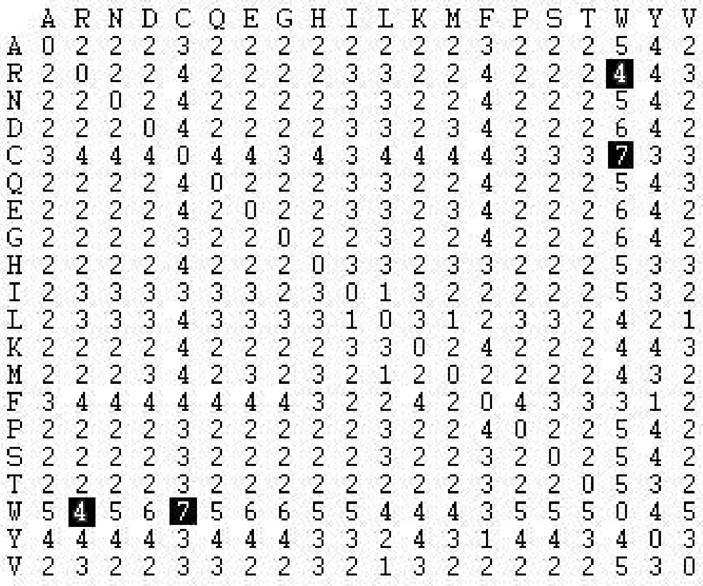
Errata for “A Metric Model of Amino Acid Substitution”

Weijia Xu

This errata is for the paper “A Metric Model of Amino Acid Substitution,” as published in *Bioinformatics*, 20 (8): 1214-1221, 2004.

Due to a programming error, two substitution values in the derived distance metric must be changed in order to form a metric rather than just one as stated in the paper. As a result, the value of the R/W pair should be 4. The updated matrix is shown below.



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

After carefully reviewing the programs and repeating the related experiments, I have concluded that this change does not significantly affect the results presented in the original paper except for Figure 2 (page 1217). The related text (page 1218), “The mPAM alignments score the same as (26 cases) or better than (17 cases) PAM in 43 of 103 queries…,” should be changed to read, “The mPAM alignments score the same as (26 cases) or better than (18 cases) PAM in 44 of 103 queries….” Please see the updated version for details.

The matrix’s violation of the metric property was first noticed and brought to our attention by Shea Clayton, a graduate student from Georgia State University. We are communicating with the journal of Bioinformatics to make a formal correction.

I apologize for this mistake. In the meantime, please use the updated version of the paper at http://www.cs.utexas.edu/users/mobios/Publications/mPAM-0704.pdf.

BIOINFORMATICS

Vol. 20 no. 8 2004, pages 1214–1221
DOI: 10.1093/bioinformatics/bth065

A metric model of amino acid substitution

Weijia Xu* and Daniel P. Miranker

Department of Computer Sciences, The Center for Computational Biology and Bioinformatics, University of Texas, Austin, TX 78712, USA

Received on August 11, 2003; revised on November 7, 2003; accepted on November 24, 2003
Advance Access publication February 10, 2004

ABSTRACT

Motivation: We address the question of whether there exists an effective evolutionary model of amino-acid substitution that forms a metric-distance function. There is always a trade-off between speed and sensitivity among competing computational methods of determining sequence homology. A metric model of evolution is a prerequisite for the development of an entire class of fast sequence analysis algorithms that are both scalable, $O(\log n)$ and sensitive.

Results: We have reworked the mathematics of the point accepted mutation model (PAM) by calculating the expected time between accepted mutations *in lieu* of calculating log-odds probabilities. The resulting substitution matrix (mPAM) forms a metric. We validate the application of the mPAM evolutionary model for sequence homology by executing sequence queries from a controlled yeast protein homology search benchmark. We compare the accuracy of the results of mPAM and PAM similarity matrices as well as three prior metric models. The experiment shows that mPAM significantly outperforms the other three metrics and sufficiently approaches the sensitivity of PAM250 to make it applicable to the management of protein sequence databases.

Contact: xwj@cs.utexas.edu

INTRODUCTION

Computational methods of biological sequence analysis usually involve a trade-off between speed, scalability and sensitivity. By sensitivity, we mean the ability of an algorithm to identify similar sequences based on evolutionary criteria rather than simple mathematical constructions of strings of letters. Scalability refers to the rate of increase in execution time as a function of the amount of data being analyzed.

The fastest and most scalable homology algorithms, SST and BLAT, first compile a sequence database into a data structure that supports a fast nearest-neighbor search in a metric-space (Giladi *et al*., 2002; Kent, 2002). Due to their choice of distance metrics, these systems are also the least sensitive, and their concomitant applicability is limited. SST, due to Giladi *et al*. (2002), uses Hamming distance and a tree-based index structure to achieve $O[m\log(n)]$ scalability, where m is the length of the query sequence and n is the length or size of the database. Giladi *et al*. (2002) report an SST execution speed of 27 times the execution speed of BLAST2 for sequence assembly on databases of 120 000 nt and estimate a 200-fold speed-up over BLAST2 on megabasepair databases. BLAT, due to Kent (2002), is based on a simple edit distance supported by hashing to achieve $O(m)$ scalability using $O(n)$ memory. Kent (2002) reported BLAT execution speeds 40 times those of WU-TBLASTX.

