J Mol Evol (1993) 36:290–300

Journal of Molecular Evolution

A Protein Alignment Scoring System Sensitive at All Evolutionary Distances

Stephen F. Altschul

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Summary. Protein sequence alignments generally are constructed with the aid of a "substitution matrix" that specifies a score for aligning each pair of amino acids. Assuming a simple random protein model, it can be shown that any such matrix, when used for evaluating variable-length local alignments, is implicitly a "log-odds" matrix, with a specific probability distribution for amino acid pairs to which it is uniquely tailored. Given a model of protein evolution from which such distributions may be derived, a substitution matrix adapted to detecting relationships at any chosen evolutionary distance can be constructed. Because in a database search it generally is not known a priori what evolutionary distances will characterize the similarities found, it is necessary to employ an appropriate range of matrices in order not to overlook potential homologies. This paper formalizes this concept by defining a scoring system that is sensitive at all detectable evolutionary distances. The statistical behavior of this scoring system is analyzed, and it is shown that for a typical protein database search, estimating the originally unknown evolutionary distance appropriate to each alignment costs slightly over two bits of information, or somewhat less than a factor of five in statistical significance. A much greater cost may be incurred, however, if only a single substitution matrix, corresponding to the wrong evolutionary distance, is employed.

Key words: Homology — Sequence comparison — Statistical significance — Alignment algorithms — Pattern recognition

Sequence alignment has become a widely used tool for the discovery and understanding of functional and evolutionary relationships among proteins. The earliest work in this area concentrated on "global" comparisons, aligning, for example, complete globin chains (Needleman and Wunsch 1970; Sellers 1974; Sankoff and Kruskal 1983). More recently the focus has shifted to "local" comparisons in which only small regions of the sequences compared need participate (Smith and Waterman 1981; Goad and Kanehisa 1982; Sellers 1984). Because proteins of related function often share only such isolated regions of similarity, e.g., near an active site, the most popular database search programs have all relied upon measures of local similarity (Lipman and Pearson 1985; Pearson and Lipman 1988; Altschul et al. 1990).

The local similarity of two sequences may be evaluated in a variety of manners. The simplest of these is the length of the longest matching segment, perhaps allowing for a specified number or proportion of mismatches (Arratia et al. 1986; Karlin and Ost 1988; Arratia and Waterman 1989). More sensitive measures take account of relationships among the amino acids, typically by assigning a score to every aligned pair of residues. Much thought has been given to choosing such a "substitution matrix" best able to distinguish biologically meaningful from chance similarities (McLachlan 1971; Dayhoff et al. 1978; Schwartz and Dayhoff 1978; Feng et al. 1985; Rao 1987; Risler et al. 1988; Altschul 1991; States et al. 1991; Gonnet et al. 1992; Henikoff and Henikoff 1992; Jones et al. 1992). Among such scoring systems, the most widely used and

most closely studied have been the log-odds matrices based upon the "PAM" model of protein evolution (Dayhoff et al. 1978; Schwartz and Dayhoff 1978). Recently it has been argued that *any* substitution matrix used for assessing local alignments is implicitly a log-odds matrix, best adapted for distinguishing local alignments characterized by specific frequencies for aligned amino acid pairs (Karlin and Altschul 1990; Altschul 1991). Accepting for the sake of analysis the PAM evolutionary model, it remains the case that when seeking protein similarities, the degree to which two segments have diverged will not be known a priori. Therefore, unless PAM matrices suitable for a range of evolutionary distances are used, one may easily miss biologically meaningful alignments that could have been distinguished from random noise (Altschul 1991).

While the importance of using scores tailored for different evolutionary distances has been recognized frequently (Schwartz and Dayhoff 1978; Collins et al. 1988; Altschul 1991; States et al. 1991; Gonnet 1993), the statistical issues presented by such multiple tests have not been well studied. The basic problem can be understood by imagining using a large number of different PAM matrices to compare the same two sequences. If only a single such matrix were used, the statistical significance of the highest-scoring segment pair may be easily bounded (Karlin and Altschul 1990). But selecting as well the PAM matrix that optimizes the alignment score introduces a new degree of freedom. An alignment that would have been statistically significant had only a single score matrix been used may no longer be surprising.

In this paper we propose a formal definition for an "All-PAM" scoring system that makes minimal a priori assumptions concerning the evolutionary distance between the sequences in the alignments sought. We study the distribution of optimal values of this scoring system when applied to random sequences so that the significance of high-scoring alignments can be properly assessed. Finally, we describe how the BLAST algorithms (Altschul et al. 1990) can be modified to accommodate an approximation of All-PAM scores, and provide examples where such scores are useful for distinguishing biologically meaningful similarities.

Maximal Segment Pairs and Their Statistical Significance

The measure of local sequence similarity we study is based upon a *substitution matrix* which assigns a score to every pair of amino acids. Given two protein sequences, we define their *maximal segment pair* (MSP) as those two equal-length segments which when aligned have maximal aggregate score. These two segments may be of any length that maximizes their score. The score of the maximal segment pair we take as our measure of local similarity.

In order to assess how great a score can be expected to occur by chance, we need a model of chance. We will employ a simple model in which the residues of a protein are chosen independently, with amino acid i occurring with probability p_i. A sequence fitting this model will be called a *random protein sequence*.

It has been argued that, given this random model, any substitution matrix useful for defining MSP scores is implicitly a log-odds matrix (Karlin and Altschul 1990; Karlin et al. 1990; Altschul 1991). In other words, if amino acid i aligned with amino acid j has score s_{ij}, then there are a specific set of *target frequencies* q_{ij} for which the scores can be written

$$s_{ij} = \log \frac{q_{ij}}{p_i p_j} \tag{1}$$

These scores are the ones best suited for distinguishing local alignments in which amino acids i and j are aligned with frequency q_{ij} (Karlin and Altschul 1990; Karlin et al. 1990; Altschul 1991; Dembo and Karlin 1991).

