# A Bayesian Evolutionary Distance for Parametrically Aligned Sequences

Pankaj Agarwal* and David J. States
Institute for Biomedical Computing
Washington University, Box 8036
700 South Euclid Avenue
St. Louis, MO 63110
{agarwal,states}@ibc.wustl.edu
Voice: (314) 362-2134 Fax: (314) 362-0234

April 16, 1996

Running title: Bayesian Evolutionary Distance.
Keywords: Evolutionary Distance, Pairwise Sequence Alignment, Parametric Alignment, DNA Sequence, Bayes rule.

**Journal of Computational Biology, 3 (1), 1–17, 1996**
**Please check journal for final copy**

**Abstract**

There is an inherent relationship between the process of pairwise sequence alignment and the estimation of evolutionary distance. This relationship is explored and made explicit. Assuming an evolutionary model and given a specific pattern of observed base mismatches, the relative probabilities of evolution at each evolutionary distance are computed using a Bayesian framework. The mean or the median of this probability distribution provides a robust estimate of the central value. The evolutionary distance has traditionally been computed as zero for an observed homology of 20 bases with no mismatches; we prove that it is highly probable that the distance is greater than 0.01. The mean of the distribution is 0.047, which is a better estimate of the evolutionary distance.

Bayesian estimates of the evolutionary distance incorporate arbitrary prior information about variable mutation rates both over time and along sequence position, thus requiring only a weak form of the molecular-clock hypothesis.

The endpoints of the similarity between genomic DNA sequences are often ambiguous. The probability of evolution at each evolutionary distance can be estimated over the entire set of alignments by choosing the best alignment at each distance and the corresponding probability of duplication at that evolutionary distance. A central value of this distribution provides a robust evolutionary distance estimate. We provide an efficient algorithm for computing the parametric alignment, considering evolutionary distance as the only parameter.

These techniques and estimates are used to infer the duplication history of the genomic sequence in *C. elegans* and in *S. cerevisae*. Our results indicate that repeats discovered using a single scoring matrix show a considerable bias in subsequent evolutionary distance estimates.

# 1 Introduction

Biological sequences evolve by complex processes. The frequency of observed substitutions has been used to estimate the number of mutations and the elapsed time since the divergence of the two sequences (Zuckerkandl and Pauling, 1965; Fitch and Margoliash, 1967). However, the relationship between the time of divergence of two genes and the number of accepted mutations per site is not linear. Furthermore, the extent of the region of sequence similarity is interrelated with the estimate of its evolutionary distance. These relationships are explored and made explicit in a Bayesian framework capable of representing mutation rates that vary with both time and site.

A number of methods have been proposed for estimating evolutionary distance in nucleotide sequences (Jukes and Cantor, 1969; Kimura, 1980; Tajima and Nei, 1984)[1]. They vary mainly in the number of different nucleotide pairs considered; the simplest model only counts the number of matches and mismatches, while a more complex model distinguishes between transitions and transversions, and others account for GC content (relative base frequencies).

The *Point Accepted Mutation (PAM)* model is commonly employed for protein evolution (Dayhoff et al., 1979). It provides a series of scoring matrices (PAM1, PAM2, ..., PAM500) optimized to find protein homology at that specific evolutionary distance. PAM120 provides maximum sensitivity for identifying alignments with evolutionary distance 1.20. In general, PAMn is best suited for identifying alignments with evolutionary distance $n/100$. Using a Markov mutational model, similar matrices can also be derived for nucleotides (Fitch and Margoliash, 1967; States et al., 1991). $M_1$ is an example of a mutational probability matrix for nucleotides, and the corresponding scoring matrix (PAM1) is $S_1$ with the scores in bit units (assuming uniform base composition).

$$M_1 = \begin{array}{c|cccc} & a & c & g & t \\ \hline a & 0.99 & 0.002 & 0.006 & 0.002 \\ c & 0.002 & 0.99 & 0.002 & 0.006 \\ g & 0.006 & 0.002 & 0.99 & 0.002 \\ t & 0.002 & 0.006 & 0.002 & 0.99 \end{array} \qquad S_1 = \begin{array}{c|cccc} & a & c & g & t \\ \hline a & 1.99 & -6.97 & -5.38 & -6.97 \\ c & -6.97 & 1.99 & -6.97 & -5.38 \\ g & -5.38 & -6.97 & 1.99 & -6.97 \\ t & -6.97 & -5.38 & -6.97 & 1.99 \end{array}$$

The diagonal probabilities for the one point accepted mutation probability matrix ($M_1$) are 0.99. In this example, the probability of a transition (a↔g and c↔t = 0.006) is 3 times that of a transversion (a↔c, a↔t, c↔g, and g↔t = 0.002). The element in the $i^{th}$ row and $j^{th}$ column of this matrix $M_{nij}$ provide the probability of the nucleotide specified in the $i^{th}$ row being substituted by the one specified in the $j^{th}$ column. These probability matrices are converted to symmetrical log odds score matrices, *the PAM matrices for nucleotides*, with scores $S_{ij}$'s. The score for aligning base $i$ with base $j$ at $n$ PAM's is $S_{nij}$, and

$$S_{nij} + S_{nji} = \log_2 \frac{p_i M_{nij}}{p_i p_j} + \log_2 \frac{p_j M_{nji}}{p_j p_i}$$

$p_i$ is the probability of occurrence of base $i$. For the *C. elegans* sequence ($p_A = p_T = 0.32$ and $p_C = p_G = 0.18$). The scores are made symmetrical, because it is normally not possible to infer the direction of evolution.

$$S_{nij} = S_{nji} = \frac{1}{2} \log_2 \frac{M_{nij} M_{nji}}{p_i p_j}$$

[1] See Gojobori et al.(1990) and Zharkikh (1994) for reviews.

```
aa
||
aa
```
Alignment A

```
aaa
|||
aaa
```
Alignment B

| Alignment | Probability of evolution | Probability of chance observation | Relatedness Odds | Score (bits) | Relatedness prob. density |
|---|---|---|---|---|---|
| A | $(0.99)^2 = 0.98$ | $(0.25)^2 = 0.0625$ | $(\frac{0.99}{0.25})^2 = 15.7$ | $\log_2 15.7 = 3.97$ | $\frac{15.7}{15.7+62.1} = 0.2$ |
| B | $(0.99)^3 = 0.97$ | $(0.25)^3 = 0.0156$ | $(\frac{0.99}{0.25})^3 = 62.1$ | $log_2 62.1 = 5.96$ | $\frac{62.1}{15.7+62.1} = 0.8$ |

Table 1: Computing the probability density of two sequence being related due to evolution at PAM 1. At PAM 1 there is 0.99 probability that a nucleotide remains unchanged. The probability of chance observation does not depend upon the PAM distance. This calculation considers two alternative hypothesis: that either "aa" and "aa" are related or "aaa" and "aaa" are related. Not surprisingly, the probability density favors the longer alignment (i.e. the one with 3 identities).

The PAM matrices provide explicit probabilities of a sequence evolving from another sequence at a particular evolutionary distance. These matrices are easy to compute, and can accommodate arbitrarily complex mutational models accounting for 12 independent mutational rates (6 if symmetry is assumed, because it is usually not possible to infer the direction of evolution), along with arbitrary initial base compositions (3 parameters)[2]. Thus, these matrices can account for up to 15 different parameters[3]. Given all these parameters, closed-form solutions for distance estimates are difficult. However, using the PAM matrices, it is possible to numerically compute the score, as well as the relative probability of the sequences being related due to evolution, at each distance.

Most methods use the number of the scoring matrix that maximizes the alignment score to estimate the evolutionary distance between two sequences. This corresponds to the mode of the probability distribution of the evolutionary distance. We propose using either the mean or the median as a more robust estimate of the actual evolutionary distance.

The probability of two sequences being related due to evolution with a specified alignment can be related to the score of that alignment. This is illustrated by considering two simple alignments.

The alignments in table 1 provide two competing hypothesis that either "aa" is related to "aa" or "aaa" is related to "aaa", i.e. an alignment with 2 identities or an alignment with 3 identities. In this simple case, clearly the alignment with the 3 identities is more significant. However, some of the terminology and issues are clarified by this example, and the probability density quantifies the notion that the longer alignment is more significant. Even though the probability of evolution of "aa' to "aa" is higher, the probability of "aaa" and "aaa" being *related* due to evolution is higher. This distinction between the probability of evolution and the probability density of two sequence being related due to evolution forms the basis of the PAM matrices (Dayhoff et al., 1979). The score for a pair amino acids is computed as the log of the probability of the pair being related due to evolution.

