

Amino Acid Substitution Matrices from an Information Theoretic Perspective

Stephen F. Altschul

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, MD 20894, U.S.A.

(Received 1 October 1990; accepted 12 February 1991)

Protein sequence alignments have become an important tool for molecular biologists. Local alignments are frequently constructed with the aid of a "substitution score matrix" that specifies a score for aligning each pair of amino acid residues. Over the years, many different substitution matrices have been proposed, based on a wide variety of rationales. Statistical results, however, demonstrate that any such matrix is implicitly a "log-odds" matrix, with a specific target distribution for aligned pairs of amino acid residues. In the light of information theory, it is possible to express the scores of a substitution matrix in bits and to see that different matrices are better adapted to different purposes. The most widely used matrix for protein sequence comparison has been the PAM-250 matrix. It is argued that for database searches the PAM-120 matrix generally is more appropriate, while for comparing two specific proteins with suspected homology the PAM-200 matrix is indicated. Examples discussed include the lipocalins, human α_1 B-glycoprotein, the cystic fibrosis transmembrane conductance regulator and the globins.

Keywords: homology; sequence comparison; statistical significance; alignment algorithms; pattern recognition

1. Introduction

General methods for protein sequence comparison were introduced to molecular biology 20 years ago and have since gained widespread use. Most early attempts to measure protein sequence similarity focused on global sequence alignments, in which every residue of the two sequences compared had to participate (Needleman & Wunsch, 1970; Sellers, 1974; Sankoff & Kruskal, 1983). However, because distantly related proteins may share only isolated regions of similarity, e.g. in the vicinity of an active site, attention has shifted to local as opposed to global sequence similarity measures. The basic idea is to consider only relatively conserved subsequences; dissimilar regions do not contribute to or subtract from the measure of similarity. Local similarity may be studied in a variety of ways. These include measures based on the longest matching segments of two sequences with a specified number or proportion of mismatches (Arratia *et al.*, 1986; Arratia & Waterman, 1989), as well as methods that compare all segments of a fixed, predefined "window" length (McLachlan, 1971). The most common practice, however, is to consider segments of all lengths, and choose those that optimize a

similarity measure (Smith & Waterman, 1981; Goad & Kanehisa, 1982; Sellers, 1984). This has the advantage of placing no *a priori* restrictions on the length of the local alignments sought. Most database search methods have been based on such local alignments (Lipman & Pearson, 1985; Pearson & Lipman, 1988; Altschul *et al.*, 1990).

To evaluate local alignments, scores generally are assigned to each aligned pair of residues (the set of such scores is called a substitution matrix), as well as to residues aligned with nulls: the score of the overall alignment is then taken to be the sum of these scores. Specifying an appropriate amino acid substitution matrix is central to protein comparison methods and much effort has been devoted to defining, analyzing and refining such matrices (McLachlan, 1971; Dayhoff *et al.*, 1978; Schwartz & Dayhoff, 1978; Feng *et al.*, 1985; Rao, 1987; Risler *et al.*, 1988). One hope has been to find a matrix best adapted to distinguishing distant evolutionary relationships from chance similarities. Recent mathematical results (Karlin & Altschul, 1990; Karlin *et al.*, 1990) allow all substitution matrices to be viewed in a common light, and provide a rationale for selecting particular sets of "optimal" scores for local protein sequence comparison.

2. The Statistical Significance of Local Sequence Alignments

Global alignments are of essentially no use unless they can allow gaps, but this is not true for local alignments. The ability to choose segments with arbitrary starting positions in each sequence means that biologically significant regions frequently may be aligned without the need to introduce gaps. While, in general, it is desirable to allow gaps in local alignments, doing so greatly decreases their mathematical tractability. The results described here apply rigorously only to local alignments that lack gaps, i.e. to segments of equal length from each of the two sequences compared. Some recent database search tools have focused on finding such alignments (Altschul & Lipman, 1990; Altschul *et al.*, 1990). However, the statistics of optimal scores for local alignments that include gaps (Smith *et al.*, 1985; Waterman *et al.*, 1987) are broadly analogous to those for the no-gap case (Karlin & Altschul,

far as possible would always tend to increase its score; this violates the idea of seeking local alignments. Substitution matrices used in other contexts, such as global alignments (Needleman & Wunsch, 1970) or local alignments using windows (McLachlan, 1971), need not satisfy these constraints. However, unless otherwise stated, it will be assumed below that any substitution matrix satisfies the two conditions described.