DEFINITION. *A metric space is a set of objects with a binary distance function d, satisfying the following for every three objects x, y and z:*

(i)	$d(x,y) \geq 0$ and $d(x,y) = 0$ iff $x = y$;	(*Positivity*)
(ii)	$d(x,y) = d(y,x)$;	(*Symmetry*)
(iii)	$d(x,z) + d(y,z) \geq d(x,y)$.	(*Triangle Inequality*)

Algorithm designers have leveraged the triangle inequality to produce entire classes of data structures to speed up metric-space search (Chavez *et al*., 2001). Intuitively, the triangle inequality says that if two objects are similar to a third object, they cannot be too dissimilar to one another. Algorithmically, the triangle inequality allows subsets of very similar data to be organized and clustered. If a new data element is sufficiently dissimilar to a given element in a cluster, then similarity with the remaining elements in the cluster may be ruled out, without any additional similarity comparisons. One algorithmic class is equivalent to materializing a hierarchical clustering of a dataset off-line as a searchable tree-based data structure. These methods are similar in structure to the index trees intrinsic to the architecture of database management systems and tend toward the same scalable, $O(\log n)$, on-line search time[1] (Brin, 1995). SST is in this class; BLAT is an instance of a hash-based algorithm.

These approaches form a sharp contrast with the generality of a Smith–Waterman local-alignment algorithm which compares each possible pairing of sequence elements and

*To whom correspondence should be addressed.

[1] That is as the database grows in n, the computation required to locate related data grows in proportion to $\log n$. For example, in $\log_2$, if it takes 20 units of computation to locate a data object among one million objects, it will only take 30 units of computation to locate a data object among one billion objects.

Bioinformatics 20(8) © Oxford University Press 2004; all rights reserved.

assesses, by virtue of a weight matrix, the probability that one element may be replaced by another. A Smith–Waterman alignment requires $O(n * m)$ computation. BLAST, which can be viewed as an approximation to Smith–Waterman, begins with a linear scan of the database, $O(n)$, to construct the hot-spot index of exact matches of length k. Those exact matches are then extended to form the local alignment results. The expected-time complexity of BLAST is approximately $aW + bn + cnW/20^w$, where w is the number of words generated and n is the number of residues in the database (Altschul *et al.*, 1990).

Since the growth rate of sequence databases now exceeds the growth rate of processor speeds (Benson *et al.*, 2002), the consequent degradation in the performance of BLAST requires investigation of search methods that organize the database offline in order to speed on-line search. SST and BLAT embody this organization by storing overlapping k-mers (sequence fragments of length k) of sequences. Although SST and BLAT were initially developed to support sequence assembly, these methods can also be used in heuristic algorithms to deduce a local alignment from the returned k-mers (Gusfield, 1997). However, Hamming distance and simple edit distance are mathematically convenient sequence metrics that only minimally reflect evolutionary criteria. BLAT continues to be used to research problems that involve evolutionarily close sequences, e.g. comparison of the human and mouse genomes (Kent, 2002; Rouchka *et al.*, 2002; Cox *et al.*, 2002; Hedenfalk *et al.*, 2003). Further, these metric-distance functions (metrics) rely on the small alphabet size of nucleic acids to maintain computational feasibility, and thus these approaches have been limited to direct nucleotide to nucleotide comparisons.

If the similarity of a pair of k-mers of protein sequences could be qualified by evolutionary criteria that also formed a metric, it follows that a fully sensitive sequence homology search could be conducted without comparing all pairs of sequence elements. In other words, one can anticipate that the emergence of an evolutionary sequence metric will open research into SST- and BLAT-like algorithms that maintain existing speed and scalability while approaching the general sensitivity of the Smith–Waterman algorithm. This problem was first posed by Sellers (1974).

It is fair to ask, ‘Can the biology of evolution be modeled as a metric?’ The accepted mutation rate between pairs of amino acids is usually asymmetric. The stipulation of the triangle inequality is a significant restriction on possible mathematical models. The PAM family of amino-acid substitution matrices (Dayhoff *et al.*, 1978) has defied efforts to identify a simple, effective algebraic normalization to convert it into metric space (Taylor and Jones, 1993; Linial *et al.*, 1997).

We derive a metric amino-acid substitution matrix (hereafter mPAM) that reflects evolutionary bias by revisiting the mathematics used to derive the PAM matrices as well as the original data (Dayhoff *et al.*, 1978). Rather than being concerned with the frequency of substitutions, we compute the expected time between substitutions. An amino-acid pair with a high substitution rate should take less time to appear than a pair with a lower substitution rate. Thus, more similar sequences will score closer to zero, one of the requirements of a metric.

