J. Mol. Biol. (1991) 219, 555–565

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_1B-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, 1990; Karlin *et al.*, 1990), where more precise results are available. Therefore, one may hope that many of the basic ideas presented below will generalize to local alignments that include gaps.

Formally, we assume that the aligned amino acids a_i and a_j are assigned the substitution score s_{ij}. Given two protein sequences, the pair of equal length segments that, when aligned, have the greatest aggregate score we call the Maximal Segment Pair (MSP†). An MSP may be of any length; its score is the MSP score.

Since any two protein sequences, related or unrelated, will have some MSP score, it is important to know how great a score one can expect to find simply by chance. To address this question one needs some model of chance. The simplest is to assume that in the two proteins compared, the amino acid a_i appears randomly with the probability p_i. These probabilities are chosen to reflect the observed frequencies of the amino acids in actual proteins. For simplicity of discussion we will assume both proteins share the same amino acid probability distribution; more generally, one can allow them to have different distributions. A random protein sequence is simply one constructed according to this model.

For the sake of the statistical theory, we need to make two crucial but reasonable assumptions about the substitution scores. The first is that there be at least one positive score and the second is that the expected score $\sum_{i,j} p_i p_j s_{ij}$ be negative. Because we permit the length of a segment pair to be adjusted to optimize its score, both these assumptions are necessary also from a practical perspective. If there were no positive scores, the MSP would always consist of a single pair of residues (or none at all, if this were permitted), and such an alignment is not of interest. If the expected score for two random residues were positive, extending a segment pair as 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:

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

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 considered essentially equivalent. Inspection of equation (1) reveals that multiplying all scores by a also has the effect of dividing λ by a. The parameter λ may, therefore, be viewed as a natural scale for any scoring system; its deeper meaning will be discussed below.

Given two random protein sequences as described above, how many distinct, or "locally optimal" (Sellers, 1984) MSPs with score at least S are expected to occur simply by chance? This number is well approximated by the formula:

$$KN e^{-\lambda S} \qquad (2)$$

where N is the product of the sequences' lengths, and K is an explicitly calculable parameter (Karlin & Altschul, 1990; Karlin *et al.*, 1990). When comparing a single random sequence with all the sequences in a database, setting N to the product of the query sequence length and the database length (in residues) yields an upper bound on the number of distinct MSPs with score at least S that the search is expected to yield.

3. Optimal Substitution Matrices for Local Sequence Alignment

Formula (2) allows us to tell when a segment pair has a significantly high score. However, it does not assist in choosing an appropriate substitution matrix in the first place. A second class of results, however, has direct bearing on this question. These state that among MSPs from the comparison of random sequences, the amino acids a_i and a_j are aligned with frequency approaching $q_{ij} = p_i p_j e^{\lambda s_{ij}}$ (Arratia *et al.*, 1988; Karlin & Altschul, 1990; Karlin *et al.*, 1990; Dembo & Karlin, 1991).

Given any substitution matrix and random protein model, one may easily calculate the set of target frequencies, q_{ij}, just described. Notice that by the definition of λ in equation (1), these target frequencies sum to 1. Now among alignments representing distant homologies, the amino acids ar

† Abbreviations used: MSP, Maximal Segment Pair; Ig, immunoglobulin.

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:

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

In other words, the score for an amino acid pair can be written as the logarithm to some base of that pair's target frequency divided by the background frequency with which the pair occurs. Such a ratio compares the probability of an event occurring under two alternative hypotheses and is called a likelihood or odds ratio. Scores that are the logarithm of odds ratios are called log-odds scores. Adding such scores can be thought of as multiplying the corresponding probabilities, which is appropriate for independent events, so that the total score remains a log-odds score.

Log-odds matrices have been advocated in a number of contexts, (Dayhoff *et al.*, 1978; Gribskov *et al.*, 1987; Stormo & Hartzell, 1989). The widely used PAM matrices (Dayhoff *et al.*, 1978), for instance, are explicitly of this form. Other substitution matrices, though based on a wide variety of rationales, are all log-odds matrices, but with implicit rather than explicit target frequencies. Therefore, while one may criticize the method described by Dayhoff *et al.* for estimating appropriate target frequencies (Wilbur, 1985), the most direct way to derive superior matrices appears to be through the refined estimation of amino acid pair target and background frequencies rather than through any fundamentally different approach.