A set of scores is changed in no essential way if all are multiplied by a positive constant: clearly the optimal alignment will not be changed. Such multiplication affects only the base of the logarithm in equation (1), not the target frequencies. In the language of information theory, scores that are expressed as natural logarithms are said to have the units of *nats* (Hamming 1986). Because it is easier to do mental calculations using powers of 2, it is frequently more convenient to use logarithms to the base 2, and scores based on such logarithms are said to be expressed in *bits*. The difference between these units, like that between feet and meters, is simply a constant factor: 1 bit $\approx$ 0.693 nat. Because the formulation of the mathematical concepts in this paper is simpler when natural logarithms are employed, it will always be assumed below, except when otherwise stated, that substitution scores are normalized (multiplied by a suitable constant) so as to be expressed in nats. Specifying a set of target frequencies then completely specifies a set of scores. How these frequencies can be chosen will be discussed in the next section.

Given normalized scores, the statistical theory for high scoring alignments takes on a particularly simple form. When two random sequences are compared, the number of distinct, or locally optimal (Sellers 1984), segment pairs expected to have score at least x is well approximated by the formula

$$KN\, e^{-x} \tag{2}$$

where K is a calculable constant dependent on the scoring system, and N is the product of the sequences' lengths (Karlin and Altschul 1990; Karlin et al. 1990, Dembo and Karlin 1991). More precisely, the MSP scores take on an *extreme value* distribution (Gumbel 1958) whose cumulative distribution function is given by the formula

$$\exp[-e^{-\lambda(x-u)}] \tag{3}$$

where the *scaling factor* λ is 1, and the *characteristic value* u is given by the formula

$$u = \ln KN \tag{4}$$

The PAM Model of Protein Evolution

Log-odds scores for protein sequence comparison were first proposed by Dayhoff et al. (1978), though in the context of global sequence alignment. In order to calculate an appropriate set of target frequencies, they constructed from a study of homologous protein families the so-called PAM (for point-accepted mutation) model of protein evolution. The model presents protein evolution as a stochastic process in which at each time a given amino acid residue has certain fixed probabilities of mutating into any of the other residues. One PAM is defined as the amount of evolutionary change required for an expected 1% of all residues (suitably weighted by their relative frequencies) to mutate. The manner in which the particular numbers in the PAM model were derived has been criticized (Wilbur 1985), and a great deal of additional data has accumulated since the model was proposed. Therefore it is certainly possible that better numbers than those derived by Dayhoff et al. (1978) can be found. This paper discusses a general manner in which any PAM-like model may be generalized, and for this purpose the values used by Dayhoff et al. (1978) will suffice. In fact, these numbers have proved very successful over the years (Feng et al. 1985).

Given the PAM model, it is easy to calculate the frequency q_{ij} with which any pair of amino acids are expected to correspond when two homologous proteins separated by a particular degree of evolutionary change are properly aligned. These target frequencies are then used to calculate "PAM scores" for the chosen evolutionary distance, using equation (1) above. For years the most popular scoring systems for protein sequence comparison were based upon PAM-250 scores (Dayhoff et al. 1978), i.e., scores derived from the target frequencies corresponding to 250 PAMs of protein evolution. More recently it has been argued that for database searches PAM-120 scores are generally more effective, and that in fact several different PAM matrices should be employed (Altschul 1991). Below we extend this line of reasoning to propose an "All-PAM" scoring system and to study its associated statistics.

The All-PAM Scoring System

A given PAM matrix is adapted to finding segments that have diverged by a specific amount of evolutionary change. Since in general it is not known what evolutionary distance is appropriate to an alignment that has yet to be found, it is important to use a variety of PAM matrices when comparing two sequences or searching a database with a query sequence (Collins et al. 1988). It has been argued that for many purposes two or three different PAM matrices will suffice (Altschul 1991). Nevertheless, one may imagine trying all possible PAM matrices in order to find that one which will maximize an alignment's score. This in effect defines a new scoring system, but one with certain subtle aspects that will be considered below.

Given a set of scores for PAM distance d and a pair of random protein sequences, the number of distinct segment pairs expected by chance to have a score of at least x is given by formula (2) above, where K_d (the K corresponding to PAM distance d) is readily calculated (Karlin and Altschul 1990). The value of K_d varies with PAM distance, decreasing as d increases. Therefore, if one uses two distinct PAM matrices to compare the same random protein sequences, the matrix of lower PAM distance is expected to yield more high-scoring segment pairs. Assuming we have no reason to favor one PAM distance to another, we should correct for this statistical effect by adjusting the PAM scores so that they all have the same asymptotic random distribution. This can be accomplished easily by subtracting $\ln K_d$ from the score of any segment pair. In other words, if S_d is the score for a segment pair using a PAM-d matrix, define the *corrected* score as

$$S'_d = S_d - \ln K_d \tag{5}$$

Notice that this does *not* change the scores for aligning pairs of amino acids. Rather, every alignment can be thought of as "starting" with the score $-\ln K_d$. Such corrected scores have identical asymptotic extreme value distributions, with characteristic value $\ln N$ and scale factor 1.

One difficulty with the approach just described is that since K_d gets arbitrarily small as d grows, the corrected score for great PAM distances may begin at arbitrarily large values. This is not really a problem because, as is argued in Appendix A, for a given finite-sized comparison there is a greatest PAM distance it makes sense to consider. Letting D

be this largest PAM distance, we are in a position to define the All-PAM score A of an alignment:

$$A = \max_{d \leq D} S'_d = \max_{d \leq D} (S_d - \ln K_d) \qquad (6)$$

While the utilization of a range of PAM matrices to search for local protein similarities has been advocated previously (Collins et al. 1988; Altschul 1991; Gonnet 1993), equation (6) introduces the $\ln K_d$ correction, as well as the formal definition of an All-PAM similarity measure. We proceed below to study the statistical behavior of this function.

