

Optimal Scoring Matrices for Estimating Distances Between Aligned Sequences

Gaston Gonnet and Chantal Korostensky
e-mail: gonnet@inf.ethz.ch

March 12, 2004

Abstract

Distance information is usually obtained from aligned sequences. Computation of distances plays a crucial role in many aspects of bioinformatics, in particular in phylogenetic studies. In this paper, we show that any method estimating distances directly from scoring matrices is not consistent. Consistency is the property of a method which guarantees that given enough data it will construct the correct tree. Then we show how to correct this problem and to obtain consistent methods from scoring matrices. Following this line, we give an algorithm to compute the most effective/powerful estimator (the one with the lowest variance) for distances. Finally we illustrate all the steps of these computations with a complete example. We derive an optimal scoring matrix to estimate the distances between human mitochondrial DNA sequences. This optimal scoring matrix is interesting in itself, as there is quite a bit of interest in estimating the phylogenies of humans based on mtDNA.

1 Introduction

1.1 Model of Evolution

The model we consider here is a Markovian model of evolution [1], which assumes that amino acids mutate independently of each other, with probabilities which depend only on the amino acids and on the amount of evolution. In mathematical terms we can describe the model of evolution via mutation matrices: a mutation matrix, denoted by M , describes the probabilities of amino acid mutations for a given period of evolution.

$$Pr\{\text{amino acid } i \longrightarrow \text{amino acid } j\} = M_{ij} \quad (1)$$

The value $1 - M_{ii}$ represents the probability of mutating away from i . Amino acids appear in nature with different frequencies. These frequencies are denoted by f_i and correspond to the steady state of the Markov process defined by the matrix M , that is, the vector f is any of the columns of M^∞ or the eigenvector of M whose corresponding eigenvalue is 1 ($Mf = f$). When we find a mutation in aligned sequences, we typically cannot distinguish which one mutated into which, or that a third amino acid mutated into them. This implies a simple symmetry relation for the entries of M :

$$f_j \cdot M_{ij} = f_i \cdot M_{ji} \quad (2)$$

M describes mutations over a given period of evolution. In order to proceed, we must quantify this amount of change in a mathematically meaningful way. Dayhoff et. al.[4] introduced the term PAM (point accepted mutation) unit.

Definition 1.1 A 1-PAM unit is the amount of evolution which will change, on average, 1% of the amino acids. In mathematical terms, this is expressed as a matrix M such that $\sum_{i=1}^{20} f_i(1 - M_{ii}) = 0.01$, where f_i is the frequency of the i^{th} amino acid.

If we have a probability or frequency vector p , the product Mp gives the probability vector after an evolution equivalent to a 1-PAM unit. After k units of evolution (a k -PAM evolution), a frequency vector p will be changed into the frequency vector $M^k p$.

The probability for any amino acid X mutating into A and into B is (see Figure 1):

$$\sum_{X=1}^{20} f_X M_{AX}^{d_A} M_{BX}^{d_B} = f_A M_{BA}^d = f_B M_{AB}^d \quad (3)$$

where $d = d_A + d_B$. This is derived using the symmetry relation 2. We see that the probability depends only on the sum of the distances, not on each one.

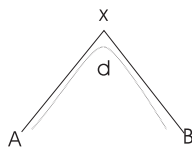


Figure 1: Amino acids A and B are d PAM apart and originate from a common ancestor x .

2 Dayhoff Scores and Evolutionary Trees

There are many tree construction methods based on PAM distances [8, 9, 11, 12, 2]. PAM distances are known to be consistent, that is, the more reliable the distances are (usually when more data is available), the more reliable the tree construction is. If an infinite amount of data would be available, then the correct evolutionary tree could be constructed [5, 7]. A question usually posed is, could scores derived from pairwise sequence alignments be used for evolutionary tree construction? Are they consistent?