Substitution Matrices for Global Alignments

While we have been considering substitution matrices in the context of local sequence comparison, they may be employed for global alignment as well (Needleman & Wunsch, 1970; Sellers, 1974; Schwartz & Dayhoff, 1978). There is a fundamental difference, however, between the use of such matrices in these two contexts. For global alignments, as previously, multiplying all scores by a fixed positive number has no effect on the relative scores of different alignments. But adding a fixed quantity a to the score for aligning any pair of residues (and $a/2$ to the score for aligning a residue with a null) likewise has no effect. Scoring systems that may be transformed into one another by means of these two rules are, for all practical purposes, equivalent. Unfortunately, the new transformation means that no unique log-odds interpretation of global substitution matrices is possible, and it is 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.

5. Local Alignment Scores as Measures of Information

Multiplying a substitution matrix by a constant changes λ but does not alter the matrix's implicit target frequencies. By appropriate scaling, one may therefore select the parameter λ at will. Writing the matrix in log-odds form, such scaling corresponds merely to using a different implicit base for the logarithm. One natural choice for λ is 1, so that all scores become natural logarithms. Perhaps more appealing is to choose $\lambda = \ln 2 \approx 0{\cdot}693$, so that the base for the log-odds matrix becomes 2. This lends a particularly intuitive appeal to formula (2). Setting the expected number of MSPs with score at least S equal to p, and solving for S, one finds:

$$S = \log_2 \frac{K}{p} + \log_2 N. \qquad (4)$$

For typical substitution matrices, K is found to be near 0·1, and an alignment may be considered significant when p is 0·05. Therefore the right-hand side of equation (4) generally is dominated by the term $\log_2 N$. In other words, the score needed to distinguish an MSP from chance is approximately the number of bits needed to specify where the MSP starts in each of the two sequences being compared. (One bit can be thought of as the answer to a single yes-no question; it is the amount of information needed to distinguish between 2 possibilities. It becomes apparent that, in general, $\log_2 N$ bits of information are needed to distinguish among N possibilities.)

For comparing two proteins of length 250 amino acid residues, about 16 bits of information are required; for comparing one such protein to a sequence database containing 4,000,000 residues, about 30 bits are needed. When cast in this light, alignment scores are not arbitrary numbers. By appropriate scaling (multiplying by $\lambda/0{\cdot}693$) they take on the units of bits, and rough significance calculations can be performed in one's head. Furthermore, when so normalized, different amino acid substitution matrices may be directly compared.

6. The Relative Entropy of a Substitution Matrix

The above review of previous results has provided us with the necessary tools for the analysis that follows. The ultimate goal is to decide which substitution matrices are the most appropriate for database searching and for detailed pairwise sequence comparison.

Given a random protein model and a substitution matrix, one may calculate the target frequencies q_{ij} characteristic of the alignments for which the matrix is optimized. A useful quantity to consider is the average score (information) per residue pair in these alignments. Assuming the substitution matrix is normalized as described above, this value is simply:

$$H = \sum_{i,j} q_{ij} s_{ij} = \sum_{i,j} q_{ij} \log_2 \frac{q_{ij}}{p_i p_j}. \quad (5)$$

Notice that H depends both on the substitution matrix and on the random protein model. In information theoretic terms, H is the relative entropy of the target and background distributions. The origin of the name need not be of concern. The important point is that, for an alignment characterized by the target frequencies q_{ij}, H measures the average information available per position to distinguish the alignment from chance. Intuitively, the higher the value of the relative entropy of target and background distributions, the more easily they are distinguished. For a high value of H, relatively short alignments with the target distribution can be distinguished from chance, while, if the value of H is lower, longer alignments are necessary.

It is interesting to examine the PAM model of molecular evolution (Dayhoff *et al.*, 1978) from this standpoint. From a study of mutations between a large number of closely related proteins, Dayhoff and co-workers proposed a stochastic model of protein evolution. The amount of evolutionary change that yields, on average, one substitution in 100 amino acid residues they called one PAM. Using their model, one may easily calculate the frequency with which any two amino acid residues are paired in an accurate alignment of two homologous proteins that have diverged by any given amount of evolutionary change. These target frequencies may then be used to construct log-odds matrices and, in particular, the widely used PAM-250 matrix. Dayhoff *et al.* (1978) originally proposed this matrix for the global alignment of two sequences suspected to be homologous, but it has since been used to search protein databases for local alignments to a query sequence (Lipman & Pearson, 1985; Pearson & Lipman, 1988). One may therefore inquire whether 250 PAMs yield reasonable target frequencies for database searches.