The score of an alignment can be computed in standard units (for example, in bits) (Altschul,

[2]The rows should sum to 1; thus there are 12 independent rates, not 16. The final base compositions are completely determined by the initial compositions and the PAM matrix.

[3]Often, for inferring homology between portions of the same genome, 2 parameters are enough: the transition/transversion ratio and the GC content. This assumes that the genomes have stable GC content.

1991). The score is computed from the log of the relatedness odds, because the log is taken to base 2, the score units are bits. The probability density of two sequences being related is proportional to the exponential of its score. Each extra bit of score implies that the log odd probability of the two sequence being related is twice as much. An alignment with a score of $k$ extra bits has $2^k$ times more information, and is $2^k$ times more likely to be due to evolution (under the model).

It is possible to compute the best alignment at each evolutionary distance and use the score of the alignment to compute the relative probability of evolution at that distance. Figure 1 displays a sketch of the probability distribution associated with the sequence having been duplicated at various distances.

In section 2, we explore these probability distributions for evolutionary distances, along with a Bayesian framework that provides robust distance estimates, especially for short sequence homologies.

Current evolutionary distance estimates rely on a given fixed alignment. Although the endpoints of an alignment for homologous gene sequences can often be precisely determined (corresponding to the endpoints of the gene), determining the precise endpoints for DNA sequences is often impossible, because there are several competing hypotheses regarding the extent of the duplication. We propose an estimate for evolutionary distance that considers the probability of duplication computed from the score of the best alignment at each distance. The estimate for the evolutionary distance should depend upon the confidence in the alignment, and in section 3 we suggest a solution.

Homology is inferred by sequence comparison, which involves computing the probability of evolution of the two sequences from a common ancestral sequence, given a particular evolution model (Altschul et al., 1994). In most cases, only a single matrix (for example, PAM120 for amino acids) is used to compute this probability. It has been suggested that the chances of detecting homology could be improved by scoring with a few different matrices for amino acids (for example, PAM $5, 30, 70, 120, 180,$ and $250$) (Altschul, 1993). In section 4, we extend this idea to detecting homology in nucleotide sequences. We show that for nucleotide comparisons it is unnecessary to select some matrices. It is possible to score ungapped alignments at all evolutionary distances with only a small computational overhead, for a class of matrices that we term *well-decaying*.[4]

The three techniques of using the Bayesian evolutionary distance, considering alternative alignments at various PAM distances, and computing the scores efficiently for a range of PAM distances all tie in together in estimating the duplication activity in a 3.66 Mb contiguous sequence from *C. elegans* and in three chromosomes from *S. cerevisae*. The experimental results are discussed in section 5.

## 2 Bayesian Evolutionary Distance

Unbiased estimates of the evolutionary distance (Tajima, 1993) are optimal in the asymptotic sense that given $S$ homologies with evolutionary distance $d$, the average of the estimated distance tends to $d$, as $S$ tends to infinity. We suggest using Bayesian estimates instead of the current maximum likelihood estimates. These estimates of evolutionary distance are asymptotically identical for long sequences (with the same choice of parameters), but the Bayesian estimates for short sequences are larger.

[4] We also show that for some other frequently used *non* well-decaying matrices, the errors in evolutionary distance and score estimates are negligible in assuming the matrices to be well-decaying.

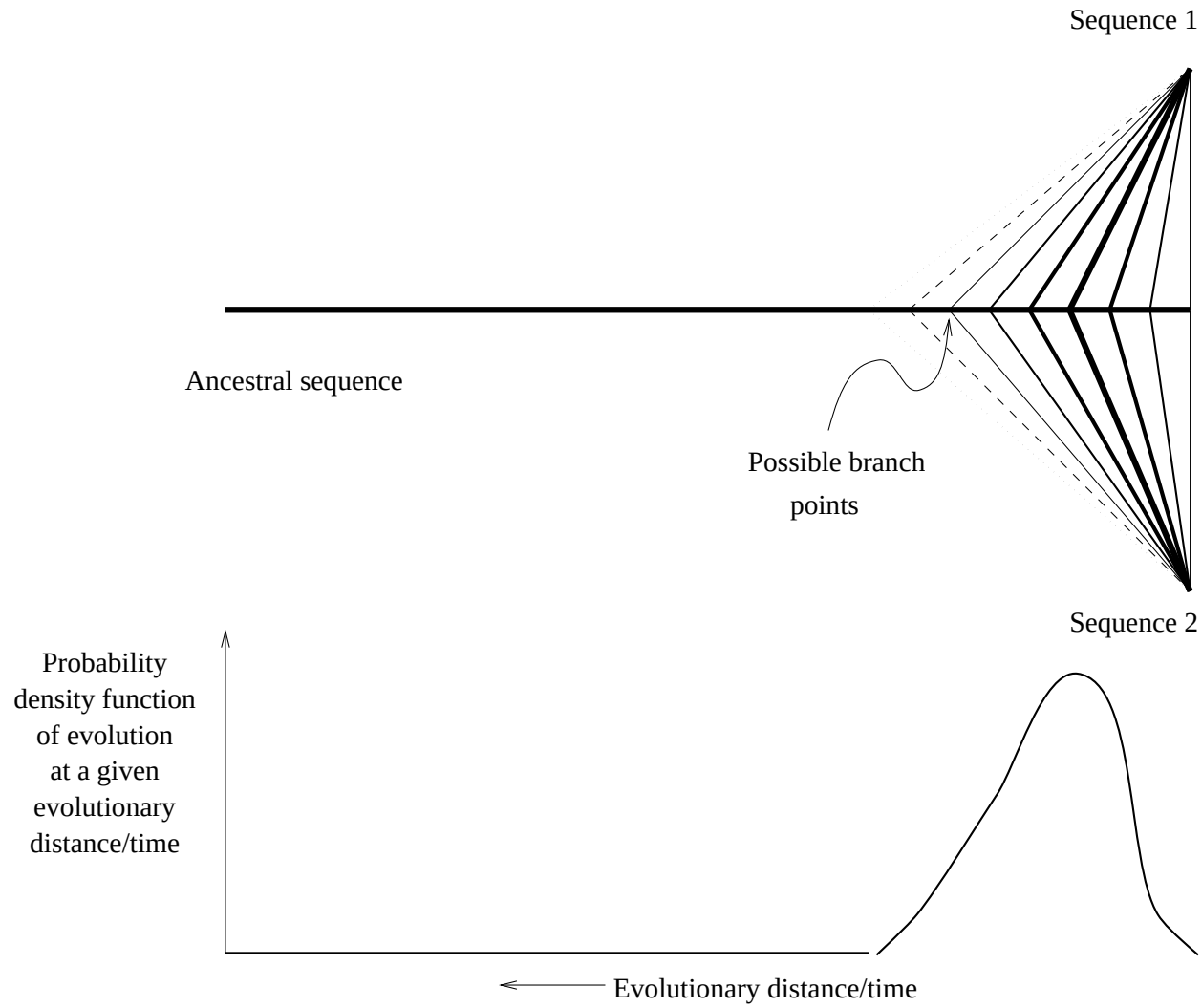



Figure 1: Cartoon of the associated probability distribution of evolution at various distances/time periods. The thickness of the various branching lines corresponds roughly to the probability of divergence at that evolutionary distance/time. It is difficult to ascertain a specific branch point; instead a probability density is associated with the branch point being at a given distance/time.

Given a homology of length $n$, we can evaluate the probability of observing $k$ mismatches at evolutionary distance[5] $x = 0.01, 0.02, \ldots M$. Let $P(k|x)$ ($\sum_{k=0}^{n} P(k|x) = 1$) be the probability of observing $k$ mismatches given that evolutionary distance is $x$, and let $P(x|k)$ be the probability of evolutionary distance being $x$ given that we observed $k$ mismatches. From conditional probabilities or Bayes rule, we get:

$$P(x|k) = \frac{P(k|x)P(x)}{P(k)}$$

We assume a uniform *a priori* distribution of the evolution distance, i.e. the homology is equally probable at each distance (in the set of distances $D = 0.01, 0.02, \ldots M$). As $\sum_{x \in D} P(x) = 1 \Rightarrow P(x) = \frac{1}{100M}\ \forall x$ and given that:

$$P(k) = \sum_{x \in D} P(k|x)P(x)$$

$$\Rightarrow P(k) = \frac{1}{100M} \sum_{x \in D} P(k|x)$$

$$\Rightarrow P(x|k) = \frac{P(k|x)}{\sum_{x \in D} P(k|x)} \tag{1}$$

The mean and the median of the $P(x|k)$ distribution provide robust estimates of the central value.