To compute scores for aligning sequences, the mutation matrix is transformed into a matrix termed a *Dayhoff matrix*. The Dayhoff matrix, D , is related to a 250-PAM mutation matrix by

$$D_{AB} = 10 \cdot \log_{10} \frac{(M^{250})_{AB}}{f_A} \quad (4)$$

The entries in the Dayhoff matrices are the logarithm of the probability that the two amino acids evolved from a common ancestor as opposed to being random sequences. This results from the comparison of two events (see Figure 1):

- a) that the two sequences are independent of each other, and hence an arbitrary position with amino acid A aligned to another arbitrary position with amino acid B has the probability equal to the product of the individual frequencies

$$Pr\{\text{alignment of } A \text{ and } B \text{ by chance}\} = f_A f_B \quad (5)$$

- b) that the two sequences have evolved from some common ancestral sequence after some amount, d_A , and d_B of evolution.

$$\begin{aligned}
Pr\{A \text{ and } B \text{ descended from a common } x\} &= \sum_x f_x Pr\{x \rightarrow A\} Pr\{x \rightarrow B\} \\
&= \sum_x f_x (M^{d_A})_{Ax} (M^{d_B})_{Bx} \\
&= \sum_x f_B (M^{d_A})_{Ax} (M^{d_B})_{xB} \\
&= f_B (M_{AB}^d = f_A (M^d)_{BA}
\end{aligned}$$

$$D_{AB} = 10 \log_{10} \left(\frac{Pr\{A \text{ and } B \text{ descended from a common } x\}}{Pr\{\text{alignment of } A \text{ and } B \text{ by chance}\}} \right) \quad (6)$$

2.1 Monotonicity of scores and distances

Naturally we would assume that a lower PAM distance corresponds to a higher score, i.e. a higher probability that the sequences are related. First we show that this assumption is true when the lengths of the sequences are the same and we ignore deletions. To verify if smaller PAM distances correspond to larger scores, we have to look at the expected value of the score $S_D(d)$ as a function of the PAM distance d . We fix the length of the sequences to be n . The expected score $S_D(d)$ for each aligned amino acid is:

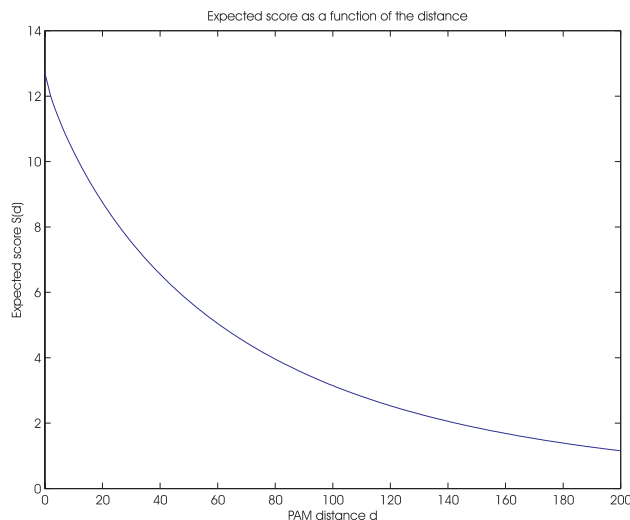


Figure 2: The expected score $S(d)$ as a function of the PAM distance d .

$$S_D(d) = \sum_{B=1}^{20} f_B \sum_{A=1}^{20} M_{AB}^d \cdot D_{AB} \quad (7)$$

where D_{AB} is the score of amino acids A and B . If we want the expected score for the whole sequences, we have to multiply this value by n . Calculating $S_D(d)$ for each d from 0 to 200, we get the plot in Figure 2.