We validate mPAM by testing its accuracy using a controlled yeast sequence query benchmark in conjunction with Smith–Waterman alignment (Smith and Waterman, 1981). The benchmark comprises 103 sequence queries whose true positive hits have been identified by human experts (Schaffer *et al.*, 2001). We compare the accuracy of mPAM with that of the PAM matrices as well as three other metric matrices detailed below. Since nearly all sequence homology algorithms operate by dividing the database sequences and/or the query sequences into k-mers (Smith–Waterman being the notable exception), we evaluate the relative performance of the mPAM and PAM250 matrices on randomly generated sets of short sequences. The results indicate that mPAM, with metric space indexing algorithms, can be a general solution to the task of building a protein sequence database with $O(\log n)$ search performance.

Of the many efforts to develop amino-acid substitution matrices, we have identified two approaches that have resulted in metrics (Fitch, 1966; Taylor and Jones, 1993). The genetic-code matrix was derived by examining the differences in the nucleotide sequences of codons (Fitch, 1966). More precisely, the entry in the substitution matrix for any amino-acid pair is defined as the minimum edit-distance between their codons. Taylor and Jones (1993) propose and evaluate a variety of methods for projecting similarity matrices into metric space. They report their inter-row distance method applied to PAM250 as the most effective. Thus in our evaluation, we compare the genetic-code and inter-row matrices to PAM250 as well as simple-edit distance.

As a historic note, Needleman and Wunsch’s (1970) classic paper on formulating global alignment used simple-edit distance, which is a metric. Sellers (1974) then proved that if the substitution matrix of a set of characters forms a metric, then the weighted edit distance between sequences of those characters is also a metric. Waterman *et al.* (1976) extended Sellers results to include gaps.

Smith and Waterman (1981) then introduced their local-alignment algorithm and concomitantly the use of probability measures, *in lieu* of metrics, as the basis of substitution matrices. In what has become a *de facto* standard, Dayhoff *et al.* (1978) introduced log-odds statistics as the basis of the entries of substitution matrices. The PAM and BLOSUM log-odds matrices are in dominant use, with PAM matrices preferred when evolutionary criteria are involved (Gonnet and Benner, 1996; Henikoff and Henikoff, 1992).

If log-odds matrices, such as PAM matrices, are used to weight edit-distance, the result is not a metric. Log-odds reward more similar sequences with higher scores, an intuitively appealing result that reverses metric order; in a metric nearly identical objects must be close to 0 distance apart. Further, log-odds scoring matrices contain negative values violating metric positivity (i).

SYSTEMS AND METHODS

The point accepted mutation model

The PAM family of matrices details a Markovian model of evolution (Baldi *et al.*, 1994). The model was derived from the observation of 71 groups of closely related proteins. The protein sequences were aligned, and a phylogenetic tree, including putative ancestral sequences, was computed using maximum parsimony. The accepted mutations through paths in the tree were counted to reflect evolutionary substitution rates. Furthermore, the counts for the 1-PAM probability matrix were normalized to achieve a substitution rate of 1%. Multiplying the 1-PAM matrix by itself N times yields the N PAM probability matrix. This model has the following assumptions that resemble those in molecular clock theory:

(1) Amino acids mutate independently of each other.

(2) The probability of mutation depends only on the amino acid and the amount of evolution.

A value in the PAM probability matrix has the following meaning:

$$\begin{aligned} M_{ij} &= p(j \to i \mid j) \\ &= p\ (j \text{ mutated to } i \text{ within 250 PAM evolution} \\ &\qquad \text{distance} \mid \text{occurrence of } j). \end{aligned}$$

From the accepted mutation matrix, a relatedness odds matrix is defined as:

$$\begin{aligned} R_{ij} &= p[(j \to i \mid j) \mid i] = \frac{p(j \to i \mid j)}{f_i} = \frac{M_{ij}}{f_i} \\ &= p(j \text{ mutated to } i \text{ per occurrence of } i, j), \end{aligned}$$

where f_i is the observed frequency of the amino acid i.

The commonly used PAM250 similarity matrix is a log-odds matrix, P, derived from the relatedness odds matrix for 250 PAM distance.

$$P_{ij} = 10 \log R_{ij}.$$

In the corresponding log-odds matrix, each value has the following biological meanings:

(a) Each value P_{ij} corresponds to the log of the likelihood of how closely amino acids i and j are related compared with independent events. The log values make for an additive model.

(b) The value of P_{ii} varies for different amino acids. This value gives the likelihood that an amino acid remains unchanged over time. It corresponds to the varying mutability of different amino acids.

(c) A high positive value of P_{ij} corresponds to a high likelihood that amino acids j and i are related. Dissimilarity is quantified as a negative value. This scoring scheme makes it easier to pick up maximum local similarity because the value zero, which corresponds to random chance, acts as a cut-off score.

Matrix evaluation method

For validation purposes we compare the accuracy of the different matrices using Smith–Waterman local alignments. To do so, the metric matrices must be converted into similarity matrices. (Correctness of the Smith–Waterman algorithm requires the substitution matrix have a mix of positive and negative values.) The conversion method involves calculating, for each matrix, the median matrix element value and subtracting it from each matrix element, which we determined produced the best results. For mPAM, the resulting sign of the value provides similar meaning as the sign of the entry in a PAM matrix.