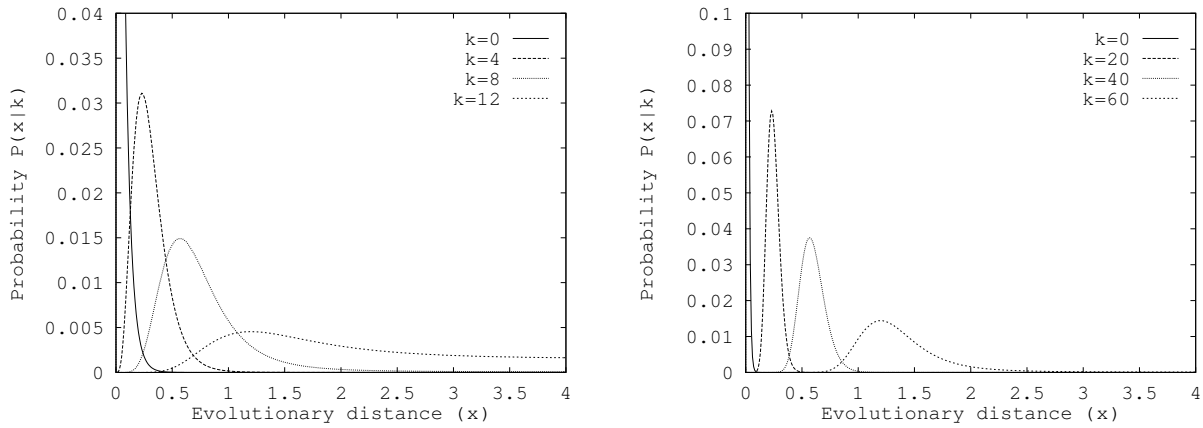
$$d_{mean}(k, n) = E(x|k) = \sum_{x \in D} x P(x|k)$$

Figure 2 plots the probability $P(x|k)$ of the evolutionary distance being $x$ given that $k$ mismatches are observed. These plots assume that the four nucleotides are equally likely in the sequence and that all mutations are equally probable; however, using the probabilities derived from the PAM matrices, similar plots can be made for any mutation and nucleotide frequencies. The PAM distances are assumed to be between 1 and 400 (corresponding to evolutionary distance 0.01 and 4.00). The upper PAM limit arises from two considerations: theoretically, it may go back to the existence of life on earth, but practically it is limited by the evolutionary relationships that can be discovered using sequence alignment techniques. The practical limit is upper bounded by PAM400.

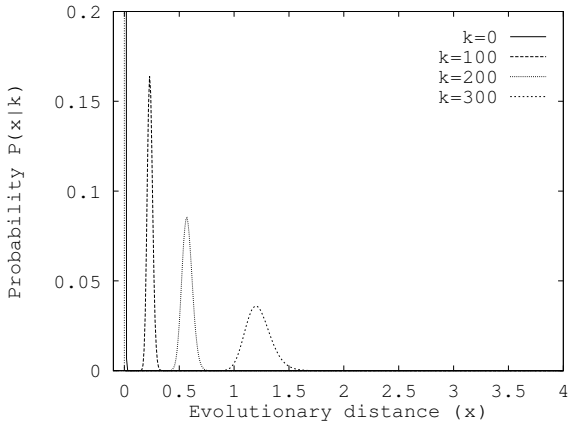
Notice that even though the plots in figure 2 are unimodal for $k < 3n/4$, they are not symmetrical about the mode; consequently, the mean is distinct from the mode. Furthermore, the distributions for the longer sequences have lower variances, thus providing more reliable distance estimates.

For $k \geq 3n/4$, the alignments are *not* distinguishable from random alignments, and any evolutionary distance estimate based on such an alignment is questionable. The probability of evolution

[5] It is possible to consider any specific discretization of the evolutionary distance. In this paper, evolutionary distance increase in steps of 0.01 corresponding to 1 PAM.




$n = 20$ $n = 100$




$n = 500$

Figure 2: $P(x|k)$ plotted as a function of $x$ for various values of $k$ for sequences of various lengths. $x$ was assumed to be between 0.01 and 4.0. The area under each curve is 1.

| | n= 20 | | | n = 100 | | | n = 500 | | |
|---|---|---|---|---|---|---|---|---|---|
| k/n | $d_{mean}$ | $d_{Ta}$ | $d_{JC}$ | $d_{mean}$ | $d_{Ta}$ | $d_{JC}$ | $d_{mean}$ | $d_{Ta}$ | $d_{JC}$ |
| 0.00 | $0.047 \pm 0.053$ | 0 | 0.00 | $0.006 \pm 0.010$ | 0 | 0.00 | $0.000 \pm 0.001$ | 0 | 0.00 |
| 0.20 | $0.311 \pm 0.156$ | 0.22 | 0.23 | $0.247 \pm 0.057$ | 0.23 | 0.23 | $0.235 \pm 0.025$ | 0.23 | 0.23 |
| 0.40 | $0.792 \pm 0.459$ | 0.54 | 0.57 | $0.599 \pm 0.112$ | 0.56 | 0.57 | $0.577 \pm 0.047$ | 0.57 | 0.57 |
| 0.60 | $2.011 \pm 0.925$ | 1.07 | 1.21 | $1.392 \pm 0.417$ | 1.17 | 1.21 | $1.230 \pm 0.116$ | 1.2 | 1.21 |
| 0.80 | $2.797 \pm 0.762$ | 2.51 | - | $3.187 \pm 0.551$ | 4.59 | - | $3.554 \pm 0.336$ | 29.8 | - |
| 1.00 | $3.098 \pm 0.618$ | 60 | - | $3.519 \pm 0.365$ | $10^{11}$ | - | $3.812 \pm 0.164$ | $10^{60}$ | - |

Table 2: The various distance estimates as a result of observing k mismatches in an alignment of length $n$. The true distance is unknown. $d_{mean}$: the expectation of the distribution of $P(x|k)$ versus $x$, the standard deviation is derived from the same distribution; $d_{JC}$: Jukes-Cantor distance; $d_{Ta}$: Tajima's unbiased estimate is equal to the mode of the distribution of $P(x|k)$ versus $x$.

between the "homologous" sequences increases monotonically with distance; thus the mean, mode, and median of the true distribution are all infinite.

The mean distance ($d_{mean}$) depends upon the sample size (or the length of the homologous sequence). Therefore, an observation of zero mismatches in a homology of size 20 bases is best characterized by distance estimate 0.05, while an observation of zero mismatches in a homology of 100 bases provides a distance estimate of 0.006 (table 2). Not observing a substitution in a homology of 20 bases is insufficient evidence to conclude that the evolutionary distance is zero, but for a large homology of 500 bases with zero mismatches, the probability of the evolutionary distance being significantly greater than zero is infinitesimal. This is a significant difference between Bayesian and other prevalent evolutionary distance estimates.

The Jukes-Cantor ($d_{JC} = -\frac{3}{4}\log(1 - \frac{4}{3}\frac{k}{n})$) measure is computed using the fraction of observed mismatches ($k/n$). Thus, the estimated distance is the same for two pairs of sequences with 10% mismatch, irrespective of the lengths of the sequences ($n$). Moreover, as $n \to \infty$, $d_{JC} \to d_{mean}$.