An Extreme Value Theory for All-PAM Scores

As discussed above, each corrected PAM score S'_d has an extreme value distribution with characteristic value $u = \ln N$. No formal theory for the statistics of the All-PAM scores defined in equation (6) has yet been established. Nevertheless, analogy with other related scores (Smith et al. 1985; Arratia et al. 1986; Altschul and Erickson 1986; Arratia et al. 1988; Arratia and Waterman 1989; Karlin and Altschul 1990; Mott 1992) strongly suggests that they also should follow an extreme value distribution. The key question is to what extent optimizing the score by allowing the PAM distance to vary increases the characteristic value of the optimal local alignment.

One way to approach this problem is to imagine that optimizing over PAM distances between 0 and D is essentially equivalent to optimizing over an effective number E of independent scoring systems. It is argued in Appendix B that, whatever the precise form of the PAM model and whatever the choice of D, in the limit E should be proportional to $\sqrt{\ln N}$. Optimizing over E independent scoring systems is equivalent to multiplying the effective search space by E; it should therefore increase the characteristic value of the optimal alignment by $\ln E$. This implies that a formula for the characteristic value u of All-PAM scores should take the form

$$u = \ln N + \frac{1}{2} \ln \ln N + C \qquad (7)$$

where the constant C will depend upon the PAM model and the range of PAM distances over which the All-PAM score is defined. (See Appendix B.)

Equation (7) has been established analytically for certain simple cases (S. Karlin, personal communication), but in the present context is only an hypothesis, and therefore can benefit from empirical support. Thus we have tested this model by random simulation, as described in the following section.

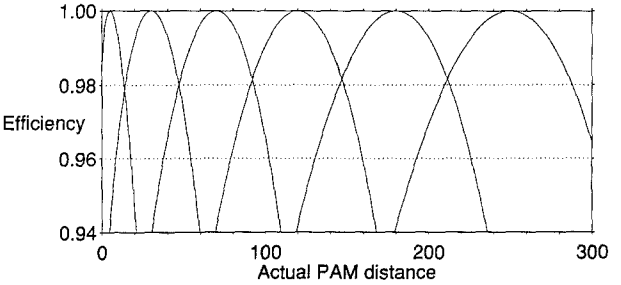

Fig. 1. The efficiency of PAM-d scores (d equals 5, 30, 70, 120, 180, and 250) as a function of the actual PAM distance of the sequences being compared. PAM-d scores have their maximum efficiency (1.0) when the actual PAM distance equals d.

A Monte Carlo Simulation

While we wish to study the behavior of All-PAM scores, no efficient algorithm for truly optimizing PAM-d scores over a continuum of possible PAM distances d has been described. We have therefore approximated All-PAM scores by maximizing S'_d over a relatively small set of selected distances. An appropriate set of distances can be chosen by considering the *efficiency* of various PAM matrices as a function of actual PAM distance (Altschul 1991; States et al. 1991). The efficiency of a PAM-d matrix at PAM distance x is defined as the ratio of the expected PAM-d to the expected PAM-x score, for a proper alignment of sequences actually separated by PAM distance x. Because the highest expected score at this distance is always the PAM-x score, the efficiency of PAM-d scores is always less than or equal to 1.0. Intuitively, it measures what proportion of the potential information available at PAM distance x is captured by using a PAM-d matrix.

The efficiency curves for PAM-d scores, with d equal to 5, 30, 70, 120, 180, and 250, are graphed in Fig. 1. It is evident that at virtually all PAM distances in the realistically detectable range 0 to 290 (Appendix A), the best of these six scoring systems is at least 98% efficient. In the Monte Carlo simulation discussed below, All-PAM scores were therefore approximated by maximizing over PAM scores for these six discrete distances.

We modeled protein sequences by strings of independently chosen amino acids, selected with the frequencies used in the PAM model of Dayhoff et al. (1978). For a given pair of lengths m and n (yielding a search space of size $N = mn$), we generated two random protein sequences and found their approximate All-PAM score, as described above. Using the method of moments (Altschul and Erickson 1986), we estimated u and λ for the extreme value distribution corresponding to lengths m and n from 20,000 such random scores. For m fixed at 248 and varying n, these estimates $\hat{u}$ and $\hat{\lambda}$ are given in Table

Table 1. The characteristic value of All-PAM MSP scores as a function of search space size[a]

n	ln ln N	$\hat{u}$	$\hat{u} - \ln N$	$\hat{\lambda}$
248	2.40	11.999	0.972	0.97
309	2.42	12.250	1.003	0.96
387	2.44	12.495	1.023	0.98
489	2.46	12.731	1.025	0.97
619	2.48	12.979	1.037	0.97
788	2.50	13.237	1.054	0.97
1,007	2.52	13.490	1.062	0.98
1,295	2.54	13.746	1.066	0.95
1,673	2.56	14.011	1.075	0.97
2,172	2.58	14.296	1.099	0.97
2,836	2.60	14.565	1.101	0.97
3,723	2.62	14.858	1.122	0.97

[a] Two random protein sequences of lengths $m = 248$ and n were chosen, and their All-PAM MSP score was approximated by maximizing S' over PAM distances 5, 30, 70, 120, 180, and 250. From 20,000 such scores, estimates $\hat{u}$ and $\hat{\lambda}$ of the characteristic value and scale factor for an extreme value distribution were obtained by the method of moments (Altschul and Erickson 1986). Values of $\hat{u}$, $\hat{u} - \ln N$, and $\hat{\lambda}$ are reported as a function of ln ln N, where $N = mn$. The standard error for each is 0.01

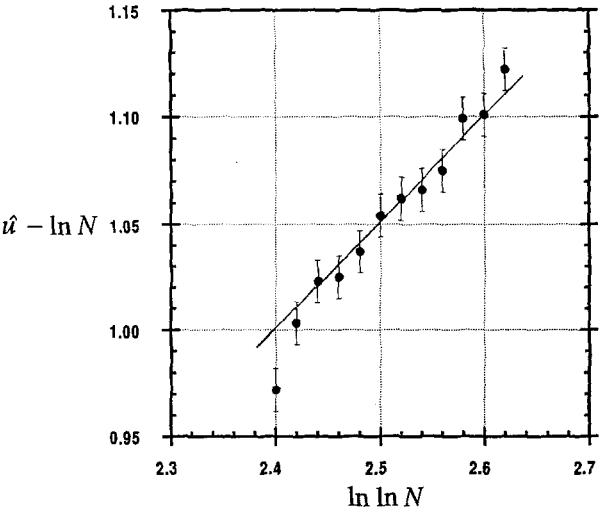

Fig. 2. The difference between the estimated characteristic extreme value $\hat{u}$ of All-PAM scores and ln N, as a function of ln ln N, where N is the size of the search space. $\hat{u}$ was estimated by the method of moments (Altschul and Erickson 1986) from 20,000 random values. Error bars represent the standard error of the estimates. The theory represented by equation (7) implies these points should lie on a line with slope 1/2. The best such line for fitting those points with ordinate greater than 2.40 is plotted.