The test was conducted on a yeast protein database with 6433 protein sequences using a Linux machine (SUSE 8.0; dual AMDXP 1800+ processors with 2 GB memory). The query set contains 103 sequences whose true positive hits have been identified by human experts and whose curation is continually refined (Schaffer *et al.*, 2001). The yeast database and query set as well as the set of true positive hits were downloaded from ftp.ncbi.nlm.nih.gov/pub/impala/blastest in August 2002.

For the 103 queries, we computed Smith–Waterman local alignment scores using the following matrices: mPAM, PAM250, PAM70, identity matrix (simple-edit distance), genetic code matrix and inter-row distance matrix. For comparison, we also ran the same search on a stand-alone version of BLAST using the PAM250 and PAM70 matrices. Our Smith–Waterman implementation was coded in Java.

We used receiver-operating characteristic (ROC) scores to compare the accuracy of each matrix (Gribskov and Robinson, 1996). For each entry in the database, the local alignment score was calculated. The result is sorted by decreasing alignment score. The ROC_{50} value is computed by comparing the result list with the list of true positive hits. The ROC value has been computed as the following:

$$\mathrm{ROC}_n = \frac{1}{nT} \sum_{i=0}^{n} t_i, \qquad (1)$$

where t_i is the number of true positive hits ranked ahead of the i-th false positive, and T is the total number of true positives.

RESULTS

Matrix derivation results

In the PAM derivation, Dayhoff *et al.* (1978) normalized accepted mutation frequency data to form the concept of a

1-PAM distance. This suggests that there may be other normalizations of the data resulting in different models but with the same evolutionary bias. Our approach is to address this normalization with respect to time period rather than frequency. Since more frequent substitutions correspond to shorter time periods, more similar sequences would be scored with smaller values, consistent with a metric. Thus, to derive a metric substitution matrix, we address two sub-problems:

(a) how to define a symmetric mutability between a pair of amino acids;

(b) how to convert the probabilities into expected time with respect to metric properties.

The accepted mutation probability, M_{AB}, for amino acids A and B given by PAM is the observed probability that amino acid B could be substituted by amino acid A within a given evolutionary distance. This asymmetric matrix is derived from experimental data. For the first problem, mutual mutability of amino acids A and B corresponds to the relatedness between amino acid A and amino acid B. To establish an unbiased view, we assume that if there were no mutations, all amino acids would occur in nature with uniform probability (Wootton and Federhen, 1993). We define the mutual mutability of two amino acids as the probability that the two amino acids evolve to the same amino acid by accepted mutation within given evolutionary distance. Hence, the relatedness of any two amino acids is associated with not only the probability of direct substitution between them but also intervening substitutions. The probability that amino acid A and amino acid B could be the same through accepted mutation is defined as the summation of the probability that amino acid A and amino acid B could be mutated to the same amino acid in the same period of evolutional time (2). Hence, we define

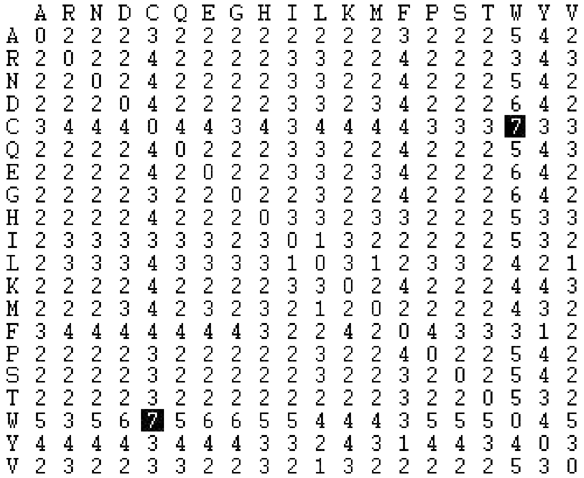
$$\begin{aligned} p(a,b) &= \sum_x [p(a \to x \mid a) * p(b \to x \mid b)] \\ &= \sum_x (M_{xa} * M_{xb}) \quad \text{if a} \neq \text{b} \\ p(a,b) &= 1 \quad \text{if a} = \text{b}, \end{aligned} \tag{2}$$

where x is any amino acid.

The assumption that any amino-acid mutation happens randomly at a constant rate makes its probability distribution random and memoryless, which is often modeled as an exponential distribution. The exponential probability distribution function is

$$F(t) = 1 - \mathrm{e}^{-\lambda t}, \tag{3}$$

where λ is the constant that corresponds to the accepted mutation rate. Equation (3) is a standard representation of the exponentional probability distribution. $F(t)$ is defined as the probability that an event, amino-acid substitution in this case, occurs within time interval t (Casella and Berger, 2002).



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

Fig. 1. The mPAM250 matrix: the related expected time (based on 250 PAM distance as one unit) for one amino acid to replace another, per Equation (8) and the PAM250 matrix. The illustrated result for C/W was the actual result decreased by 1 to maintain metric properties.