The function is monotonic and decreasing, meaning that a larger PAM distance d does correspond to a lower score $S_D(d)$. The graph is only a "visual" proof. To prove this formally, we have to analyze the derivative of $S_D(d)$ with respect to d . To compute this derivative M^d can be rewritten as:

$$M^d = U\Lambda^d U^{-1} \quad (8)$$

where Λ is a diagonal matrix containing the eigenvalues of M . In Λ^d each diagonal element is λ_i^d . $S_E(d)$ can now be rewritten as (for any matrix E):

$$S_E(d) = \sum_{B=1}^{20} f_B \sum_{A=1}^{20} (U\Lambda^d U^{-1})_{AB} \cdot E_{AB} \quad (9)$$

if we multiply everything out, then we can rewrite $S_E(d)$ as

$$S_E(d) = \sum_{B=1}^{20} f_B \sum_{A=1}^{20} E_{AB} \sum_{i=1}^{20} U_{Bi} U_{iB}^{-1} \lambda_i^d = \sum_{i=1}^{20} T_{ii} \lambda_i^d \quad (10)$$

where T_{ii} is the i^{th} diagonal entry in the matrix $T = U^{-1} F E^T U$ (F is a diagonal matrix with the frequencies $F_{ii} = f_i$). Its derivative is

$$S'_E(d) = \sum_{i=1}^{20} T_{ii} \lambda_i^d \ln \lambda_i \quad (11)$$

and can be computed and verified to be negative for all d in the range 0..250, for the Dayhoff matrix in [10]. This follows from all the T_{ii} being positive and all $\lambda_i \leq 1$.

2.2 All scoring matrices are inconsistent to build phylogenetic trees

Imagine a tree as in Figure 3. In this tree, the PAM distances follow:

$$d_{AB} + d_{CD} < d_{AC} + d_{BD} \quad (12)$$

If scores were consistent to build trees, then the sum of the first two scores must be larger than the sum of the other two scores (see Figure 3):

$$S_{AB} + S_{CD} > S_{AC} + S_{BD} \quad (13)$$

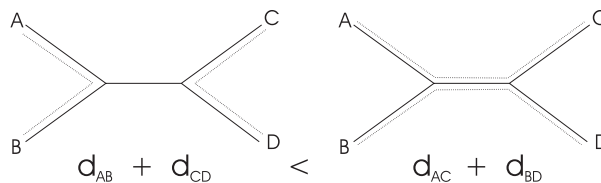


Figure 3: If Dayhoff scores are consistent, then $S_{AB} + S_{CD} > S_{AC} + S_{BD}$

As we have seen the expected score $S_D(d)$ as a function of d is not a straight line, which means that the scores are not consistent. We now show this with a nice geometrical proof.

Theorem 2.1 $d_{AB} + d_{CD} < d_{AC} + d_{BD}$ does not imply $S_{AB} + S_{CD} > S_{AC} + S_{BD}$