1, and $\hat{u} - \ln N$ is plotted in Fig. 2 as a function of ln ln N. The standard error for each $\hat{u}$ and $\hat{\lambda}$ is 0.01.

Equation (7) implies that the quantity $\hat{u} - \ln N$ should be a linear function of ln ln N, with slope 1/2. Except for the smallest N, Fig. 2 shows that the data fit the theory remarkably well. Linear regression on the points with ln ln $N \geq 2.42$ yields a slope of 0.55 ± 0.05. The line plotted in Fig. 2, $\hat{u} - \ln N = 0.5 \ln \ln N - 0.199$, is the best-fitting one with slope 1/2. Similar results were obtained for other values of m (data not shown).

Because ln ln N is such a slowly growing function, it is not easy to detect the hypothesized deviation of $\hat{u}$ from ln N with increasing N. From one point to the next in Fig. 2, the y-coordinate is expected to increase by only 0.01, but even with 20,000 random trials per point this is the standard error for y. Nevertheless, over the domain investigated, the slope of the line becomes well defined. Thus, given our domain is appropriate for detecting asymptotic effects, and given the correct formula for $\hat{u}$ does contain a log-log term, we can confidently conclude that its coefficient is much closer to 1/2 than it is to either 1 or 0. Appendix B, of course, argues that the coefficient is exactly 1/2.

Of less relevance to the main thrust of this paper, but nevertheless of some interest, is the idea that while in the infinite limit the scale factor λ for the extreme value distribution of All-PAM scores should approach 1, for finite N a better estimate of λ is $1 - (1/2u)$. This is discussed briefly in Appendix B and is supported by the Monte Carlo estimates of λ recorded in Table 1.

A closer approximation to true All-PAM scores than the one studied above would yield slightly higher scores, and thus a slightly higher value of C than the -0.20 found above. Random simulation as well as theoretical considerations (Appendix B) suggest that for true All-PAM scores over the range 0 to 290, C would be increased by about 0.15 to -0.05.

Practical Approximations to All-PAM Scores

It is instructive here to return to the motivation for equation (7): the idea that the effective number E of independent PAM matrices embodied by All-PAM scores is proportional to $\sqrt{\ln N}$. The fact that, for the All-PAM range studied, C in (7) is about -0.05, means that E can be well approximated as

$$E \approx 0.95 \sqrt{\ln N} \tag{8}$$

From equation (8), we calculate that when two proteins of average length are compared, $N \approx 60{,}000$ and $E \approx 3.2$. Similarly, when a protein of average length is compared with a comprehensive protein sequence database of typical current size, $N \approx 10^{10}$ and $E \approx 4.6$. These numbers are important for deciding when the theory developed here has fruitful application. For instance, if All-PAM scores for protein database searches are approximated by taking the maximum of $E^* < 5$ distinct PAM-d matrices, then treating these matrices as independent, i.e., estimating u as ln (E^*N), is conservative but still superior to equation (8). On the other hand, if

All-PAM scores are estimated by taking the maximum of five or more corrected PAM-d scores, equation (8) provides a tighter upper bound on the effective number of independent comparisons actually performed. As described above, 98% efficiency can be achieved over the detectable PAM range by using the six PAM matrices 5, 30, 70, 120, 180, and 250 (Fig. 1). For typical database searches these, or any greater number of matrices, count as fewer than five independent scoring systems. It should be noted that nothing is wrong with optimizing PAM distance "after the fact." In other words, given an alignment found by any method, one may vary d at will in order to maximize the alignment's S'_d score and still obtain a valid estimate of this All-PAM score's statistical significance.

For database searching, All-PAM MSP scores can be calculated in reasonable time by specially designed microchips (Chow et al. 1991; Hughey 1991; White et al.1991), by parallel architecture machines (Coulson et al. 1987), or by fast heuristic approximation algorithms such as BLAST (Altschul et al. 1990). For the last of these, running the same search many times using a large number of different PAM matrices can result in an unacceptable decrease in speed. This can be avoided by using at first only a small number of widely spaced PAM matrices, e.g., PAM-30, -120, and -250. These three matrices alone should be able to locate any segment pair with potentially significant All-PAM score. Once the relevant alignments have been located, they may be rescored (with possible attendant trimming or extension) by using a more tightly spaced set of matrices. This strategy allows All-PAM scores to be approximated closely in only a small multiple of the time required for a single PAM-d score.

The use of All-PAM scores implies an ignorance concerning the evolutionary distance inherent in the similarities sought; the data are used to select an optimal distance. The estimation of this unknown parameter must be paid for with lost statistical significance. As discussed above, for typical database searches, the choice of an optimal PAM distance from the range 0 to 290 costs a factor of about 4.6 in significance, or about 2.2 bits of information. The single set of PAM-120 scores is superior to such All-PAM scores if one knows for certain that the similarities sought have PAM distances between 70 and 170. The problem arises for similarities with implicit PAM distance outside this range. PAM-120 scores are, respectively, only 82% and 50% efficient for alignments with true PAM distances 30 and 250. Thus significant alignments at these two distances, with 36 bits of potential (i.e., All-PAM) information, are left, respectively, with PAM-120 scores of only 30 and 18 bits. Employing All-PAM scores can therefore be seen as a decision always to sacrifice about two bits of information so as to avoid occasional losses of much greater magnitude.