Tajima (1993) has suggested using an unbiased estimate for the distance ($d_{Ta}$), which corrects some problems caused by the use of logarithms in the Jukes-Cantor distance. The Tajima estimate corresponds to the mode of the distribution in equation 1. It varies with both the observed fraction of mismatches $\frac{k}{n}$ and $n$. However, for constant $\frac{k}{n}$, increasing $n$ increases $d_{Ta}$, which is contrary to the relationship between $n$ and $d_{mean}$.

$$d_{Ta} = \sum_{i=1}^{k} \frac{1}{i}(\frac{4}{3})^{i-1} \frac{k!}{(k-i)!} \frac{(n-i)!}{n!}$$

Table 2 displays the various distance estimates as a result of observing k mismatches in an alignment of length n. For values of $k < 3n/4$, $d_{mean} \geq d_{Ta} \geq d_{JC}$. For $k \geq 3n/4$, $d_{JC}$ cannot be computed, and the true $d_{mean}$ is infinity; however, artificially limiting the PAM distances to be less than 400 provides finite estimates for $d_{mean}$. In any case, the utility of computing distance estimates for alignments with greater than 75% mismatches is dubious.

Other distances, as described by Kimura (1980) and Tajima and Nei (1984), account for different sets of parameters. These depend mainly upon the observed relative frequencies of the various substitutions, but do *not* consider the sample size $n$. There are also measures of similarity (and

evolutionary distance) of sequences, based on $k$-tuple composition, that do not require sequence alignment (Blaisdell, 1986).

## Variable rates of substitution

The molecular-clock hypothesis has been frequently criticized for its dependence on a uniform mutation model (Wilbur, 1985). The Bayesian method can be readily extended to account for variable rates of substitution at different sites or at different time periods. Altschul (1991) has observed that all scoring matrices can be normalized to provide scores in bits. This technique is essential to combining scores from different matrices. The site for each base can be scored with a different scoring matrix, and the probability of the aligned pair of bases being related due to evolution at any evolutionary distance can be computed. The probability of two sequences being related due to evolution is the product of the probabilities for each base assuming that the sites evolve independently. (This is similar to the scoring used for sequence alignments.) These relative probabilities of evolution at each evolutionary distance can then be used to compute either the mean or the median evolutionary distance. Thus, this method can be employed to incorporate higher mutation rates at the silent codon sites and non-transcribed regions; furthermore, it can amalgamate information from genes with different mutation rates.[6]

Information about variable mutation rates over time may also be incorporated into the scoring matrices. This is achieved by remapping the matrices to the evolutionary distances. We illustrate this by a simple example. Assume that the normal mutation rates are such that the PAM1 is about a million years ago (in other words, the mutation rate is $10^{-8}$ per base per year). Let us also assume that the mutation rates were twice as high between 50 and 60 million years ago. Thus, for time period $n \leq 50$ million years, the corresponding matrix is PAMn, and the distance is $n/100$, but for $50 < n \leq 60$ the corresponding scoring matrix is PAM $(n + (n - 50))$, and the distance is $(2n - 50)/100$. Therefore, information about variable mutation rates over time is easily incorporated by renumbering the PAM matrices. The various rates of evolution at different sites and time periods are taken as priors and not estimated from the limited data set.

This provides a technique for an overall estimation of the evolutionary distance between large sections of genomes (or even entire genomes) where different regions of the genome might have had different rates of evolution.

## Prior probability

Prior information about the probability of the evolution for certain time periods can be directly included as the prior probability in the Bayesian computation (which is assumed to be uniform in the absence of any other evidence). An example of a sources of this prior information is paleontology; another one is punctuated evolution with high probability of evolution associated with the speciation time periods. In addition, knowledge about the time period of divergence may be used to limit the range of considered PAM's.

[6]Other techniques for considering variable rates of mutation over sites have also been proposed (Yang, 1994).

## 3 Uncertainty in Sequence Alignments

In general, the precise evolutionary origins of sequences are not known, and there are multiple hypotheses regarding the possible alignment of two sequences. Alignments frequently vary with the choice of the parameters. Parametric sequence alignment is a technique for efficiently discovering the highest scoring alignment over a range of values for a set of parameters. The most common parameters are the match-mismatch scores and gap opening-continuing penalties (Fitch and Smith, 1983; Waterman et al., 1992; Waterman, 1994; Vingron and Waterman, 1994; Tillier, 1994; Gusfield et al., 1994). The regions having identical scores in this alignment space are convex polygons and can be discovered efficiently (in constant time per region). However, the number of regions is $O(n^3)$ for local alignments on sequences of length $n$. The maximum likelihood alignment of DNA sequences over a range of parameters has also been considered (Bishop and Thompson, 1986; Thorne et al., 1991; Thorne et al., 1992; Allison et al., 1992). In addition to providing the optimal alignment, they provide estimates of the evolutionary parameters. The maximum likelihood alignment requires searching the multi-dimensional likelihood surface for the maximum, which is an expensive numerical computation.

Analysis of suboptimal alignments has revealed that often homologous sequences have numerous alignments with scores close to the maximum score, making it impossible to determine the *true* alignment by considering only the scores (Saqi and Sternberg, 1991; Zuker, 1991).

Most current evolutionary distance estimates rely on a *true* alignment. In the absence of prior knowledge about which alignment is correct, we consider the relative probability of each alignment being correct. These probabilities may be utilized to estimate a mean for a number of parameters, including the evolutionary distance and the number of substitutions of a specific type. In particular, the mean evolutionary distance is a weighted sum of the evolutionary distances. The weights correspond to the probability of the best alignment at each distance. Instead of considering only the best alignment at each distance, one could consider all the possible suboptimal alignments at each distance. However, the advantage of the additional computation (because there are an exponential number of gapped alignments and a cubic number of ungapped alignments) is not evident.

The previous section discussed a technique for estimating the probabilities of evolution over a range of evolutionary distances given a specific ungapped alignment. We extend the technique to considering different alignments at each distance. We also provide an algorithm for efficiently computing the best alignment at every distance.

Consider an example of similar sequence fragments from the *C. elegans* cosmid C30C11[7] at offsets 30272 and 8522 respectively. There are two hypotheses regarding their duplication (table 3).

- Alignment A: There was a short duplication of 25 nucleotides with 2 observed mismatches.
- Alignment B: There was a longer duplication of 45 bases with 11 observed mismatches.

Alignment A is the first 25 bases of alignment B. The two alignments were discovered using PAM matrices, with a transition to transversion ratio of 1.5 and assuming a stable GC content of the *C. elegans* genome (35.7%). Figure 3 contains plots of the scores and corresponding probabilities of the two alignments as a function of the evolutionary distances. Sequence homologies are often

