice-Hall, New York, 1983.

[17] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.

[18] R.L. Tatusov, E.V. Koonin, and D.J. Lipman. A genomic perspective on protein families. *Science*, 278:631,637, 1997.