A Biological Example

The value of All-PAM scores is perhaps best observed in a database search employing as query sequence a member of a large and diverse protein superfamily. For such a sequence, proteins related at a wide range of PAM distances are likely simultaneously to be present. All-PAM scores are of use, however, whenever the relationships sought are of unknown implicit distance.

We used the BLAST algorithm (Altschul et al. 1990) to search the PIR protein sequence database (Release 31; Barker et al. 1990) with soybean leghemoglobin c 1 (PIR code GPSYC1; Hyldig-Nielsen et al. 1982), using amino acid substitution matrices for PAM distances 5, 30, 70, 120, 180, and 250. From these searches many proteins exhibited statistically significant similarities to the query. Alignments illustrating three such similarities are shown in Fig. 3, along with the All-PAM score for each rounded to the nearest bit; the PAM matrix from which the All-PAM score derives is indicated in brackets. In the context of the search performed $N \approx 1.5 \times 10^9$; from the formulas developed above one may conclude that an All-PAM score of 37 bits or greater is statistically significant (P value ≤ 0.05). All three alignments shown in Fig. 3 represent true homologies and meet this significance criterion.

How do these three homologous pairs of sequence fare under the individual scoring systems whose maximum constitutes the All-PAM score? For a database search with N as before, and using a single substitution matrix, a corrected PAM-d score of 35 bits is statistically significant; as discussed above, about two fewer bits are required than for All-PAM statistics. Table 2 records, for each PAM-d matrix employed, S'_d for the three homologous sequence pairs. It is evident that at no individual distance d do all three alignments simultaneously appear significant. For instance, if PAM-120 scores are used, the alignments involving the soybean leghemoglobin fragment (PIR code B32711; Stougaard et al. 1987) and the bacterial hemoglobin (PIR code GGZLB; Wakabayashi et al. 1986) both appear highly significant, but the long and weak alignment involving sea cucumber globin (PIR code S15979; Mauri et al. 1991) is missed.

In most database searches, homologies at such a range of evolutionary distances will not be found, and it will perhaps appear "after the fact" that a single specific PAM-d matrix was the correct one to use. The challenge, however, is to select this matrix before the search is performed. As argued earlier

```
(a)  All-PAM score:  48 bits  [PAM-30]        P-value:  4 x 10^-5

             GPSYC1    3 FTEKQEALVSSSFEAFKANIP 23
                         FT  Q ALV SS EAFK N P
             B32711    2 FTAQQDALVGSSYEAFKQNLP 22


(b)  All-PAM score:  41 bits  [PAM-120]       P-value:  0.004

GPSYC1     92 HAQKAVTDPQFVVVKEALLKTIKEAVGGNWSDELSSAWEVAYDELA 137
              H Q  V       V   LL  IKE  G    D    AW  AY   A
GGZLB      85 HCQAGVAAAHYPIVGQELLGAIKEVLGDAATDDILDAWGKAYGVIA 130


(c)  All-PAM score:  37 bits  [PAM-250]       P-value:  0.05

GPSYC1     47 LANGVDPTNPKLTGHAEKLFALVRDSAGQLKTNGTVVADAALVSIHA  93
              L      T      HA    AL         T       A L   H
S15979     59 LSPAELRTSRQMHAHAIRVSALMTTYIDEMDTEVLPELLATLTRTHD 105

GPSYC1     94 QKAVTDPQFVVVKEALLKTIKEAVGGNWSDELSSAWEVAYDELAAAI 140
                V           L   IK   G         AW
S15979    106 KNHVGKKNYDLFGKVLMEAIKAELGVGFTKQVHDAWAKTFAIVQGVL 152
```

Fig. 3. The All-PAM maximal segment pairs for comparisons of soybean leghemoglobin *c* 1 (PIR code GPSYC1) with (**a**) a soybean leghemoglobin fragment (PIR code B32711); (**b**) a bacterial hemoglobin (PIR code GGZLB); and (**c**) a sea cucumber globin (PIR code S15979). Identical residues are echoed on the central line of each alignment. *P* values are based on All-PAM statistics and a search space size of $N \approx 1.5 \times 10^9$.

Table 2. Corrected PAM-*d* scores for optimal local alignments of soybean leghemoglobin *c* 1 (PIR code GPSYC1) with various other protein sequences[a]

PIR code	Corrected PAM-*d* score (in bits) for *d* equal to					
	5	30	70	120	180	250
B32711	36	48	46	40	33	27
GGZLB	16	18	37	41	40	35
S15979	17	18	19	29	35	37

[a] B32711, soybean leghemoglobin fragment; GGZLB, bacterial hemoglobin; S15979, sea cucumber globin

(Altschul 1991), the PAM-120 matrix is probably the best single choice for general purposes, but as seen here it can have its failings. The more conservative course therefore is to employ All-PAM scores for the realistically detectable range of evolutionary distances.

Discussion and Conclusion

The idea of seeking biologically interesting similarities by using a range of different PAM matrices was first investigated, in the global alignment context, by the originators of the PAM model of molecular evolution (Schwartz and Dayhoff 1978). This proposal has been frequently advocated since (Collins et al. 1988; Altschul 1991; States et al. 1991; Gonnet 1993). In this paper we have attempted to give the idea a formal embodiment in the form of All-PAM scores and to study the statistics of this measure of sequence similarity. The study has yielded several conclusions. If one does not know a priori the appropriate PAM distance for the similarities sought, there is a price for estimating this parameter from the data. In the context of a typical protein sequence database search, choosing an optimal PAM distance in the range 0–290 costs a factor of about 4.6 in statistical significance (2.2 bits). If one is unwilling to pay this price—for example, if one uses only PAM-120 scores—one may gain slightly for the bulk of similarities in the PAM-70 to PAM-170 range, but those far outside this range may be missed completely.