Assuming the model described by Dayhoff *et al.* (1978), Table 1 lists the relative entropy H implicit in a range of PAM matrices. As argued above, distinguishing an alignment from chance in a search of a typical current protein database using an average length protein requires about 30 bits of information. Accordingly, for an alignment of segments separated by a given PAM distance, one can calculate the minimum length necessary to rise above background noise; these lengths are recorded in Table 1. For instance, at a distance of 250 PAMs, on average only 0·36 bit of information is available per alignment position. To be statistically significant, such an alignment would need to have a length greater than about 83 residues. Many biologically interesting regions of protein similarity are much shorter than this, and accordingly need a stronger signal to be detected. A local alignment of length 20 residues will need about 1·5 bits per alignment position, while one of length 40 residues will need about 0·75 bit. Table 1 shows that such alignments will not be detectable if their constituent

Table 1
The relative entropy H *of PAM matrices*

PAM distance	*H* (bits)	Min. significant length (30 bits)	PAM distance	*H* (bits)	Min. significant length (30 bits)
0	4·17	8	180	0·60	51
10	3·43	9	190	0·55	55
20	2·95	11	200	0·51	59
30	2·57	12	210	0·48	63
40	2·26	14	220	0·45	68
50	2·00	15	230	0·42	73
60	1·79	17	240	0·39	78
70	1·60	19	250	0·36	83
80	1·44	21	260	0·34	89
90	1·30	24	270	0·32	94
100	1·18	26	280	0·30	100
110	1·08	28	290	0·28	107
120	0·98	31	300	0·27	113
130	0·90	34	310	0·25	120
140	0·82	37	320	0·24	127
150	0·76	40	330	0·22	134
160	0·70	43	340	0·21	141
170	0·65	47	350	0·20	149

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 M employed	Actual PAM distance D of segments							
	40	80	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
320	75 to 192	34 to 124
360	94 to 233	42 to 149

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 within 87% of the optimal. In Table 3, we list the range of critical lengths over which various PAM

Table 4
The PAM-120 matrix with scores in half bits

A	3																			
R	−3	6																		
N	0	−1	4																	
D	0	−3	2	5																
C	−3	−4	−5	−7	9															
Q	−1	1	0	1	−7	6														
E	0	−3	1	3	−7	2	5													
G	1	−4	0	0	−5	−3	−1	5												
H	−3	1	2	0	−4	3	−1	−4	7											
I	−1	−2	−2	−3	−3	−3	−3	−4	−4	6										
L	−3	−4	−4	−5	−7	−2	−4	−5	−3	1	5									
K	−2	2	1	−1	−7	0	−1	−3	−2	−2	−4	5								
M	−2	−1	−3	−4	−6	−1	−4	−4	−4	1	3	0	8							
F	−4	−4	−4	−7	−6	−6	−6	−5	−2	0	0	−6	−1	8						
P	1	−1	−2	−2	−3	0	−1	−2	−1	−3	−3	−2	−3	−5	6					
S	1	−1	1	0	−1	−2	−1	1	−2	−2	−4	−1	−2	−3	1	3				
T	1	−2	0	−1	−3	−2	−2	−1	−3	0	−3	−1	−1	−4	−1	2	4			
W	−7	1	−5	−8	−8	−6	−8	−8	−5	−7	−5	−5	−7	−1	−7	−2	−6	12		
Y	−4	−6	−2	−5	−1	−5	−4	−6	−1	−2	−3	−6	−4	4	−6	−3	−3	−1	8	
V	0	−3	−3	−3	−2	−3	−3	−2	−3	3	1	−4	1	−3	−2	−2	0	−8	−3	5
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

matrices are appropriate for detailed pairwise sequence comparison. As a single matrix, the PAM-200 spans the most typical range of local alignment lengths, i.e. 16 to 62 residues. Alternatively, if two different matrices are to be used, the PAM-80 and PAM-250, which together span alignment lengths 6 to 85 residues, or the PAM-120 and PAM-320 matrices, which span lengths 9 to 124 residues, appear to be appropriate pairs.

Since it is convenient to express substitution matrices as integers, and since a probability factor of 2 between score levels is too rough, the units for the PAM-120 matrix shown in Table 4 are half bits. The scores in the original PAM-250 matrix (Dayhoff *et al.*, 1978) were scaled as $10 \times \log_{10}$. Because $10/(\ln 10) \approx 3/(\ln 2)$ to within 0·4%, a unit score in that matrix can be thought of as approximately one-third of a bit.

8. Biological Examples

As discussed, the particular PAM matrix that best distinguishes distant homologies from chance similarities found in a database search depends on the nature of the homologies present, and this cannot be known *a priori*. However, it is frequently the case that distantly related proteins will share isolated stretches of relatively conserved amino acid residues, corresponding to active sites or other important structural features. It has been observed that in general the mutations along genes coding for proteins are not Poisson-distributed (Uzzell & Corbin, 1971; Holmquist *et al.*, 1983), suggesting that short, conserved regions are to be expected. As shown in Table 3, this means that the widely used PAM-250 matrix generally will not be optimal for locating distant relationships.