The probability density function is

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}(1 - \mathrm{e}^{-\lambda t}) = \lambda \mathrm{e}^{\lambda t}. \tag{4}$$

The expected time, t, for an event to happen is

$$E(t) = \int_0^\infty t * f(t)\,dt = \int_0^\infty t\lambda \mathrm{e}^{\lambda t}\,dt = \frac{1}{\lambda}. \tag{5}$$

Since the probability given by the PAM250 probability matrix is that for one amino acid to mutate to another amino acid at the same position based on 250 acceptable mutations per 100 amino acids, we define the time needed for 250 acceptable mutations per 100 amino acids as one mPAM time unit. It follows that the accepted mutation rate, λ, for each pair of amino acids a, b is given by

$$\lambda_{(a \to b)} = -\ln[1 - p(a,b)]. \tag{6}$$

The expected mean time between two successive events is

$$T_{(a \leftrightarrow b)} = \frac{1}{\lambda_{(a \leftrightarrow b)}} = -\frac{1}{\ln[1 - p(a,b)]}. \tag{7}$$

Thus, the elements of the distance matrix are calculated by the following:

$$\begin{aligned} D\ (a,b) &= -\frac{1}{\ln\left(1.0 - \sum[p(a \to x)p(b \to x)]\right)} && \text{if } a \neq b, \\ D\ (a,b) &= 0 && \text{if } a = b. \end{aligned} \tag{8}$$

The matrix in Figure 1, mPAM, is the set of normalized solutions for Equation (8) with respect to the PAM250 matrix, except for the pair of entries representing the solution for

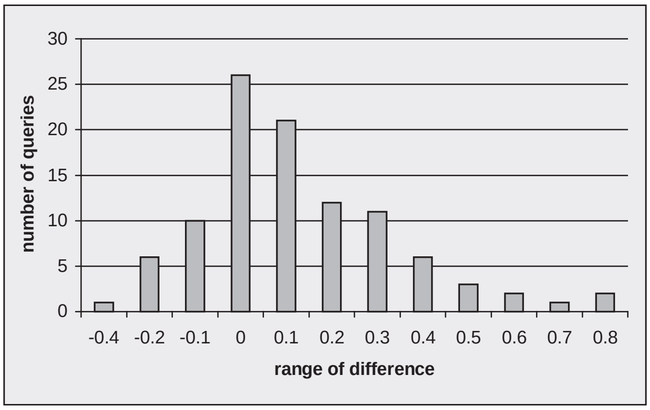


Fig. 2. The difference between ROC$_{50}$ using mPAM and PAM250. The x-dimension shows the range of difference. For example, the bar at 0.2 shows the number of queries on which the PAM250 has better ROC$_{50}$ than mPAM within 0.1–0.2 difference. Negative values indicate that mPAM has better performance than PAM250. The bar at 0 shows that there are 26 queries for which mPAM and PAM250 have the same ROC$_{50}$ value.

cysteine (C) and tryptophan (W). The solution for cysteine and tryptophan is 8, the largest distance between any two amino acids. However, this single value causes three of the 1140 possible triangle inequalities to fail. Decreasing this single distance to 7 results in a metric-matrix and maintains cysteine and tryptophan as the unique pair of most infrequently substituted amino acids. Hence, the alteration will not have any impact on the rank of distance between amino-acid pairs. We round to a single digit since this was the resolution of the results in the source data (Dayhoff *et al.*, 1978). In the Discussion section, we detail a conjecture per the general ability of this construction to produce metrics.

Matrix evaluation result

Figure 2 plots the comparison of ROC$_{50}$ scores for each query with respect to mPAM and PAM250. The mPAM ROC$_{50}$ score was subtracted from PAM ROC$_{50}$. A negative difference indicates a more accurate query result is returned by mPAM. The mPAM alignments score the same as (26 cases) or better than (17 cases) PAM in 43 out of 103 queries, or nearly 43% of the total number of queries. Among 60 cases where PAM250 outperformed mPAM, 21 queries score just 0.1 less accurate. A challenge in integrating the results across the queries is that the size of the true positive sets is different, preventing simple merging of the true positive sets. The size of the true positive sets ranges from 1 to 123 and averages 10. For a sequence query with a total of 10 true positive results, just one disagreement in the number of true positive hits ranked ahead of the first false positive result can induce a 0.1 difference in the ROC$_{50}$ score.

In order to assess the overall performance, we averaged the ROC plots among the 103 queries. Figure 3 is a plot of the results. We see that mPAM performs very close to PAM70 (Table 1) in net performance. The inter-row distance matrix is the best of the three prior metric substitution matrices. PAM250 is nearly the best PAMx matrix for this benchmark. Although mPAM has almost the same performance as PAM70, it is at a slight disadvantage compared with PAM250. However, by examining the knees of the curves, we see that the discrepancy between the inter-row matrix and mPAM is about 50% larger than the discrepancy between mPAM and PAM250.

In practice, many sequence homology search methods construct the database by breaking sequences into fragments of fixed length. Query sequences may also be broken into fixed-length fragments. For example, the fragment length (hot-spot) for BLAST defaults to 11 nt. The SST and BLAT packages were analyzed for fragment lengths 4–10 and 8–16, respectively (Altschul *et al.*, 1990; Giladi *et al.*, 2002; Kent, 2002). A search is conducted by matching query fragments with database fragments using global alignment. The matching fragments are chained together to form a complete local alignment. In these approaches, the quality of fragment matching is vital to the final result.