Some mention should be made here of the "nonlinear similarity functions" proposed earlier by the author (Altschul and Erickson 1986). The motivation for these functions was similar in spirit to that for the All-PAM scores studied above. However, both theory and experience now suggest that All-PAM scores are superior for the purpose of revealing biologically significant similarities. It is worth noting that statistical studies of the "nonlinear similarity functions" revealed the presence of a log-log term in the growth of their characteristic values (Altschul and Erickson 1986; Altschul and Erickson 1988; Waterman and Gordon 1990). It is likely that an analysis such as that outlined here could be applied successfully to those functions as well.

We have discussed All-PAM scores only in the context of local alignments lacking gaps. Unfortunately, no good statistical theory yet exists to permit the relaxation of this restriction. Certain results suggest, however, that the general spirit of our anal-

ysis should apply as well to alignments with gaps (Smith et al. 1985; Waterman et al. 1987; Mott 1992). Even when gaps are allowed, however, it should be understood that alignment scores based on simple substitution matrices, such as those studied here, exclude many potential sources of information concerning biological relatedness. If several related sequences are available to query a database, rather than just a single one, more general weight matrices or "profiles" can often be of use (Taylor 1986; Gribskov et al. 1987; Patthy 1987). Even in the context of pairwise alignments, complex scoring systems may be able to extract more evidence for biological relatedness than can simple substitution matrices (Argos 1987). However, algorithms for optimizing such measures can require considerably more computation time. The heuristic BLAST programs are able to search typical protein sequence databases for optimal All-PAM alignments in a matter of minutes (Altschul et al. 1990), while when run on a similar machine, programs based on more complicated measures may require as much as 2 weeks (Vogt and Argos 1992).

While the analysis in this paper has been tied to the PAM model of protein evolution, most of the ideas developed should apply to any singly parametrized set of scores used either for sequence comparison (e.g., the recently developed system of Henikoff and Henikoff 1992) or for "single sequence analysis" (Karlin and Altschul 1990). Such parametrized scores arise naturally from an evolutionary model, but they also occur in other contexts. For example, one may seek regions of a protein rich in a given residue type (Karlin et al. 1991), but not know a priori what target frequency to expect; a parametrized range of scores is then appropriate.

In summary, the theory developed in this paper elucidates some of the choices implicit in the selection of a specific amino acid substitution matrix, or of All-PAM scores. It is hoped that this will lead to better informed selection of scoring systems for macromolecular sequence comparison, and therefore to more sensitive detection of biologically interesting similarities. While these ideas have not been shown to extend beyond the particular context in which they have been presented, their general spirit should find wider application.

Appendix A: Approximate Upper Bounds on Detectable PAM Distances

The greater the PAM distance by which two segments have diverged, the less information each aligned pair of residues carries, and the longer the segment pair need be for its signal to rise above background noise. This intuition is captured quantitatively by the *relative entropy* H_d corresponding to PAM distance d (Altschul 1991). In short, the expected PAM-d score per properly aligned residue pair is given by the formula

$$H_d = \sum_{i,j} q_{ij} \ln \frac{q_{ij}}{p_i p_j} \tag{9}$$

where the q_{ij} are the PAM-d target frequencies (Altschul 1991). High-scoring random segment pairs will also be characterized by these frequencies (Karlin and Altschul 1990; Karlin et al. 1990; Dembo and Karlin 1991), and thus by the relative entropy H_d. Employing this quantity, a rough analysis of the greatest PAM distance D detectable in a comparison of two sequences of length m is possible.

For large d, the correction constant K_d employed above approaches H_d (S. Karlin, personal communication). Therefore two sequences of length m that are related at PAM distance d, have an expected S'_d given by

$$mH_d - \ln K_d \approx mH_d - \ln H_d \tag{10}$$

Since the characteristic maximal S'_d expected to occur by chance from the comparison of two such sequences is $\ln N = \ln m^2$, we require that

$$mH_d - \ln H_d > \ln m^2 \tag{11}$$

if the score for the true relationship is to rise above background noise. Elementary calculus shows that the left-hand side of (11) decreases monotonically with decreasing H_d, when $H_d > 1/m$. (Beyond this point, arbitrarily large but misleading scores S'_d can be achieved simply by choosing d large, making H_d arbitrarily small. It can be shown, however, that the correction term $-\ln K_d \approx -\ln H_d$ in the definition of S'_d is valid only in the limit of large m and that H_d should therefore always be kept larger than $1/m$.) Simple substitution shows that for $H_d = (\ln m)/m$, the bound of inequality (11) is already violated. We will therefore take the PAM distance D that yields

$$H_D = \frac{\ln m}{m} \tag{12}$$

as an effective upper bound on detectable PAM distances. Notice that if two sequences of unequal lengths are compared, and m is the length of the shorter sequence, the bound of (12) is still conservative. Using this equation for a typical protein length of $m = 250$, we get $H_d = 0.022$ nats, which corresponds to a PAM distance $D \approx 680$.

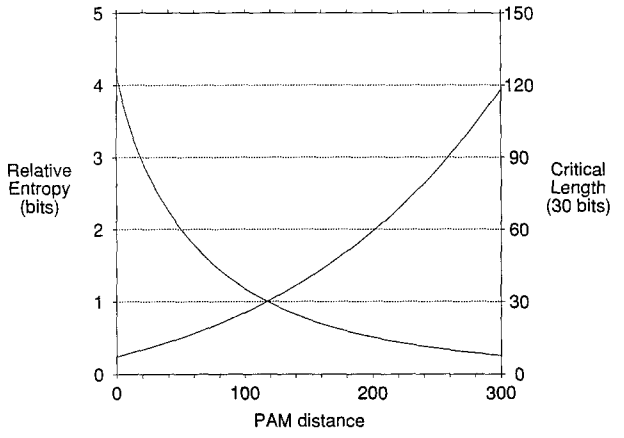

Fig. 4. The relative entropy and the length of the shortest alignment to attain significance, as a function of PAM distance. Relative entropy decreases with increasing PAM distance, is expressed in bits, and is measured by the scale to the left of the graph. Critical length—the length in residues of the shortest alignment expected to achieve a score of 30 bits—increases with increasing PAM distance and is measured by the scale on the right of the graph.