In the examples below, we compare the PAM-250 and PAM-120 scores for MSPs representing distant relationships to four different query sequences. In all cases, we consider relationships near the limits of what can be distinguished from chance in a search of the PIR protein sequence database (Release 26·0; 7,348,950 residues). It will be noticed that the highest chance PAM-250 scores are consistently slightly smaller than the highest chance PAM-120 scores. This is primarily attributable to the fact that the parameter K discussed above is about half as large for the former scores as for the latter. Furthermore, since neither the PIR database nor a given query sequence ever precisely fits the random protein model described by Dayhoff *et al.* (1978), the parameter λ varies slightly from one comparison to another. Therefore, while we will treat the PAM-120 scores from Table 4 as half bits, and the PAM-250 scores of Dayhoff *et al.* (1978) as one-third bits, it should be noted that this is always a slight approximation.

(a) *Lipocalins*

We used the BLAST program (Altschul *et al.*, 1990) to search the PIR database with human apolipoprotein D precursor (PIR code LPHUD; Drayna *et al.*, 1987), using both the PAM-250 (Dayhoff *et al.*, 1978) and PAM-120 (Table 4) substitution matrices. Human apolipoprotein D precursor is a 189 residue glycoprotein that belongs to the lipocalin (α_2-microglobulin) superfamily, which contains proteins that exhibit a wide range of functions related to their ability to bind small hydrophobic ligands. The similarities among these proteins and their biological roles have been analyzed (Peitsch & Boguski, 1990), and crystal structures are available for several members of the superfamily (Cowan *et al.*, 1990). Three proteins in the superfamily are rat androgen-dependent epididymal protein (PIR code

Table 5

Three MSPs representing distant relationships, from searches of the PIR protein sequence database (release 26·0) with human apolipoprotein D precursor (PIR code LPHUD)

PIR code	Optimal PAM-250 alignment			Optimal PAM-250 score (bits)	Optimal PAM-120 score (bits)
LPHUD	25	LGKCPNPPVQENFDVNKYLGRWYEI	49		
SQRTAD	12	LAAGTEGAVVKDFDISKFLGFWYEI	36	27.0	33.5
A32202	27	HDTVQPNFQQDKFLGRWY	44	25.7	33.5
HCHU	28	NIQVQENFNISRIYGKWYNL	47	23.0	30.5
Highest chance alignment score:				27.0	29.0
PIR code of sequence involved:				S00758	S00758

LPHUD, human apolipoprotein D precursor; SQRTAD, rat androgen-dependent epididymal 18·5 K protein precursor; A32202, rat prostaglandin-D synthase; HCHU, human α_1-microglobulin/inter-α-trypsin inhibitor precursor; S00758, human surface glycoprotein CD16 precursor.

SQRTAD; Brooks *et al.*, 1986), rat prostaglandin-D synthase (PIR code A32202; Urade *et al.*, 1989) and human α_1-microglobulin (PIR code HCHU; Kaumeyer *et al.*, 1986). The second of these has only recently been recognized as a member of the superfamily (M. S. Boguski & M. C. Peitsch, personal communication); it is the first such member with known catalytic activity (Urade *et al.*, 1989).

Using PAM-250 scores, the maximal segment pair for each of these sequences when compared to LPHUD is shown in Table 5. These local similarities correspond to one of two motifs that are conserved throughout the superfamily (Boguski & States, 1990). The scores for the three alignments are 27·0, 25·7 and 23·0 bits, respectively. However, the highest score from a protein in the database unrelated to LPHUD is 27·0 bits, involving human surface glycoprotein CD16 precursor (PIR code S00758; Simmons & Seed, 1988). The PAM-250 matrix therefore fails to separate the homologous alignments shown from background noise. In contrast, using the PAM-120 matrix of Table 4, the scores for the three alignments jump to 33·5, 33·5 and 30·5 bits, respectively. (The 1st 7 alignment positions for LPHUD-SQRTAD shown in Table 5 are dropped in an optimal PAM-120 alignment, as are the 1st 3 positions for the LPHUD-A32202 alignment.) This raises their scores above that of the best chance PAM-120 alignment (29·0 bits), again involving human surface glycoprotein CD16 precursor. Notice that in both cases the estimate that about 30 bits are needed clearly to distinguish an MSP from chance is valid. For this query sequence, no relationship is found using the PAM-250 matrix that is missed by the PAM-120.