[7] Genbank Accession L09634 L18807

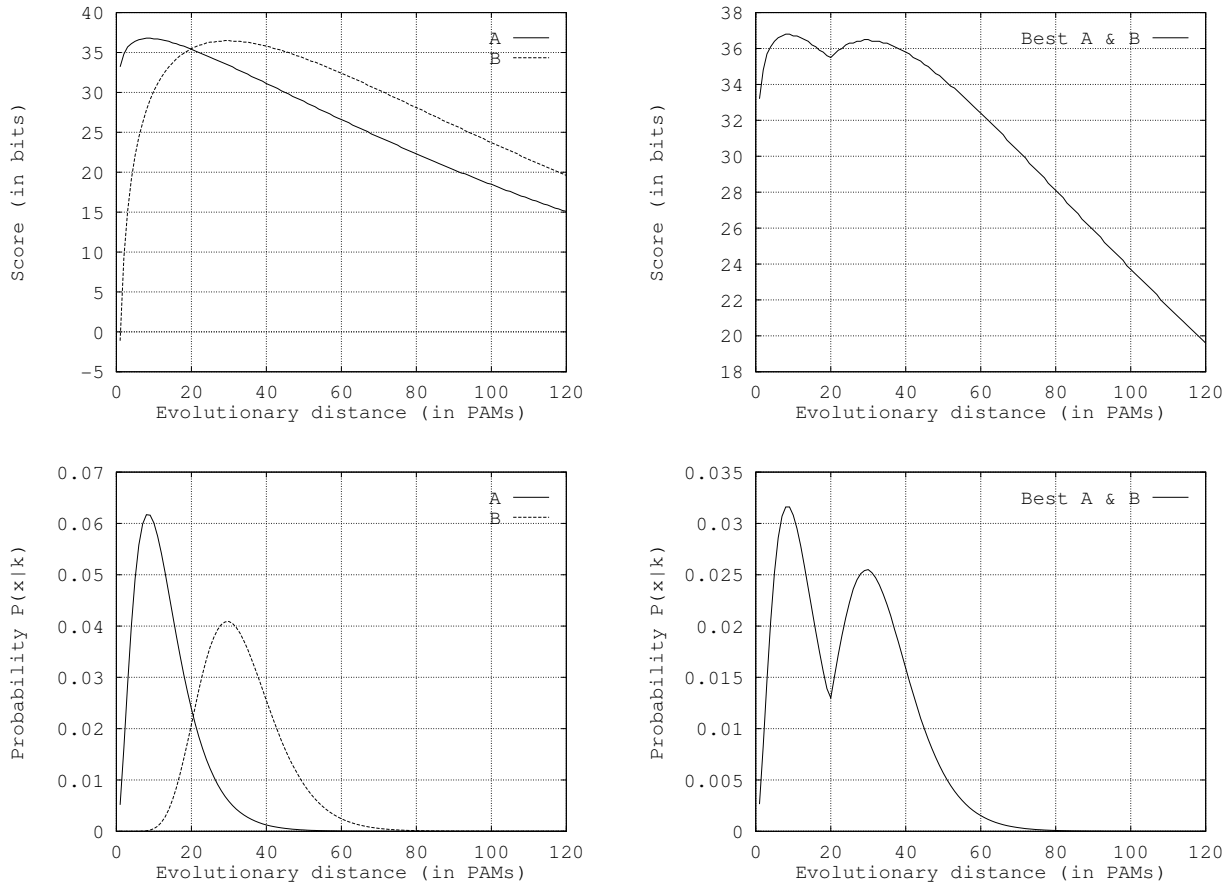



Figure 3: The variation of the scores of two possible alignments as a function of the evolutionary distance. Alignment A is optimal (higher-scoring and thus a higher probability of being related due to evolution) below PAM19, while B is optimal beyond PAM19. The last column in the table has the evolutionary distances estimated by considering the better alignment of A and B at each distance. The probabilities curves are normalized to have unit area under them.

```
tacagtactcgttaaaggcgcacac
||||||||||  ||||||||||||
tacagtactctctaaaggcgcacac
```

Alignment A

```
tacagtactcgttaaaggcgcacacccgtttgtatttaacgataa
||||||||||  ||||||||||||    |  |||| ||| | ||
tacagtactctctaaaggcgcacactttctcttattcaacaaaaa
```

Alignment B

| Method | Alignment A | Alignment B | Best alignment at each distance |
|---|---|---|---|
| $d_{JC}$ | 0.085 | 0.296 | 0.207 |
| $d_{mean}$ | 0.132 | 0.332 | 0.243 |

Table 3: Alignment averaging: Computing the mean distance over two possible alignments. The probability of each evolutionary distance ($P(x|k)$) is estimated as the exponential of the score of the best alignment using a PAM$x$ scoring matrix. The same probabilities are used for the two distance estimates. The distance estimate for Jukes-Cantor is a step function that changes from 0.085 for alignment A to 0.296 for alignment B at $\approx$ PAM19.

found using a single scoring matrix $(+5, -4)$[8] corresponding to approximately PAM47. The PAM47 matrix would identify alignment B as it scores higher at PAM47. Notice that the score is not highest at PAM47, and a bias regarding the evolutionary distance is introduced as a result of conducting the search utilizing a specific scoring matrix. Thus, most single matrix searches would fail to identify alignment A (whose score at PAM8 is the highest). Consequently, the evolutionary distance estimate would be based only on alignment B. Our proposed method determines the best alignment (with its probability) at each PAM, and computes the expected PAM from its probability distribution (as illustrated in figure 3 and table 3).

The extent of an ungapped sequence alignment often changes if it is scored with matrices corresponding to distinct evolutionary distances. The frequency of these changes is plotted in figure 4. The mean of the number of different ungapped alignments observed is approximately 4 when the search is conducted with matrices from PAM1 to 120. Thus, a search at a single PAM would have failed to evaluate more than 75% of the ungapped alignments identified at other PAM's, many of which contribute significantly to the evolutionary distance estimates.

We have demonstrated a technique to account for uncertainty in the alignment. The alignments we have considered are ungapped (no insertions or deletions). However, the same technique can be used for gapped alignments provided we can estimate the evolutionary distance for the various alignments as well as estimate the probability of each of those alignments being correct.

[8]The matches are scored as +5 and the mismatches scored as −4. This is the default scoring in BLASTN 1.4.8 (Altschul et al., 1990).

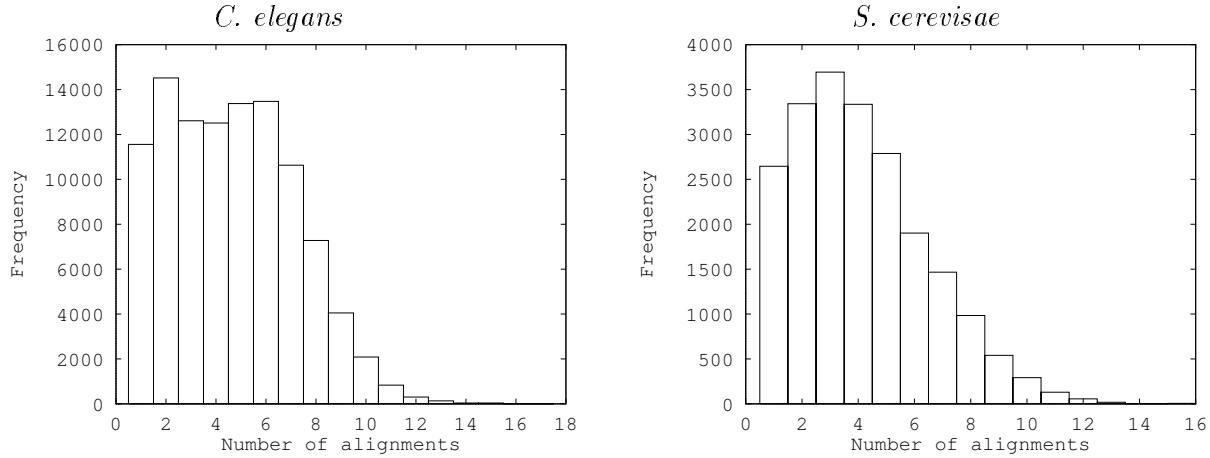



Figure 4: The frequency of the different number of observed ungapped alignments. These sets of alignments correspond to the set of repeats in a 3.66 Mb contiguous sequence from *C. elegans* and chromosomes III, VIII, and XI (1.54 Mb) from *S. cerevisae*. The range of distance used was PAM1 to PAM120.

# 4 Scoring using multiple matrices

Calculating $d_{mean}$ and accounting for the uncertainty in alignment require the computation of alignments over a range ($D$) of evolutionary distances. Sensitive searching of molecular sequence databases also requires the use of multiple matrices. Obvious algorithms using multiple scoring matrices need computation time that increases linearly with the number of matrices ($D$) used. We present an implementation with only a logarithmic overhead for a large class of scoring matrices. In addition, we show empirically that for most other scoring matrices, we can still compute the alignments and distances with only a logarithmic overhead; however, the computed distances have small, insignificant errors.

We consider a simple case of parametric sequence alignment. The only parameter considered is evolutionary distance. The score of the best alignment varies continuously as a function of the evolutionary distance. However, it has points of transition, and the derivative does not exist at the distance where the best alignment changes (illustrated in figure 3). Thus, the maxima of this function is computationally expensive to determine.[9] Fortunately, the length of the alignment is piece-wise constant (but not continuous), and given the length (and endpoints), the score can be computed in time proportional to the length of the alignment for a given scoring matrix. If we isolate all the distance intervals over which the alignment length is constant, then for each interval the score at each PAM can be determined in time proportional to the sum of the length of the interval and the alignment length. These distance intervals (over which the alignment length is fixed) are difficult to isolate if the alignment length is not a monotone function of the distance. For a certain class of matrices, termed *well-decaying*, the length of the alignment is a non-decreasing function of the evolutionary distance[10], and the cost of searching using a series of $N$ scoring matrices

[9] This is similar to the technique of finding the maximum of the likelihood surface employed by Thorne et al.(1991; 1992), but we are able to characterize our likelihood surface and provide a logarithmic algorithm for computing its maximum.

[10] Even for the matrices for which this is not strictly valid, we show that almost no sensitivity is lost by making

grows approximately as $\log N$ (because the number of distinct alignment lengths grows as $\log N$).