The foregoing analysis assumes that a homology stretching the entire length of the protein is present, and of course that it does not involve gaps. In practice, it is rare that regions of such length are conserved over great evolutionary distances. It is much more typical for shorter regions, crucial in some way to function, to be conserved, and for such regions only smaller PAM distances will be detectable. For typical database searches an uncorrected MSP score of at least 30 bits generally is required to distinguish a protein similarity from background noise (Altschul 1991). For any PAM distance d, the relative entropy H_d of equation (9) allows us to analyze how long an MSP need be to achieve such a score. In Fig. 4 is graphed H_d (expressed in bits) as a function of PAM distance, along with the minimum alignment length required on average to attain 30 bits of information. It can be seen that by PAM distance 290, an MSP needs to be over 100 residue pairs long to achieve significance. While as described above greater PAM distances theoretically are detectable, this is an effective upper bound for real protein database searches, and the one we employ in this paper.

Appendix B: The Characteristic Extreme Value of All-PAM Scores

Comparing two random proteins of lengths m and n, let $\mathbf{A}_c$ and $\mathbf{A}_d$ be the maximal segment pairs at the two PAM distances c and d. As described in the text, the corrected scores for these alignments should have identical asymptotic extreme value distributions with characteristic value ln N, where $N = mn$. For c and d very different, these alignments should be essentially independent. If S is the maximum of their scores, S should then be extreme value distributed, with characteristic value ln $2N$ (Gumbel 1958). On the other hand, if $c = d$, S trivially has characteristic value ln N. For c close but not identical to d, we expect an intermediate case: S should be extreme value distributed with characteristic value ln EN for some number E between 1 and 2.

Maximizing a continuous range of corrected PAM-d scores, with d between 0 and D, confronts us with an analogous situation. In a small neighborhood of a given d the optimal alignments are all correlated, so their maximum score should not differ much from S'_d. A large range of d, however, should encompass a number of essentially independent alignments. Maximizing S'_d should again be equivalent to maximizing over an effective number E of independent scoring systems. It is hypothesized here that E should not be fixed but should grow proportionally with $\sqrt{\ln N}$. More specifically, for a continuous range of PAM distance indexed by the parameter x, it is hypothesized that in the limit of large N, E is given by the formula

$$E = \int \sqrt{(J_x \ln N)/2\pi H_x}\, dx = \sqrt{\ln N} \int \sqrt{J_x/2\pi H_x}\, dx \qquad (13)$$

where H_x is the relative entropy of PAM-x scores defined in (9), J_x is the *Fisher information* (Fisher 1925) of the target frequency distribution at distance x, and the integration is of course performed over the relevant range of PAM distances. The motivation for this equation will be outlined below. Here we merely observe that equation (7) in the text follows from equation (13), with C given by

$$C = \ln \int \sqrt{J_x/2\pi H_x}\, dx \qquad (14)$$

We also note that this formula for C is invariant, as it must be, under any alternative parametrization of the chosen range of PAM distance.

This hypothesis embodied in equation (13) originates from the idea that for a continuous, parameteized set of correlated random variables (in this case the alignments $\mathbf{A}_x$) there should be some quantity that behaves like a "density" of independent random variables. The most suitable candidate for this quantity derives from the Fisher information J_x (Fisher 1925) for the target frequency distribution $\mathbf{q}_x$ at PAM distance x. In the limit of large N, an optimal local alignment for this distance is expected to have length $(\ln N)/H_x$, and its aligned amino acid pairs are expected to occur with the target frequencies $\mathbf{q}_x$ (Karlin and Altschul 1990; Karlin et al. 1990; Dembo and Karlin 1991). Since the Fisher informa-

tion for n independent identically distributed (i.i.d.) samples of a random variable is n times the Fisher information for an individual sample (Cover and Thomas 1991), the Fisher information for $\mathbf{A}_x$ should approach $(J_x \ln N)/H_x$. Intuitively, this quantity is a measure of the rate at which the alignments $\mathbf{A}_x$ become independent of one another. While the Fisher information should correlate with the sought "density" of independent random variables, a scaling argument shows that only its square root provides a consistent measure under any change in parametrization. (Alternatively, the second derivative of the efficiency curves shown in Fig. 1, at their maximum, can be shown to equal $-J_x/H_x$. Since the maximum score is always approximately $\ln N$, the "density of curves" required to lose no more than ϵ bits of information is therefore proportional to $\sqrt{(J_x \ln N)/H_x}$. Short of the factor $\sqrt{2\pi}$ in the denominator, this motivates equation (13). This final factor can be derived from the limiting case that an isolated PAM matrix should count as one independent random variable.

The arguments above in support of equations (13) and (14) are completely intuitive and nonrigorous. Nevertheless, they help motivate the $1/2 \ln \ln N$ term in equation (7), which can be established for certain simple cases (S. Karlin, personal communication). The arguments generalize naturally to sets of scores parametrized by k independent variables, for which they imply a $k/2 \ln \ln N$ growth term in equation (7); this also can be established rigorously in some simple cases (S. Karlin, personal communication).

Using equation (14) to estimate C for the effective PAM range 0–290 discussed in Appendix A yields $C \approx -0.05$. This is slightly larger than the value of $C \approx -0.20$ found by the Monte Carlo simulation in the text. That simulation, however, is based upon an approximation to All-PAM scores using matrices for the six discrete PAM distances 5, 30, 70, 120, 180, and 250. True All-PAM scores would have yielded a somewhat greater value for C. While a rigorous theory is perhaps unlikely to validate equation (14) in detail, its success here suggests it may nevertheless have a place as a useful approximation.

In the limit of large N we expect All-PAM MSP scores to approach an extreme value distribution with characteristic value u given by equations (7) and (14), and scale factor $\lambda = 1$. It should be noted, however, that for finite N a small correction to λ is appropriate. This derives from the fact that, independent of N, a higher All-PAM score implies a greater number of "effective independent trials." Properly accounting for this suggests that for All-PAM scores λ should be approximately $1 - (1/2u)$, which of course approaches 1 in the limit of large N. The Monte Carlo estimation of λ described in the text and recorded in Table 1 supports this correction.