(b) *Human α_1B-glycoprotein*

We searched the PIR database with human α_1B-glycoprotein (PIR code OMHU1B; Ishioka *et al.*, 1986), a plasma glycoprotein of unknown function, and a member of the immunoglobulin superfamily. Using the PAM-250 matrix, the only protein in the database with an MSP that rises above background noise is pig Po2 F protein (PIR code PL0030; Van de Weghe *et al.*, 1988), which achieves a score of 32·3 bits. As shown in Table 6, the score for this known homology (Van de Weghe *et al.*, 1988) rises to 45·0 bits when the PAM-120 matrix is used instead. In addition, two proteins with immunoglobulin domains, kinase-related transforming protein precursor (PIR code S00474; Qiu *et al.*, 1988) and human Ig κ chain precursor V-III region (PIR code K3HUVH; Pech & Zachau, 1984), achieve scores of 29·0 and 28·5 bits, respectively. Table 6 illustrates that both these similarities are only just distinguishable from chance, and that using the PAM-250 matrix both similarities drop in score by at least four bits.

(c) *The cystic fibrosis transmembrane conductance regulator*

The cause of cystic fibrosis has been traced to mutations in a protein that bears striking similarity to many proteins involved in the transport of substances across the cell membrane (PIR code A30300; Riordan *et al.*, 1989). Characteristic features of the protein are two nucleotide (ATP)-binding folds (Higgins *et al.*, 1986). When the PIR database is searched with A30300, many related

Table 6

Three MSPs representing distant relationships, from searches of the PIR protein sequence database (release 26·0) with human α_1B-glycoprotein (PIR code OMHU1B)

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

OMHU1B, human α_1B-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 hygromycin* 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 VSKDINLEIQDGEFVVFVGPSGCGKSTLLRMIAGLETVTSGDL 60	28.3	40.0
BVECUA	11 THNLKNINLVIPRDKLIVVTGLSGSGKSSL 40	24.7	35.0
A30300	1219 YTEGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLRLLNTEGEI 1267		
QRECFH	19 FRVPGRTLLHPLSLTFPAGKVTGLIGHNGSGKSTLLKMLGR 59	29.3	35.0
QREBOT	31 DGDVTAVNDLNFTLRAGETLGIVGESGSGKSQSRLRLMGLLATNGRI 77	28.3	32.5
	Highest chance alignment score:	31.3	33.0
	PIR code of sequence involved:	A34416	A32916

A30300, Cystic fibrosis transmembrane conductance regulator; S05328, *Enterobacter aerogenes* inner membrane protein malK (Dahl *et al.*, 1989); BVECUA, *Escherichia coli* uvrA protein (Husain *et al.*, 1986); QRECFH, ferrichrome-iron transport protein fhuC (Coulton *et al.*, 1987); QREBOT, oligopeptide permease membrane protein oppD (Higgins *et al.*, 1985); A34416, fluke hydroxymethylglutaryl-CoA reductase (NADPH) (Rajkovic *et al.*, 1989); A32916, *Vibrio harveyi* acyl-protein synthetase (fragment) (Johnston *et al.*, 1989).

```
GPVF      49 SAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVVLDGKDGSIHIQ  94
                S   AHA V       L     L     H
S06134    61 ASQLRSSRQMQAHAIRVSSIMSEYIEELDSDILPELLATLARTHDL 106

GPVF      95 KGVLDPHFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
               V   H       L       G       W A
S06134   107 NKVGPAHYDLFAKVLMEALQAELGSDFNQKTRDSWAKAFSIVQAVL 152
```

Figure 1. The PAM-250 maximal segment pair of broad bean leghemoglobin I (PIR code GPVF) and sea cucumber hemoglobin I (PIR code S06134). Identical residues are echoed on the central line. PAM-250 score. 25·3 bits: length. 92 residues.

sequence (1480 residues), and partly to its composition which renders the parameter λ slightly smaller than in the previous examples.

(d) *Globins*

It is possible to find examples of long alignments representing distant relationships that are better distinguished by the PAM-250 than by the PAM-120 matrix. In practice such examples are rare for some of the reasons discussed above. The globins are one superfamily in which sequence divergence has been relatively uniform over the length of entire proteins. As a result, some sequence relationships within this superfamily become apparent only with scoring systems tailored for long but very weak alignments.