The statistical theory of MSP scores (Karlin & Altschul, 1990; Karlin *et al.*, 1990) involves a key parameter λ , which is the unique positive solution to the equation:

$$\sum_{i,j} p_i p_j e^{\lambda s_{ij}} = 1. \quad (1)$$

Notice that multiplying all the scores of a substitution matrix by some positive constant does not effect the relative scores of any subalignments. Two matrices related by such a factor can, therefore, be

paired with certain characteristic frequencies. Only if these correspond to a matrix's target frequencies, it has been argued, can the matrix be optimal for distinguishing distant local homologies from similarities due to chance (Karlin & Altschul, 1990).

Any substitution matrix has an implicit set of target frequencies for aligned amino acids. Writing the scores of the matrix in terms of its target frequencies, one has:

$$s_{ij} = \left(\ln \frac{q_{ij}}{p_i p_j} \right) / \lambda. \quad (3)$$

In other words, the score for an amino acid pair can

doubtful that any "target distribution" theorem can be proved. It may be possible to make a convincing case for a particular substitution matrix in the global alignment context, but the argument will most likely have to be different from that for local alignments (Karlin & Altschul, 1990). The same applies to substitution matrices used with fixed-length windows for studying local similarities (McLachlan, 1971; Argos, 1987; Stormo & Hartzell, 1989): a fixed quantity can be added to all entries of such a matrix with no essential effect. It is notable that while the PAM matrices were developed originally for global sequence comparison (Dayhoff *et al.*, 1978), their statistical theory has blossomed in the local alignment context.

Table 2
The average score (in bits) per alignment position when using given PAM matrices to compare segments in fact separated by a variety of PAM distances

PAM matrix <i>M</i> employed	40	80	Actual PAM distance <i>D</i> of segments						
			120	160	200	240	280	320	
40	2.26	1.31	0.62	0.10	-0.30	-0.61	-0.86	-1.06	
80	2.14	1.44	0.92	0.53	0.23	-0.02	-0.21	-0.37	
120	1.93	1.39	0.98	0.67	0.42	0.22	0.06	-0.07	
160	1.71	1.28	0.95	0.70	0.50	0.33	0.20	0.09	
200	1.51	1.16	0.90	0.68	0.51	0.38	0.26	0.17	
240	1.32	1.05	0.82	0.65	0.51	0.39	0.29	0.21	
280	1.17	0.94	0.75	0.60	0.48	0.38	0.30	0.23	
320	1.03	0.84	0.68	0.56	0.46	0.37	0.30	0.24	

segments have diverged by more than about 75 and 150 PAMs, respectively.

7. PAM Matrices for Database Searching and Two-sequence Comparison

The relative entropy associated with a specific PAM distance indicates how much information per position is optimally available. For a given alignment, one can attain such a score only by using the appropriate PAM matrix, but, of course, before the alignment is found it will not be known which matrix that is. It has therefore been proposed that a variety of PAM matrices be used for database searches (Collins *et al.*, 1988). We seek here to analyze how many such matrices are necessary, and which should be used.

Suppose one uses a matrix optimized for PAM distance *M* to compare two homologous protein segments that are actually separated by PAM distance *D*. For a range of values of *M* and *D*, the average score achieved per alignment position is shown in Table 2. Notice that for any given matrix *M*, the smaller the actual distance *D*, the higher the score. On the other hand, for a specific distance *D*, the highest score corresponds to the matrix with PAM distance *M* = *D*; this score is just the relative entropy discussed above. Using a PAM matrix with *M* near *D*, however, can yield a near-optimal score.

Table 3
Ranges of local alignment lengths for which various PAM matrices are appropriate

PAM matrix	93% efficiency range for database searching (30 bits)	87% efficiency range for 2-sequence comparison (16 bits)
40	9 to 21	4 to 14
80	13 to 34	6 to 22
120	19 to 50	9 to 33
160	26 to 70	12 to 46
200	36 to 94	16 to 62
240	47 to 123	21 to 80
280	60 to 155	27 to 101

For example, the relative entropy for *D* = 160 is 0.70 bit, but any PAM matrix in the range 120 to 200 yields at least 0.67 bit per position. In practice, how near the optimal is it important to be?

As argued above, for a given PAM distance there is a critical length at which alignments are just distinguishable from chance in a typical current database search; these lengths are recorded in Table 1. For the sake of analysis, we will assume that it is worth performing an extra search (using a different PAM matrix) only if it is able to increase the score for such a critical alignment by about two bits, corresponding to a factor of 4 in significance. Since a critical alignment has about 30 bits of information, we will therefore be satisfied using a PAM matrix that yields a score greater than 93% of the optimal achievable. Using data such as those shown in Table 2, one can calculate for which PAM distances *D* (and thus for which critical lengths) a given matrix *M* is appropriate; the results are recorded in Table 3. Our experience has shown that perhaps the most typical lengths for distant local alignments are those for which the PAM-120 matrix gives near-optimal scores, i.e. lengths 19 to 50 residues. Therefore, if one wishes to use a single standard matrix for database searches, the PAM-120 matrix (Table 4) is a reasonable choice. This matrix may, however, miss short but strong or long but weak similarities that contain sufficient information to be found. Accordingly, Table 3 shows that to complement the PAM-120 matrix, the PAM-40 and PAM-240 (or traditional PAM-250) matrices can be used. Additional matrices should improve the detection of distant similarities only marginally (i.e. raise their scores by at most 2 bits).