The array of scoring matrices most often used for nucleotide comparisons are the PAM matrices. The PAM number increases monotonically with increasing evolutionary distance.

We consider local alignments without gaps (no insertions or deletions), similar to the alignments produced by BLAST (Altschul et al., 1990). These alignments are simply obtained by walking along a single diagonal of the dynamic programming matrix. An alignment is maximal for a given scoring matrix if no extension on either side can increase its score. Every prefix and suffix of a maximal alignment has a positive score. Otherwise, the alignment minus that prefix or suffix has either a higher score or the same score (but is shorter) than the complete alignment; thus the complete alignment is no longer maximal. To prove that an alignment of length $L$ is maximal, it is sufficient to consider all extensions on either side until the score falls below zero. The length of this possible extension in the worst case can be the length of the sequence ($n$). However, for finding repeats in large genomic sequences, $L \ll n$, and the average extension required is much lower than $n$ (it is about $O(L)$). For the remainder of this section, we assume that given a specific diagonal and scoring matrix, a maximal ungapped alignment can be computed in time $O(L)$.

**Lemma 1** *The maximum score, using a range of $N$ scoring matrices, can be discovered in $O(LN)$ time.*

**Proof:** Trivially, each scoring matrix can be treated independently and the alignment rescored. □

**Lemma 2** *If the length of the alignment ($L$) is fixed over a range of ($N$) scoring matrices, then the maximum score obtained by using any of those matrices can be discovered in $O(L + \log N)$ time.*

**Proof:** The score as a function of PAM has a single maxima; thus a modified binary search will yield the maxima. The binary search is modified to check if the middle element is locally maximal, and it is only initiated if the score ($S$) is minimal at both ends, i.e. $S[1] < S[2]$ and $S[N] < S[N-1]$. We can preprocess the alignment length ($O(L)$) to count the number of each type of substitution, and fill out a substitution-count matrix. The alignment score (for a given scoring matrix) can be computed in constant time by evaluating a dot product of the substitution-count matrix and scoring matrix. □

Consider the scoring matrix at evolutionary time t, composed of some non-negative scoring entries $\{p \in P\}$ with scores $s_p \geq 0$, and negative scoring entries $\{n \in N\}$ with scores $s_n < 0$. With increasing evolutionary distance, the positive scores decay towards zero, and the negative scores rise towards zero (possibly rising beyond it). The class of scoring matrices for which equation 2 holds are termed *well-decaying* [11]. This class includes the Jukes-Cantor matrices.

$$\max_{n \in N} \frac{s_n(d+1)}{s_n(d)} < \min_{p \in P} \frac{s_p(d+1)}{s_p(d)} \tag{2}$$

**Lemma 3** *For a well-decaying set of matrices, if an alignment has a non-positive score at distance d+1, then it has a non-positive score at distance d. (Intuitively, matrices at higher PAM's are more tolerant of mismatches.)*

this assumption.

[11] In addition, for well-decaying matrices, $s(d) > 0 \Rightarrow s(d+1) > 0$.

**Proof:** Let the alignment score at evolutionary distance $d$ be $S(d)$ and at distance $d+1$ be $S(d+1)$.

$$S(d) = \sum_{p \in P} c_p s_p(d) + \sum_{n \in N} c_n s_n(d)$$

where $c_i$ is the count of the number of base pairs in the alignment scoring $s_i$.
For well-decaying matrices:

$$\max_{n \in N} \frac{s_n(d+1)}{s_n(d)} < \min_{p \in P} \frac{s_p(d+1)}{s_p(d)}$$

Let $n' \in N$ and $p' \in P$ be such that,

$$\forall\{n \in N\} \frac{s_n(d+1)}{s_n(d)} \leq \frac{s_{n'}(d+1)}{s_{n'}(d)} < \frac{s_{p'}(d+1)}{s_{p'}(d)} \leq \forall\{p \in P\} \frac{s_p(d+1)}{s_p(d)}$$

$$s_p(d)\frac{s_{n'}(d+1)}{s_{n'}(d)} \leq s_p(d+1) \quad (3)$$

$$s_n(d)\frac{s_{n'}(d+1)}{s_{n'}(d)} \leq s_n(d+1) \quad (4)$$

$$S(d+1) = \sum_{p \in P} c_p s_p(d+1) + \sum_{n \in N} c_n s_n(d+1) \leq 0$$

The sets $P$ and $N$ do not change from $d$ to $d+1$. The scores that change from negative to positive from PAM $d$ to PAM $d+1$ increase the score of an alignment ($S(d) < S(d+1)$), and thus we can eliminate them from consideration in this lemma. Furthermore, scores changing from positive to negative with increasing PAM are not permitted for well-decaying matrices.

Substituting from equations 3 and 4:

$$\Rightarrow \sum_{p \in P} c_p s_p(d)\frac{s_{n'}(d+1)}{s_{n'}(d)} + \sum_{n \in N} c_n s_n(d+1)\frac{s_{n'}(d+1)}{s_{n'}(d)} \leq S(d+1) \leq 0$$

$$\Rightarrow \frac{s_{n'}(d+1)}{s_{n'}(d)}\left(\sum_{p \in P} c_p s_p(d) + \sum_{n \in N} c_n s_n(d)\right) \leq 0$$

As $s_{n'}(d+1)/s_{n'}(d) > 0$

$$\Rightarrow c_{p'} s_{p'}(d) + c_{n'} s_{n'}(d) \leq 0$$

$$\Rightarrow S(d) \leq 0$$

□

**Theorem 1** *The maximal alignment length is a non-decreasing function of evolutionary distance for well-decaying scoring matrices.*

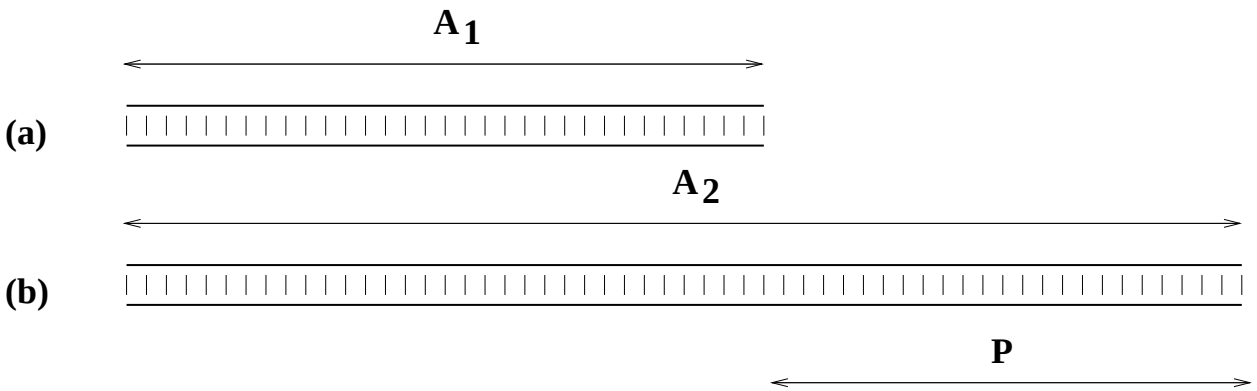



Figure 5: Alignment length increases with evolutionary distance. (a) Alignment ($A_1$) for evolutionary distance $d$. (b) The alignment ($A_2$) for evolutionary distance $d+1$ must not be smaller than $A_1$.

**Proof** by contradiction: Consider figure 5. The maximal alignment at evolutionary distance $d$ is $A_1$ and at distance $d+1$ is $A_2$. For the alignment length to decrease with increasing distance there is a piece (P) of the alignment that is part of $A_1$ and not part of $A_2$ (the alignments in the figure would need to be relabeled). The score of the piece, P, at $d$ must be positive; otherwise, $A_1$ is not a maximal alignment, as we can increase or maintain its score by dropping $P$ from the alignment. This $P$ must score non-positively at $d+1$; otherwise, $A_2$ is not maximal, as we can increase its score by adding $P$ to the alignment. Thus, $P$ is an alignment by itself, which scores positively at $d$ and non-positively at $d+1$, contradicting lemma 3. □

For well-decaying scoring matrices, the maximum score can be obtained in $O(N + I(L + \log(N/I)))$ computational time, where I is the number of different alignment lengths over the range of scoring matrices. ($0 < I \leq N$, but empirically $I \leq \log N$, see figure 4). This time bound follows from lemma 2 and the observation that it is most expensive to compute all the boundaries at which the alignment length changes, when these changes are spaced equally ($N/I$ scoring matrices) apart.