For example, searching the PIR database with broad bean leghemoglobin I (PIR code GPVF; Richardson *et al.*, 1975), the alignment with sea cucumber hemoglobin I (PIR code S06134; Suzuki, 1989), shown in Figure 1, is found having a PAM-250 score of 25·3 bits. This is almost as high as the score of the best chance MSP (26·7 bits), which involves *Salmonella typhimurium* cystathionine β-lyase (PIR code JV0020; Park & Stauffer, 1989). The alignment is 92 residue pairs long; only 14 of these pairs involve identical amino acid residues, and they are spread fairly evenly along the alignment. This particular similarity is totally obscured when PAM-120 scores are used. The best region of the alignment shown then involves residues 100 to 133 of the leghemoglobin sequence and has a score of only 13 bits, while the best chance PAM-120 alignment, involving mouse hepatitis virus E1 membrane glycoprotein (PIR code VGIHE1; Armstrong *et al.*, 1984), scores 27·5 bits. Nevertheless, as in the previous examples, a number of relationships are distinguished by the PAM-120 matrix but missed by the PAM-250.

9. Conclusion

This paper has analyzed the properties of amino acid substitution matrices in the context of local alignments lacking gaps. This is exactly the sort of alignment sought by the recently developed BLAST database search programs (Altschul *et al.*, 1990; Altschul & Lipman, 1990). We have concluded that 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 relative entropy of about one bit, e.g. the PAM-120 matrix. In order to detect short but strong homologies or long but weak ones, this matrix can be complemented by the PAM-40 and PAM-250 matrices; additional matrices should be of only marginal utility. Of course, many database search methods, such as the FASTA programs (Lipman & Pearson, 1985; Pearson & Lipman, 1988), seek local alignments with gaps, and such measures are potentially more sensitive to distant homologies. Unfortunately, if gaps with associated scores are allowed, the specific quantitative discussion above is no longer correct. Nevertheless, the general thrust of the arguments should still apply, and theory and experiment suggest that analogous results will hold for local alignments with gaps (Smith *et al.*, 1985; Waterman *et al.*, 1987; Collins *et al.*, 1988).

There are, of course, many much more involved ways for assessing local alignment than those discussed here. Scores can be assigned to aligned di-residues or tri-residues; they can depend on alignment length (Altschul & Erikson, 1986); or they can be complex combinations of various scoring methods (Argos, 1987). Protein databases may also be searched with position-dependent scores or "profiles" constructed from multiple alignments (Taylor, 1986; Gribskov *et al.*, 1987; Patthy, 1987). In certain contexts such systems may well be more sensitive than the straightforward local scoring system considered here. Two advantages of simple additive scores are their amenability to powerful algorithmic methods (Altschul *et al.*, 1990) and to rigorous statistical analysis (Karlin & Altschul, 1990; Karlin *et al.*, 1990). Such analysis may also yield insight into the properties of more complicated scoring schemes.

The author thanks Drs David Lipman, Mark Boguski and Andrew McLachlan for helpful conversations and suggestions on the manuscript.

References

Altschul, S. F. & Erickson, B. W. (1986). A nonlinear measure of subalignment similarity and its significance levels. *Bull. Math. Biol.* **48**. 617–632.

Altschul, S. F. & Lipman, D. J. (1990). Protein database searches for multiple alignments. *Proc. Nat. Acad. Sci., U.S.A.* **87**. 5509–5513.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**. 403–410.

Argos, P. (1987). A sensitive procedure to compare amino acid sequences. *J. Mol. Biol.* **193**. 385–396.

Armstrong, J., Niemann, H., Smeekens, S., Rottier, P. & Warren, G. (1984). Sequence and topology of a model intracellular membrane protein, E1 glycoprotein, from a coronavirus. *Nature (London)*. **308**. 751–752.

Arratia, R. & Waterman, M. S. (1989). The Erdos–Renyi strong law for pattern matching with a given proportion of mismatches. *Ann. Prob.* **17**. 1152–1169.

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

Arratia, R., Morris, P. & Waterman, M. S. (1988). Stochastic scrabble: large deviations for sequences with scores. *J. Appl. Prob.* **25**, 106–119.

Boguski, M. S. & States, D. J. (1990). Molecular sequence databases and their uses. In *Protein Engineering: A Practical Approach* (Rees, A. R., Wetzel, R. & Sternberg, M. J. E., eds), chap. 5. IRL Press, Oxford.

Brooks, D. E., Means, A. R., Wright, E. J., Singh, S. P. & Tiver, K. K. (1986). Molecular cloning of the cDNA for two major androgen-dependent secretory proteins of 18·5 kilodaltons synthesized by the rat epididymis. *J. Biol. Chem.* **261**, 4956–4961.