To test the quality of mPAM on fragmented representations, we randomly generated a set of sequence fragments 10 amino acids long. For each matrix, we computed all global alignments between pairs of fragments and then sorted them by score. We chose the rank list produced by PAM250 as the standard. The PAM250 list was compared with each other list, and the percentage of the same hits among the top t hits was computed (Fig. 4). The result from mPAM had good similarity to PAM250 at small distances and significantly outperformed the other metrics in general. These results suggest specifically that SST- and BLAT-like algorithms can be extended to protein sequences by replacing Hamming and edit distances with mPAM distance. These algorithms would also achieve much better sensitivity. Furthermore, when fragment matching uses exact or near-exact matching, short fragment lengths are required to retain good selectivity. With mPAM, it may be possible to use much longer fragment lengths, which will help improve speed. Sensitivity may be maintained by using a larger search radius in the fragment match (Giladi *et al.*, 2002; Myers, 1994). Detailing these effects will require effort commensurate with that put into analyzing and improving BLAST.

DISCUSSION

For the purpose of homology search, we have generated a metric amino-acid substitution matrix. To summarize, Table 1 contains the average ROC$_{50}$ scores for the results in Figure 3 and the ROC$_{50}$ score for BLAST. The disparity in ROC$_{50}$ scores between the mPAM results and the BLAST PAM250 score is just 0.05. These results are not directly comparable since BLAST uses the statistical significance (E-score) of a sequence match to prune the search in its chaining algorithm

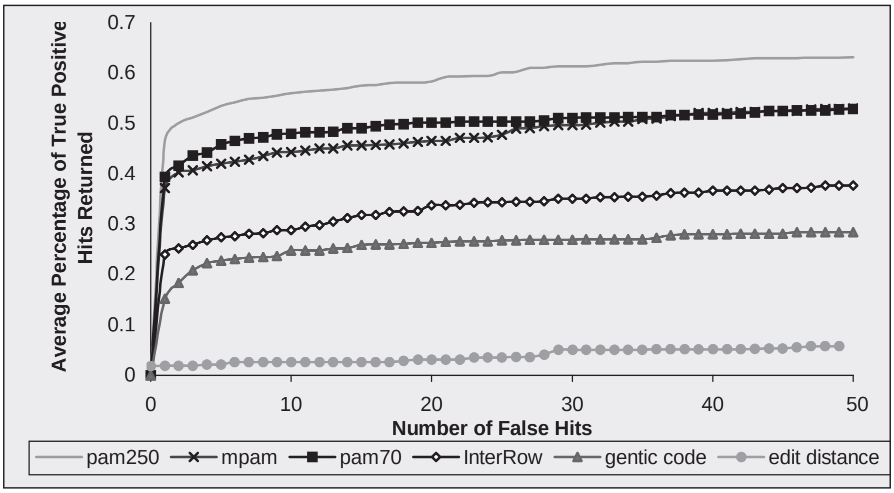


Fig. 3. Average fraction of true positive hits versus fraction of negative hits for PAM250, mPAM, PAM70, genetic-code matrix, inter-row matrix and identity matrix (edit distance) using local alignment.

Table 1. Comparison of average of ROC$_{50}$ value

Alignment method	Smith–Waterman local alignment						BLAST	
Matrix name	mPAM	PAM 250	PAM 70	Inter-row distance	Genetic code	Edit distance	PAM 250	PAM 70
Average ROC$_{50}$	0.48	0.59	0.50	0.33	0.26	0.04	0.53	0.42

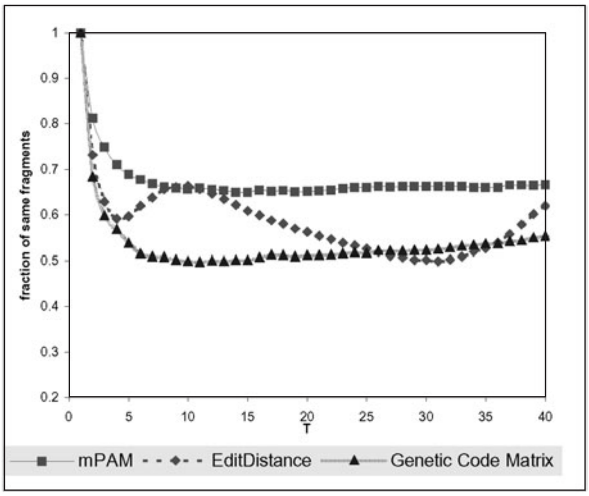


Fig. 4. Top t-score comparison among PAM250, mPAM, edit distance and genetic-code matrix. The plot shows the fraction of the same fragments for top t rank. A set of randomly generated fragments of length 10 was used.

and to rank the output. Assessing the significance of an alignment score has greatly increased the quality of homology search results (Collins and Coulson, 1990; Altschul and Gish, 1996; Altschul *et al.*, 2001). Using methodology similar to that of BLAST for chaining, we expect fragment based sequence analysis to maintain or even improve upon the accuracy we achieved using the Smith–Waterman algorithm. Consequently, one can now anticipate that sequence analysis methods like SST or BLAT may be extended and exhibit both their intrinsic speed and scalability and the generality and sensitivity of BLAST.