Acknowledgments. The author thanks Drs. Warren Gish, Gaston Gonnet, Phil Green, Samuel Karlin, David Lipman, Chris Sander, David States, and John Wilbur for helpful conversations.

References

Altschul SF (1991) Amino acid substitution matrices from an information theoretic perspective. J Mol Biol 219:555–565

Altschul SF, Erickson BW (1986) A nonlinear measure of subalignment similarity and its significance levels. Bull Math Biol 48:617–632

Altschul SF, Erickson BW (1988) Significance levels for biological sequence comparison using non-linear similarity functions. Bull Math Biol 50:77–92

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (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

Arratia R, Gordon L, Waterman MS (1986) An extreme value theory for sequence matching. Ann Stat 14:971–993

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

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

Barker WC, George DG, Hunt LT (1990) Protein sequence database. Methods Enzymol 183:31–49

Chow ET, Hunkapiller T, Peterson JC, Zimmerman BA, Waterman MS (1991) A systolic array processor for biological information signal processing. In: Proceedings of the 1991 international conference on supercomputing. ACM Press, New York, pp 216–223

Collins JF, Coulson AFW, Lyall A (1988) The significance of protein sequence similarities. Comput Appl Biosci 4:67–71

Coulson AFW, Collins JF, Lyall A (1987) Protein and nucleic acid database searching: a suitable case for parallel processing. Computer J 30:420–424

Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York, pp 326–329

Dembo A, Karlin S (1991) Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables. Ann Prob 19:1737–1755

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5, suppl 3. Natl Biomed Res Found, Washington, pp 345–352

Feng DF, Johnson MS, Doolittle RF (1985) Aligning amino acid sequences: comparison of commonly used methods. J Mol Evol 21:112–125

Fisher RA (1925) Theory of statistical estimation. Proc Cambridge Phil Soc 22:700–725

Goad WB, Kanehisa MI (1982) Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. Nucl Acids Res 10:247–263

Gonnet GH (1993) A tutorial introduction to computational biochemistry using Darwin. Manuscript in preparation

Gonnet GH, Cohen MA, Benner SA (1992) Exhaustive matching of the entire protein sequence database. Science 256:1443–1445

Gumbel EJ (1958) Statistics of extremes. Columbia University Press, New York

Gribskov M, McLachlan AD, Eisenberg D (1987) Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 84:4355–4358

Hamming RW (1986) Coding and information theory. Prentice-Hall, Englewood Cliffs, p 106

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919

Hughey RP (1991) Programmable systolic arrays. PhD Thesis, Brown University, Providence

Hyldig-Nielsen JJ, Jensen EO, Paludan K, Wiborg O, Garrett R, Jorgensen P, Marcker KA (1982) The primary structures of two leghemoglobin genes from soybean. Nucl Acids Res 10: 689–701

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 87:2264–2268

Karlin S, Bucher P, Brendel V, Altschul SF (1991) Statistical methods and insights for protein and DNA sequences. Annu Rev Biophys Biophys Chem 20:175–203

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

Karlin S, Ost F (1988) Maximum length of common words among random letter sequences. Ann Prob 16:535–563

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441

Mauri F, Omnaas J, Davidson L, Whitfill C, Kitto GB (1991) Amino acid sequence of a globin from the sea cucumber *Caudina* (*Molpadia*) *arenicola*. Biochim Biophys Acta 1078:63–67

McLachlan AD (1971) Tests for comparing related amino-acid sequences. Cytochrome *c* and cytochrome c_{551}. J Mol Biol 61:409–424

Mott R (1992) Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. Bull Math Biol 54:59–75

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. J Mol Biol 48:443–453

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

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448

Rao JKM (1987) New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. Int J Peptide Protein Res 29:276–281

Risler JL, Delorme MO, 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 JB (1983) Time warps, string edits and macromolecules: the theory and practice of sequence comparison. Addison-Wesley, Reading

Schwartz RM, Dayhoff MO (1978) Matrices for detecting distant relationships. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5, suppl 3. Natl Biomed Res Found, Washington, pp 353–358

Sellers PH (1974) On the theory and computation of evolutionary distances. SIAM J Appl Math 26:787–793

Sellers PH (1984) Pattern recognition in genetic sequences by mismatch density. Bull Math Biol 46:501–514

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Smith TF, Waterman MS, Burks C (1985) The statistical distribution of nucleic acid similarities. Nucl Acids Res 13:645–656

States DJ, Gish W, Altschul SF (1991) Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. Methods 3:66–70

Stougaard J, Petersen TE, Marcker KA (1987) Expression of a complete soybean leghemoglobin gene in root nodules of transgenic *Lotus corniculatus*. Proc Natl Acad Sci USA 84: 5754–5757

Taylor WR (1986) Identification of protein sequence homology by consensus template alignment. J Mol Biol 188:233–258

Vogt G, Argos P (1992) Searching for distantly related protein sequences in large databases by parallel processing on a transputer machine. Comput Appl Biosci 8:49–55

Wakabayashi S, Matsubara H, Webster DA (1986) Primary sequence of a dimeric bacterial haemoglobin from *Vitreoscilla*. Nature 322:481–483

Waterman MS, Gordon L (1990) Multiple hypothesis testing for sequence comparisons. In: Bell GI, Marr TG (eds) Computers and DNA. Addison-Wesley, Reading, pp 127–135

Waterman MS, Gordon L, Arratia R (1987) Phase transitions in sequence matches and nucleic acid structure. Proc Natl Acad Sci USA 84:1239–1243

White C, Singh RK, Reintjes PB, Lampe J, Erickson BW, Dettloff WD, Chi VL, Altschul SF (1991) BioSCAN: A VLSI-based system for biosequence analysis. In: Proceedings of the 1991 IEEE international conference on computer design: VLSI in computers and processors. IEEE Comp Soc Press, Los Alamitos, pp 504–509

Wilbur WJ (1985) On the PAM matrix model of protein evolution. Mol Biol Evol 2:434–447

Received June 22, 1992/Accepted August 24, 1992