Collins, J. F., Coulson, A. F. W. & Lyall, A. (1988). The significance of protein sequence similarities. *Comput. Appl. Biosci.* **4**, 67–71.

Coulton, J. W., Mason, P. & Allatt, D. D. (1987). *fhuC* and *fhuD* genes for iron(III)-ferrichrome transport into *Escherichia coli* K-12. *J. Bacteriol.* **169**, 3844–3849.

Cowan, S. W., Newcomer, M. E. & Jones, T. A. (1990). Crystallographic refinement of human serum retinol binding protein at 2 Å resolution. *Proteins*, **8**, 44–61.

Dahl, M. K., Francoz, E., Saurin, W., Boos, W., Manson, M. D. & Hofnung, M. (1989). Comparison of sequences from the malB regions of *Salmonella typhimurium* and *Enterobacter aerogenes* with *Escherichia coli* K12: a potential new regulatory site in the interoperonic region. *Mol. Gen. Genet.* **218**, 199–207.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, pp. 345–352. Nat. Biomed. Res. Found., Washington, DC.

Dembo, A. & Karlin, S. (1991). Strong limit laws of empirical functionals for large exceedences of partial sums of I.I.D. variables. *Ann. Prob.* In the press.

Drayna, D. T., McLean, J. W., Wion, K. L., Trent, J. M., Drabkin, H. A. & Lawn, R. M. (1987). Human apolipoprotein D gene: gene sequence, chromosome localization, and homology to the $\alpha_{2\mu}$-globulin superfamily. *DNA*, **6**, 199–204.

Feng, D. F., Johnson, M. S. & Doolittle, R. F. (1985). Aligning amino acid sequences: comparison of commonly used methods. *J. Mol. Evol.* **21**, 112–125.

Goad, W. B. & Kanehisa, M. I. (1982). Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. *Nucl. Acids Res.* **10**, 247–263.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 4355–4358.

Higgins, C. F., Hiles, I. D., Whalley, K. & Jamieson, D. J. (1985). Nucleotide binding by membrane components of bacterial periplasmic binding protein-dependent transport systems. *EMBO J.* **4**, 1033–1039.

Higgins, C. F., Hiles, I. D., Salmond, G. P., Gill, D. R., Downie, J. A., Evans, I. J., Holland, I. B., Gray, L., Buckel, S. D., Bell, A. W. & Hermodson, M. A. (1986). A family of related ATP-binding subunits coupled to many distinct biological processes in bacteria. *Nature (London)*, **323**, 448–450.

Holmquist, R., Goodman, M., Conroy, T. & Czelusniak, J. (1983). The spatial distribution of fixed mutations within genes coding for proteins. *J. Mol. Evol.* **19**, 437–448.

Husain, I., Van Houten, B., Thomas, D. C. & Sancar, A. (1986). Sequences of *Escherichia coli uvrA* gene and protein reveal two potential ATP binding sites. *J. Biol. Chem.* **261**, 4895–4901.

Ishioka, N., Takahashi, N. & Putnam, F. W. (1986). Amino acid sequence of human plasma α 1B-glycoprotein: homology to the immunoglobulin supergene family. *Proc. Nat. Acad. Sci., U.S.A.* **83**, 2363–2367.

Johnston, T. C., Hruska, K. S. & Adams, L. F. (1989). The nucleotide sequence of the *luxE* gene of *Vibrio harveyi* and a comparison of the amino acid sequences of the acyl-protein synthetases from *V. harveyi* and *V. fischeri*. *Biochem. Biophys. Res. Commun.* **163**, 93–101.

Karlin, S. & Altschul, S. F. (1990). 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.

Karlin, S., Dembo, A. & Kawabata, T. (1990). Statistical composition of high-scoring segments from molecular sequences. *Ann. Stat.* **18**, 571–581.

Kaumeyer, J. F., Polazzi, J. O. & Kotick, M. P. (1986). The mRNA for a proteinase inhibitor related to the HI-30 domain of inter-α-trypsin inhibitor also encodes α-1-microglobulin (protein HC). *Nucl. Acids Res.* **14**, 7839–7850.

Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–1441.

McLachlan, A. D. (1971). Tests for comparing related amino acid sequences. Cytochrome c and cytochrome c_5. *J. Mol. Biol.* **61**, 409–424.

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* **48**, 443–453.

Osorio-Keese, M. E., Keese, P. & Gibbs, A. (1989). Nucleotide sequence of the genome of eggplant mosaic tymovirus. *Virology*, **172**, 547–554.