Our results also show that mPAM outperforms other metric matrices (Fig. 3 and Table 1). As expected, the simple-edit distance is insufficient because it does not contain any evolutionary information. Although the genetic-code matrix is derived from the genetic codon, this calculation entails only an information theoretic distance and does not take evolutionary criteria into account (Haig and Hurst, 1991). The mathematical projections from similarity-based matrices to distance metrics, as proposed by Taylor and Jones (1993), necessarily produce distance distortions. To solve the global classification of protein sequences, Linial *et al.* (1997) used a distance measure between two segments of 50 amino-acid residues from their similarity score $[d(u, v) = s(u, u) + s(v, v) - 2s(u, v)]$ with small metric distortion. We expect that mPAM can be used in protein classification to form a metric space directly.

We anticipate that the explicit calculation of an expected time between amino acid substitutions will draw the criticism connected to the debate of the molecular clock theory (Zuckerkandl and Pauling, 1962). A molecular clock was also assumed to normalize the 1-PAM matrix. In general, the simplifying assumptions in the derivation of the mPAM matrix are nearly identical to those made in the derivation of the PAM family. The most obvious weakness in mPAM with respect to a molecular clock is that the relative mutability for different amino acids is not preserved in mPAM as it is in PAM (all the diagonal elements of mPAM are zero). We suspect this is an important factor in the overall difference between mPAM and PAM.

In another parallel with early work on sequence evolution, the number of mismatched nucleotides was initially used as a distance measure. To compensate for the fact that the difference count slows as the time of divergence between two sequences increases, Jukes and Cantor (1969) developed a correction for the simple-edit distance model for nucleotide sequences. The Jukes–Cantor correction as it applies to peptides is

$$D = -\left(\frac{19}{20}\right) \ln\left[1 - \left(\frac{19}{20}\right) D\right] \qquad (9)$$

or simply

$$D = -\ln(1 - D). \qquad (10)$$

In our derivation of mPAM, we recognized the resemblance of Equation (8) to Equation (10). The Jukes–Cantor model assumes that all symbols share the same constant mutation rate. Instead of using one constant rate for all of the amino-acid pairs as in (10), we used a pair-specific value (6).

We expect that there are a number of refinements that may be made to the mPAM model beyond adjusting for the relative mutability of different amino acids. In particular, it may help to explain our original speculation on why the mPAM derivation would succeed. We considered Seller's theorem that if a character-weighting matrix forms a metric, then the corresponding weighted edit distance for sequences also forms a metric (Sellers, 1974). We observed that for PAM Dayhoff first aligned a set of sequences using maximum parsimony, a distance metric, and then computed the phylogenetic tree, also based on maximum parsimony. The evolutionary distances among the leaves of such a tree form an ultra-metric (Gusfield, 1997). Under this circumstance, we speculated that the converse of Seller's theorem might hold, that traversing the tree, counting amino-acid substitutions by sequence position, should form a metric and that integrating the counts across the length of the sequence would still be a metric. Our precise conjecture, which remains to be proven, is the following: given an ultra-metric over a set of aligned sequences, there exists a metric character weighting matrix such that the weighted edit distance among the sequences will approximate the ultra-metric. If this conjecture is proven constructively, then any model of phylogenetic tree construction for sequences that produces an ultra-metric may be used to define a metric substitution matrix. Nakhleh *et al*. (2002) have some recent results that would further generalize this to allow any phylogenetic tree construction, including an assessment of the deviation of the approximation from a molecular clock. Just as the PAM matrices spawned refinements and other substitution matrices, we expect the mPAM model to be the subject of further refinement.

The protein space is a very complex, high-dimension space. None of the known scoring models can accurately capture all of the complex relations within protein space. When a similarity matrix is converted into a distance matrix, some property loss is inevitable. Hence, the mPAM may not be as sensitive as the PAM or BLOSUM series matrices when applied to a certain set of sequences. Our intention of developing such a matrix is to effectively index protein sequences in metric space. Therefore, mPAM should primarily be used in building an index structure for overlapping k-mers of protein sequences. Such an index structure can accelerate the homology search by offering fast on-line range query for searching similar k-mers followed by a heuristic chaining algorithm to form the local alignment results. In this sense, mPAM is the most sensitive amino-acid substitution matrix with metric-space properties. Moreover, there are several other important issues in sequence analysis in addition to the scoring system, such as the algorithm used to find alignment and the statistical methods used to evaluate the significance of an alignment score. Those other factors might substantially affect the results of any scoring system. The integration of mPAM or a related metric into homology search algorithms is still the subject of research.

The doubling time of the sequence content of Genebank has shrunk from 18 months to 15 months, and its rate of growth continues to accelerate (Benson *et al*., 2002). Moore's constant for the doubling of processor speeds has been stable at around 18 months for over a decade (Patterson and Hennessy, 1996). This means that the volume of biological sequence data is growing faster than Moore's law, and it has now reached a rate of growth that ensures a widening gulf between computer capacity and biological computing requirements. As a result, metric-space indexing may be the only solution to manage gigabytes of biological sequence data.