## 5 Experimental results

We have empirically evaluated the efficiency of searching using well-decaying matrices.

- The Jukes-Cantor PAM scoring matrix series is well-decaying. These matrices assume equal probability of mutation to any other base. They also assume that all the bases are equally likely.
- The bounds obtained in the previous section are asymptotic. Given that the number of matrices is limited, the bound is only useful if the constants involved are small, which is the case as even a simple implementation of searching using matrices with the well-decaying property provides a four-fold increase in speed over searching with each of the matrices independently.
- Introducing an unequal transition-transversion rate or accounting for GC content makes the set of matrices **not** well-decaying beyond a certain evolutionary distance.

| Data | # alignments | Score (in bits at best PAM) | Best PAM | Mean PAM |
|---|---|---|---|---|
| | | Root Mean Square Error | | |
| Yeast | 38489 | $8.0 \times 10^{-6}$ | 0.0002 | 0.0001 |
| Elegans | 102023 | $4.0 \times 10^{-5}$ | 0.0002 | 0.0001 |
| | | Relative Root Mean Square Error | | |
| Yeast | 38489 | $2.1 \times 10^{-9}$ | $4.4 \times 10^{-6}$ | $4.8 \times 10^{-6}$ |
| Elegans | 102023 | $7.7 \times 10^{-9}$ | $8.9 \times 10^{-6}$ | $4.8 \times 10^{-6}$ |
| | | Maximum Error | | |
| Yeast | 38489 | 0.29 | 2 | 3 |
| Elegans | 102023 | 1.85 | 5 | 4 |

Table 4: The various error rates as a result of assuming matrices to be well-decaying. The relative errors are obtained from the squared differences scaled by the datum value.

- Treating the matrices that have unequal transition-transversion rate and/or account for GC content as well-decaying rarely causes appreciable errors in the score estimates.

  We constructed the PAM scoring matrices using a transition to transversion ratio of 1.5 and the GC content (35.7%) of the *C. elegans* subset. These matrices are not well-decaying beyond 80 PAM's. We discovered repetitive sequence motifs using all these matrices. However, only 0.5% of the repeats had alignments that decreased in length with increasing PAM. The distance estimates of these alignments are affected only when the alignments have equal lengths at two different PAM's and the length changes in between these two PAM's. Table 4 provides estimates of the errors that were caused by assuming the matrices to be well-decaying. These errors are both small and rare enough to be ignored. For comparison, table 4 also includes data from the *S. cerevisae* subset.[12]

## Estimating duplication history in *C. elegans* and *S. cerevisae*

The techniques of Bayesian evolutionary distance estimation, incorporating uncertain sequence alignments, and efficient search using an array of matrices have been utilized to study the proliferation of repetitive sequence and motifs in the *C. elegans* and *S. cerevisae* genomes. All the repetitive sequences discovered are categorized according to the mean evolutionary distance estimated, considering the uncertainty in the sequence alignments. The cumulative score is plotted against the evolutionary distance in figure 6 for the *C. elegans* data. The plot shows a linear increase in the total amount of repetitive sequence between PAM1 and approximately PAM40. This suggests that genomic duplications have been taking place at an uniform rate over recent time. The number of repeats discovered beyond 40 PAM's decays gradually (1994).

The effects of conducting a search utilizing a single $(+5, -4)$ PAM47 scoring matrix, as opposed to utilizing an entire series, are illustrated in figure 7. The loss of sensitivity, in terms of the total bit score of all the significant duplications observed, is 9.6% in *C. elegans* and 12.1% for

[12]The GC content of *S. cerevisae* data was 38.5%; therefore, the matrices were quite similar to that produced for *C. elegans*.

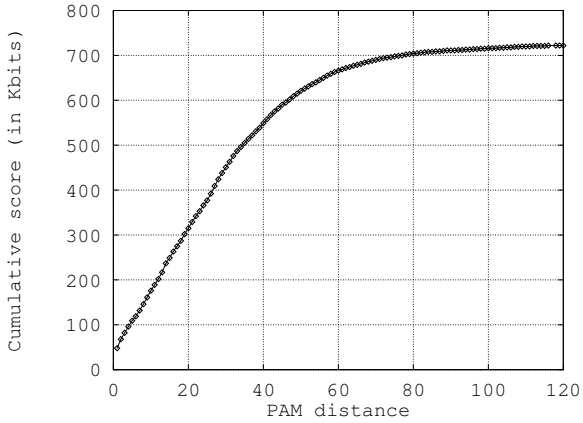



Figure 6: The amount of duplication plotted against the mean evolutionary distance for *C. elegans*.

*S. cerevisae*[13]. Furthermore, disconcerting discrepancies are observed in the evolutionary distance estimates. Repeats at certain evolutionary distances show a marked increase; at other distances they show a clear decline. The general trend shows a decline in the repeats found at the ends of the scale (low and high PAM's). Some of this decline is easily explained by the reduced sensitivity of the $(+5, -4)$ scoring matrix at low ($< 20$) and high ($> 70$) PAM's. The PAM47 matrix is less than 90% efficient below PAM20 and above PAM68 (States et al., 1991). The apparent high efficiency ($> 100\%$) observed at the PAM's close to 47 exposes a serious flaw. The PAM47 search is **not** discovering repeats that the All-PAM search failed to identify; it is only classifying them into the wrong PAM. If there are repeats with two different alignments at PAM47 and an extreme PAM, the PAM47 search identifies only the first alignment even if it scores much lower, leading to an incorrect estimation of evolutionary distance. Thus, the extent of the repeat that was discovered by $(+5, -4)$ matrix was incorrect, and any subsequent determination of the evolutionary distance will be incorrect[14].

# 6 Discussion

## Amino acids

The Bayesian estimates described in this paper can be readily extended to amino acid sequences, and the mean or median of the distribution utilized as an estimate. For amino acids sequences, ambiguity in the end points of the homology is unusual; thus the utility of the technique for considering possibly different alignments at each evolutionary distance is marginal.

The scoring technique employing multiple matrices discussed in section 4 is not as useful for amino acid sequences. There are 20 amino acids, and a substitution-count matrix has 205 entries. Often, counting all the mismatches at each PAM does not compare favorably to the length of the alignment (which is frequently less than 205). Thus, there would be little or no saving in computational time even if the PAM matrices were well-decaying. As it turns out, the PAM

[13]Details regarding the evaluation of significance for a repeat are provided by Agarwal and States (1994).

[14]Some of the differences in the plots at high PAM's, especially for *S. cerevisae*, are explained by regions with biased (high) GC content.