Park, Y. M. & Stauffer, G. V. (1989). DNA sequence of the *metC* gene and its flanking regions from *Salmonella typhimurium* LT2 and homology with the corresponding sequence of *Escherichia coli*. *Mol. Gen. Genet.* **216**, 164–169.

Patthy, L. (1987) Detecting homology of distantly related proteins with consensus sequences. *J. Mol. Biol.* **198**, 567–577.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Nat. Acad. Sci., U.S.A.* **85**, 2444–2448.

Pech, M. & Zachau, H. G. (1984). Immunoglobulin genes of different subgroups are interdigitated within the VK locus. *Nucl. Acids Res.* **12**, 9229–9236.

Peitsch, M. C. & Boguski, M. S. (1990). Is apolipoprotein D a mammalian bilin-binding protein? *New Biologist*, **2**, 197–206.

Qiu, F., Ray, P., Brown, K., Barker, P. E., Jhanwar, S., Ruddle, F. H. & Besmer, P. (1988). Primary structure of c-kit: relationship with the CSF-1/PDGF receptor kinase family–oncogenic activation of v-kit involves deletion of extracellular domain and C terminus. *EMBO J.* **7**, 1003–1011.

Rajkovic, A., Simonsen, J. N., Davis, R. E. & Rottman, F. M. (1989). Molecular cloning and sequence analysis of 3-hydroxy-3-methylglutaryl-coenzyme reductase from the human parasite *Schistosoma*

mansoni. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 8217–8221.

Rao, J. K. M. (1987). New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. *Int. J. Pept. Protein Res.* **29**, 276–281.

Richardson, M., Dilworth, M. J. & Scawen, M. D. (1975). The amino acid sequence of leghaemoglobin I from root nodules of broad bean (*Vicia faba* L.). *FEBS Letters*, **51**, 33–37.

Riordan, J. R., Rommens, J. M., Kerem, B. S., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J. L., Drumm, M. L., Iannuzzi, M. C., Collins, F. S. & Tsui, L. C. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, **245**, 1066–1073.

Risler, J. L., Delorme, M. O., Delacroix, H. & Henaut, A. (1988). Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *J. Mol. Biol.* **204**, 1019–1029.

Sankoff, D. & Kruskal, J. B. (1983). *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, MA.

Schwartz, R. M. & Dayhoff, M. O. (1978). Matrices for detecting distant relationships. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, pp. 353–358. Nat. Biomed. Res. Found., Washington, DC.

Sellers, P. H. (1974). On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.* **26**, 787–793.

Sellers, P. H. (1984). Pattern recognition in genetic sequences by mismatch density. *Bull. Math. Biol.* **46**, 501–514.

Simmons, D. & Seed, B. (1988). The Fcγ receptor of natural killer cells is a phospholipid-linked membrane protein. *Nature (London)*, **333**, 568–570.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.

Smith, T. F., Waterman, M. S. & Burks, C. (1985). The statistical distribution of nucleic acid similarities. *Nucl. Acids Res.* **13**, 645–656.

Stormo, G. D. & Hartzell, G. W., III (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 1183–1187.

Suzuki, T. (1989). Amino acid sequence of a major globin from the sea cucumber *Paracaudina chilensis*. *Biochim. Biophys. Acta*, **998**, 292–296.

Taylor, W. R. (1986). Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* **188**, 233–258.

Urade, Y., Nagata, A., Suzuki, Y. & Hayaishi, O. (1989). Primary structure of rat brain prostaglandin D synthetase deduced from cDNA sequence. *J. Biol. Chem.* **264**, 1041–1045.

Uzzell, T. & Corbin, K. W. (1971). Fitting discrete probability distributions to evolutionary events. *Science*, **172**, 1089–1096.

Van de Weghe, A., Coppieters, W., Bauw, G., Vanderkerckhove, J. & Bouquet, Y. (1988). The homology between the serum proteins PO2 in pig, Xk in horse and α_1B-glycoprotein in human. *Comp. Biochem. Physiol.* **90B**, 751–756.

Waterman, M. S., Gordon, L. & Arratia, R. (1987). Phase transitions in sequence matches and nucleic acid structure. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 1239–1243.

Wilbur, W. J. (1985). On the PAM matrix model of protein evolution. *Mol. Biol. Evol.* **2**, 434–447.

Zalacain, M., Gonzalez, A., Guerrero, M. C., Mattaliano, R. J., Malpartida, F. & Jimenez, A. (1986). Nucleotide sequence of the hygromycin B phosphotransferase gene from *Streptomyces hygroscopicus*. *Nucl. Acids Res.* **14**, 1565–1581.

Edited by F. E. Cohen