If, rather than searching a database with a query sequence, one wishes to compare two specific sequences for which one already has evidence of relatedness, the background noise is greatly decreased. As discussed above, for two proteins of typical length, about 16 bits are needed to distinguish a local alignment from chance. Accordingly, applying the same criteria as before, a matrix should be considered adequate for those PAM distances at which it yields an average score

Table 6

Table 6

Three MSPs representing distant relationships, from searches of the PIR protein sequence database (release 26-0) with human α_1 B-glycoprotein (PIR code OMHU1B)

PIR code	Optimal PAM-250 alignment	Optimal PAM-250 score (bits)	Optimal PAM-120 score (bits)
OMHU1B	1 AIFYETQPSLWAESESLLKPLANVTLCQA 30		
PL0030	1 ALFLDPPPNLWAEQSLLEPWANVTLSQS 30	32.3	45.0
OMHU1B	171 LSEPSATVTIEELAAPPPVLMHGESSQVLHPGNKVTLCVAPLS 216		
S00474	18 LRGQTATSQPSASPGEPSPPSIHPAQSELIVEAGDTLSLTCIDP 61	25.0	29.0
K3HUVH	15 LPDTTREIVMTQSPPTLSLSPGERVTLSCLASQS 48	22.0	28.5
Highest chance alignment score:		27.0	28.0
PIR code of sequence involved:		JQ0102	WGSMHH

OMHU1B, human α_1 B-glycoprotein; PL0030, pig Po2 F protein; S00474, kinase-related transforming protein (kit) precursor; K3HUVH, human Ig κ chain precursor V-III region (Vh); JQ0102, eggplant mosaic virus RNA replicase (Osorio-Keese *et al.*, 1989); WGSMHH, *Streptomyces hygroscopicus* B phosphotransferase (Zalacain *et al.*, 1986).

proteins may be identified easily using either the PAM-250 or the PAM-120 substitution matrix. However, several distant relationships present are harder to detect. In Table 7 are shown four optimal PAM-250 alignments, representing homologies to each of the two A30300 nucleotide-binding folds. None of these alignments has a PAM-250 score as great as the highest chance score of 31.3 bits. In contrast, when the PAM-120 matrix is used, the

alignments jump in score by 4 to almost 12 bits, giving all but one a score greater than the highest chance PAM-120 score of 33.0 bits. (The boundaries of the optimal alignments change slightly under the alternate scoring scheme.) No biologically significant similarity is distinguished by the PAM-250 matrix that is not found using the PAM-120. The relatively high chance scores found in this example are partly attributable to the length of the query

Table 7

Four MSPs representing distant relationships, from searches of the PIR protein sequence database (release 26-0) with cystic fibrosis transmembrane conductance regulator (PIR code A30300)

PIR code	Optimal PAM-250 alignment	Optimal PAM-250 score (bits)	Optimal PAM-120 score (bits)
A30300	438 TPVLKDINFKIERGQLLAVAGSTGAGKTSLLMMIMGELEPSEGKI 482		
S05328	18 VSKDINLEIQDGEFVVFVGPSCGKSTLLRMIALETVTSGDL 60	28.3	40.0
BVECOA	11 THNLKNINLVIPRDKLIVVTGLSGSGKSSL 40	24.7	35.0
A30300	1219 YTEGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLRLINTEGEI 1267		
QRECFH	19 FRVPGRTLLHPLSLTFPAGKVTGLIGHNGSGKSTLLKMLGR 59	29.3	35.0
QREBOT	31 DEDYTAVNDINFTLRAGETLGIYGESGSGKSOSRLRMGLLATNGRI 77	28.3	32.5

GPVF 49 SAGVVDSPKLGAAAEKVFGMVRDSAVQLRATGEVVLDGKDGSIHQ 94
S AHA V L L H
506134 61 ASQLRSSRQMAHATRVSSIMSEYIEELSDILPELLATLARTHDL 106

for protein databases of typical current size (about 1×10^7 residues), the most broadly sensitive substitution matrix should be a log-odds matrix with

Arratia, R., Gordon, L. & Waterman, M. S. (1986). An extreme value theory for sequence matching. *Ann. Stat.* **14**, 971-993.

within genes coding for proteins. *J. Mol. Evol.* **19**, 437-448.
Husain, I., Van Houten, B., Thomas, D. C. & Sorensen