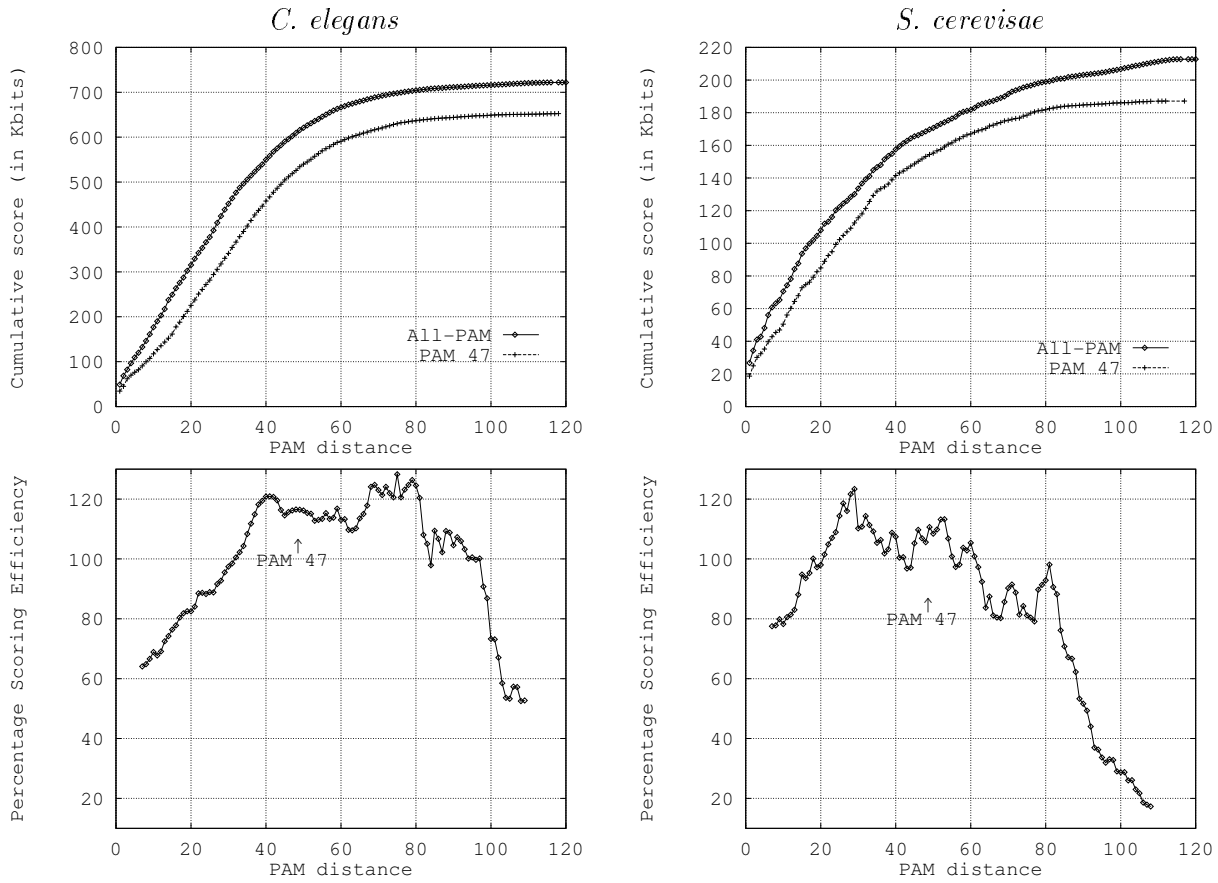



Figure 7: Amount of duplication plotted against evolutionary distance. The All-PAM plot is from a search using all the PAM matrices from 1 to 120. The PAM47 plot only includes repeats found using a PAM47 $(+5, -4)$. These sets of alignments correspond to the set of repeats in a 3.66 Mb contiguous sequence from *C. elegans* and chromosomes III, VIII, and XI (1.54 Mb) from *S. cerevisae*. The lower plots demonstrate the efficiency of the search at PAM47. Efficiency is measured by the amount of duplication found at PAM47 divided by the amount of duplication found by the All-PAM search. The lower plots have been smoothed by plotting a moving average over 15 PAM's.

matrices for amino acids are well-decaying only until PAM50, and the BLOSUM series is not well-decaying at all.

### Mutational model

The Bayesian approach to evolutionary distance estimation relies on a mutational model. However, this mutational model may vary between sites and over time. We could incorporate more precise mutational models that use dinucleotide or codon mutation rates. The maximum likelihood approach of Allison et al. (1992), Bishop and Thompson (1986), and Thorne et al. (1991; 1992) estimate both the best model and evolutionary distance; while the Bayesian estimate requires a mutation model at every base position. The advantage of the Bayesian approach is the savings in computational time, rendering it practical.

### Insertion/deletion events

The extension of these techniques to alignments with insertion/deletion events require associating a probabilty with an insertion/deletion at each evolutionary distance. It is obvious that the probability of an insertion/deletion should increase with increasing evolutionary distance. However, the dependence of the length of the insertion/deletion on the evolutionary distance is less certain. There appears to be no simple and correct way to associate probabilities with insertion/deletion events. Nevertheless, given the probabilities for gaps in alignments, they can be incorporated rather easily into the probability of evolution of one sequence from another at a given evolutionary distance. Furthermore, the best alignment can be found at each distance and the mean evolutionary distance estimated. It is computationally more expensive to discover the best gapped-alignment for each evolutionary distance; the cost is $O(NL^2)$ using N scoring matrices for two sequences, each of length L.

### Statistical significance

Altschul (1993) has proposed an empirical correction factor for estimating the statistical significance of similarities discovered using the All-PAM scoring system. Their proposed correction factor is for the score of the best alignment discovered.

A question that arises is if we consider the mean score of all the alignments (lower than the best score) — is a correction factor still required, and if so what should be its magnitude?

## Acknowledgments

We would like to thank Marcus B. Feldman and Michael Zuker for many useful discussions, Stanley Sawyer for comments on a draft version, an anonymous referee for many excellent suggestions, and Laureen C. Treacy for her dedicated proofreading. This work was supported in part by the Department of Energy (DOE) grant DE-FG02-94ER61910.

## References

Agarwal, P. and States, D. (1994). The Repeat Pattern Toolkit (RPT): Analyzing the structure and evolution of the *C. elegans* genome. In Altman, R., Brutlag, D., Karp, P., Lathrop, R.,

and Searls, D., editors, *Proceedings Second International Conference on Intelligent Systems for Molecular Biology*, pages 1–9, Menlo Park, CA. AAAI Press.

Allison, L., Wallace, C., and Yee, C. (1992). Finite-states models in the alignment of macromolecules. *J. Mol. Evol.*, 35:77–89.

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

Altschul, S. (1993). A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.*, 36:290–300.

Altschul, S., Boguski, M., Gish, W., and Wootton, J. (1994). Issues in searching molecular sequence databases. *Nature Genetics*, 6:119–129.

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

Bishop, M. and Thompson, E. (1986). Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.*, 190:159–165.

Blaisdell, B. (1986). A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Natl. Acad. Sci. USA*, 83:5155–5159.

Dayhoff, M., Schwartz, R., and Orcutt, B. (1979). A model of evolutionary change in proteins. In *Atlas of Protein Structure*, volume 5, chapter 22, pages 345–352. National Biomedical Research Foundation, Silver Spring, MD.

Doolittle, R. (1990). *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, volume 183 of *Methods in Enzymology*. Academic Press, New York.

Fitch, W. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155(760):279–284.

Fitch, W. and Smith, T. (1983). Optimal sequence alignments. *Proc. Natl. Acad. Sci. USA*, 80:1382–1386.

Gojobori, T., Moriyama, E., and Kimura, M. (1990). Statistical methods for estimating sequence divergence. In (Doolittle, 1990), chapter 33, pages 531–550.

Gusfield, D., Balasubramanian, K., and Naor, D. (1994). Parametric optimization of sequence alignment. *Algorithmica*, 12:312–326.

Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. In Munro, H., editor, *Mammalian Protein Metabloism*, volume 3, chapter 24, page 21. Academic Press, New York.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120.

Saqi, M. and Sternberg, M. (1991). A simple method to generate non-trivial alternate alignments of protein sequences. *J. Mol. Biol.*, 219:727–732.

States, D., Gish, W., and Altschul, S. (1991). Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods: A Companion to Methods in Enzymology*, 3(1):66–70.

Tajima, F. (1993). Unbiased estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.*, 10(3):677–688.

Tajima, F. and Nei, M. (1984). Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.*, 1(3):269–285.

Thorne, J., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, 33:114–124.

Thorne, J., Kishino, H., and Felsenstein, J. (1992). Inching toward reality: An improved likelihood model of sequence evolution. *J. Mol. Evol.*, 34:3–16.

Tillier, E. (1994). Maximum likelihood with multiparameter models of substitution. *J. Mol. Evol.*, 39(4):409–417.

Vingron, M. and Waterman, M. (1994). Sequence alignment and penalty choice: Review of concepts, case studies and implications. *J. Mol. Biol.*, 235:1–12.

Waterman, M. (1994). Parametric and ensemble sequence alignment algorithms. *Bull. Math. Biol.*, 56(4):743–767.

Waterman, M., Eggert, M., and Lander, E. (1992). Paramteric sequence comparisons. *Proc. Natl. Acad. Sci. USA*, 89:6090–6093.

Wilbur, W. (1985). On the PAM matrix model of protein evolution. *Mol. Biol. Evol.*, 2(5):434–447.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.*, 39:306–314.

Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.*, 39:315–329.

Zuckerkandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. *J. Theor. Biol.*, 8(2):357–366.

Zuker, M. (1991). Suboptimal sequence alignment in molecular biology: Alignment with error analysis. *J. Mol. Biol.*, 221:403–420.