ACKNOWLEDGEMENTS

We acknowledge Jacob Sarvela and Rui Mao, students in the laboratory who have participated in many discussions that contributed to this paper. We thank Professors Andrew Miranker and Edward Marcotte for their energetic and ongoing encouragement in the pursuit of this research. This research was supported in part by grants from the Texas Higher Education Coordinating Board and NSF contract no. 0241180.

REFERENCES

Altschul,S.F., Bundschuh,R., Olsen,R. and Hwa,T. (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.*, **29**, 351–361.

Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.

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

Baldi,P., Chauvin,Y., Hunkapiller,T. and McClure,M.A. (1994) Hidden markov models of biological primary sequence information. *Proc. Natl Acad. Sci., USA*, **91**, 1059–1063.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **20**, 17–20.

Brin,S. (1995) Near neighbor search in large metric spaces. *Proceedings of the 21st Conference on Very Large Database (VLDB'95)*, September 11–15, Zurich, Switzerland, pp. 574–584.

Casella,G. and Berger,R.L. (2002) *Statistical Inference*, 2nd edn. Duxbury Press, Pacific Grove, CA.

Chavez,E., Navarro,G., Baeza-Yates,R. and Marroquin,J.L. (2001) Searching in metric spaces. *ACM Comput. Surv.*, **33**, 273–321.

Collins,J.F. and Coulson,A.F. (1990) Significance of protein sequence similarities. *Methods Enzymol.*, **183**, 474.

Cox,L.A., Birnbaum,S. and VandeBerg,J.L. (2002) Identification of candidate genes regulating HDL cholesterol using a chromosomal region expression array. *Genome Res.*, **12**, 1693–1702.

Dayhoff,M.O., Schwartz,R. and Orcutt,B.C. (1978) Atlas of protein sequence and structure. **5** (Suppl. 3), 345–358.

Fitch,W.M. (1966) An improved method of testing for evolutionary homology. *J. Mol. Biol.*, **16**, 9–16.

Giladi,E., Walker,G.M., Wang,J.Z. and Volkmuth,W. (2002) SST: an algorithm for finding near-exact sequence matches in time proportional to the logarithm of the database size. *Bioinformatics*, **18**, 873–879.

Gonnet,G.H. and Benner,S.A. (1996) Probabilistic ancestral sequences and multiple alignments. *5th Scandinavian Workshop on Algorithm Theory, Springer vol. 1097 of Lecture Notes in Computer Science*, pp. 380–391.

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.

Gusfield,D. (1997) *Algorithms on Strings, Trees and Sequences Computer Science and Computational Biology*. Press Syndicate of the University of Cambridge, USA, pp. 449–454.

Haig,D. and Hurst,L.D. (1991) Quantitative measure of error minimization in the genetic code. *J. Mol. Evol.*, **33**, 412–417.

Hedenfalk,I., Ringner,M., Ben-Dor,A., Yakhini,Z., Chen,Y., Chebil,G., Ach,R., Loman,N., Olsson,H., Meltzer,P., Borg,A. and Trent,J. (2003) Molecular classification of familial non-BRCA1/BRCA2 breast cancer. *Proc. Natl Acad. Sci., USA*, **100**, 2532–2537.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Jukes,T.H. and Cantor,C. (1969) Evolution of protein molecules. In Munro,H. (ed.), *Mammalian Protein Metabolism*. Academic Press, USA, pp. 21–132.

Kent,W.J. (2002) BLAT—the BLAST like alignment tool. *Genome Res.*, **12**, 656–664.

Linial,M., Linial,N., Tishby,N. and Yona,G. (1997) Global self organization of all known protein sequences reveals inherent biological signatures. *J. Mol. Biol.*, **268**, 539–556.

Myers,E.W. (1994) A sublinear algorithm for approximate keyword searching. *Algorithmica*, **12**, 345–374.

Nakhleh,L., Roshan,U., Vawter,L. and Warnow,T., (2002) Estimating the deviation from a molecular clock. *Proceedings of the Second International Workshop on Algorithms in Bioinformatics (WABI). Lecture Notes in Computer Science (LNCS #2452)*, pp. 287–299.

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

Patterson,D.A. and Hennessy,J.L. (1996) *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann Publishers Inc., San Francisco.

Rouchka,E.C., Gish,W. and States,D.J. (2002) Comparison of whole genome assemblies of the human genome. *Nucleic Acids Res.*, **30**, 5004–5014.

Schaffer,A.A., Aravin,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.

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

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

Taylor,W.R. and Jones,D.T. (1993) Deriving an amino acid matrix. *J. Theor. Biol.*, **164**, 65–83.

Waterman,M.S., Smith,T.F. and Beyer,W.A. (1976) Some biological sequence metrics. *Adv. Math.*, **20**, 367–387.

Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.

Zuckerkandl,E. and Pauling,L. (1962) Molecular disease, evolution and genetic heterogeneity. In Kasha,M. and Pullman,B. (eds), *Horizons in Biochemistry*. Academic Press, New York, pp. 189–225.