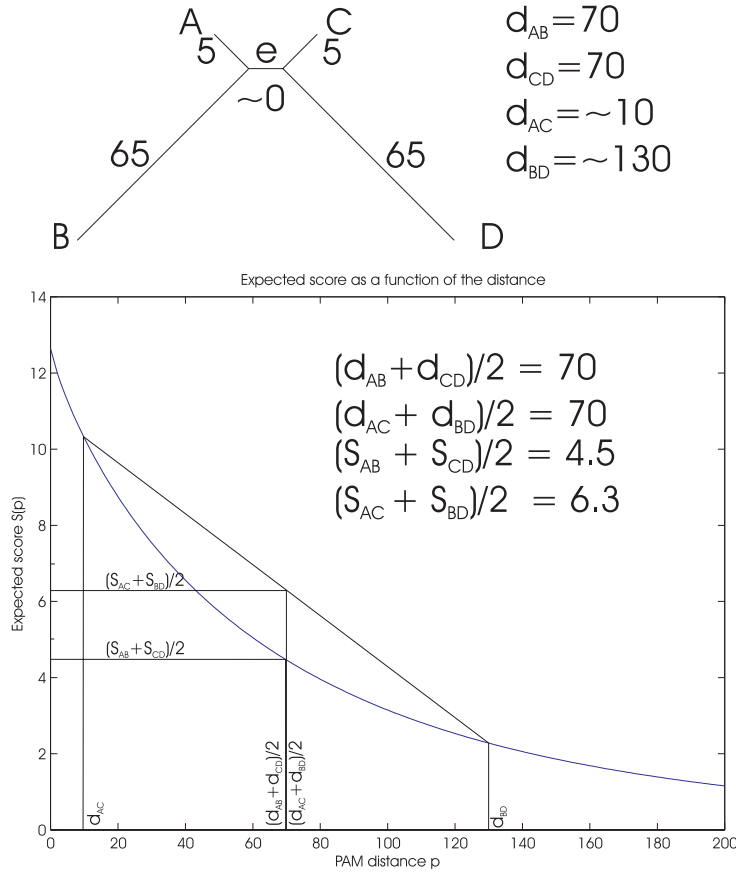


Figure 4: For certain types of trees equation 13 does not hold when equation 12 is true

Proof 2.1 To prove this, we construct a tree with four leaves, where two edges are very long and the other edges are very short (see Figure 4 on top). We choose the middle edge e to be very short. We also choose the distances so that $d_{AB} + d_{CD} = d_{AC} + d_{BD}$. For simplicity we look at $\frac{1}{2}$ of the distances and scores. As you can see in Figure 4, both sums of distances are the same: $(d_{AB} + d_{CD})/2 = 70$, and $(d_{AC} + d_{BD})/2 = 70$.

The scores can be read from the graph. The score $(S_{AC} + S_{BD})/2 \approx 6.3$ is the midpoint of the line between S_{AC} and S_{BD} . We do the same for $(S_{AB} + S_{CD})/2 \approx 4.5$. This case is simpler, as both values are the same, so we already have the midpoint. If the graph of $S(d)$ is not a straight line, there will be points where the sum of the scores at the end will be different than twice the score at the midpoint.

Now we make e slightly larger than 0 so that $(d_{AC} + d_{BD})/2$ is now slightly greater than 70, but the score $(S_{AC} + S_{BD})/2$ is still clearly larger than the score $(S_{AB} + S_{CD})/2$. Even though $d_{AB} + d_{CD} < d_{AC} + d_{BD}$, the condition for the scores, $S_{AB} + S_{CD} > S_{AC} + S_{BD}$, does not hold. If the curvature was negative, then we would move $(d_{AC} + d_{BD})/2$ to be slightly lower than 70. Unless there is no curvature, i.e. $S(d)$ is a straight line, we can always find a counter example.

The conclusion is: if we use scores derived from Dayhoff matrices to construct trees, we could obtain incorrect results, no matter how much data is available. In this example, we would decide to connect leaves AC and BD , not AB and CD . Felsenstein [6] noted a similar result for parsimony. This means that scores derived from Dayhoff matrices should not be used to construct

evolutionary trees. They are positively misleading in some cases, as is parsimony.

The result is more general, whatever scoring matrix E we use, if the expected score $S_E(d)$ as a function of d is not a straight line, we can derive counterexamples like the one in Figure 4. As $S_E(d)$ is a linear combination of exponentials in d , it will never be a straight line. So no scoring matrix E can give consistent scores usable for constructing trees.

2.3 How to Make Scores Consistent

As it turns out to be, making scores consistent is quite straightforward. For any scoring matrix E for which $S_E(d)$ is monotonic we can invert $S_E(d)$. Our distance is computed using the inverse, i.e. $d^* = S_E^{-1}(X/n)$ where X is the actual score obtained from a pair of aligned sequences $\langle a, b \rangle$.

$$X = \sum_{i=1}^n E_{a_i, b_i}$$

For n sufficiently large, X/n converges to its expected value $S_E(d)$ and d^* converges to d . Since $S_E(d)$ is a sum of exponentials in d , it is generally not possible to invert it algebraically. The computation of the inverse has to be done numerically. Let

$$S_E^k(d) = \sum_{B=1}^{20} f_B \sum_{A=1}^{20} M_{AB}^d E_{AB}^k$$

be the k^{th} moment of the scoring matrix E . The central moments of X/n are:

$$E[X/n] = S_E(d) \tag{14}$$

$$\sigma^2(X/n) = E[(X/n - S_E(d))^2] = \frac{S_E^2(d) - S_E(d)^2}{n} = \frac{\mu_2}{n} \tag{15}$$

$$E[(X/n - S_E(d))^3] = \frac{S_E^3(d) - 3S_E^2(d)S_E(d) + 2S_E(d)^3}{n^2} = \frac{\mu_3}{n^2} \tag{16}$$

$$E[(X/n - S_E(d))^k] = O(n^{\lfloor -k/2 \rfloor}) \tag{17}$$

3 Deriving the Optimal Scoring Matrix

Our next goal is to construct a scoring matrix that is optimal to estimate distances. An optimal estimator (scoring function) has the smallest variance. Let $d^* = S_E^{-1}(X/n)$ be our estimate of the distance using X . To compute the expected value of d^* and its variance, we use the Taylor expansion of $S_E^{-1}(X/n)$ around $S_E(d)$. These computations were done with a computer algebra system (Maple [3]), they are too tedious and error prone to be done by hand.

$$S_E^{-1}(X/n) = S_E^{-1}(S_E(d)) + (X/n - S_E(d))S_E^{-1'}(S_E(d)) + (X/n - S_E(d))^2 \frac{S_E^{-1''}(S_E(d))}{2} + \tag{18}$$

$$O((X/n - S_E(d))^3)$$

The derivative $S_E^{-1'}(S_E(d))$ is $\frac{1}{S_E'(d)}$, because for any functions $f(x), g(x)$, where $f(g(x)) = x$,

the derivative $f'(g(x)) = \frac{1}{g'(x)}$. The second derivative $S_E^{-1''}(S_E(d))$ is $-\frac{S_E''(d)}{S_E'(d)^3}$, so

$$S_E^{-1}(X/n) = d + (X/n - S_E(d)) \frac{1}{S_E'(d)} - (X/n - S_E(d))^2 \frac{S_E''(d)}{2S_E'(d)^3} + O((X/n - S_E(d))^3) \tag{19}$$

Taking expected values we find that

$$E[d^*] = d + E[X/n - S_E(d)] \frac{1}{S'_E(d)} - E[(X/n - S_E(d))^2] \frac{S''_E(d)}{2S'_E(d)^3} + O\left(\frac{1}{n^2}\right) \quad (20)$$

$$= d - \sigma^2(X/n) \frac{S''_E(d)}{2S'_E(d)^3} + O\left(\frac{1}{n^2}\right) \quad (21)$$

$$= d - \frac{S''_E(d) - S_E(d)^2}{n} \frac{S''_E(d)}{2S'_E(d)^3} + O\left(\frac{1}{n^2}\right) \quad (22)$$

since $E[X/n - S_E(d)] = E[X/n] - S_E(d) = 0$. Notice that when $n \rightarrow \infty$ then $E[d^*] \rightarrow d$ and hence $E[d^*]$ is an unbiased estimator (we could correct it to any order of n if desired). The variance $\sigma^2(d^*) = E[(d^* - E[d^*])^2]$ can be derived as follows: with the asymptotic value of $E[d^*]$ we compute the Taylor series of $(d^* - E[d^*])^2$ in powers of $(X/n - S_E(d))$. Then we take expected values, replacing the powers of $(X/n - S_E(d))^k$ by the central moments described in equations 17, 18 and 19. Truncating after two terms we obtain

$$\sigma^2(d^*) = \frac{\mu_2}{S'_E(d)^2 n} - \frac{S''_E(d) \left(\frac{S''_E(d) \mu_2^2}{4} + \mu_3 S'_E(d)^2 \right)}{S'_E(d)^6 n^2} + O\left(\frac{1}{n^3}\right) \quad (23)$$

The most effective/powerful estimator is the one among all possible estimators which has minimal variance. Since the variance is

$$\sigma^2(d^*) = \frac{F(d)}{n} + O\left(\frac{1}{n^2}\right) \quad (24)$$

where

$$F(d) = \frac{\mu_2}{S'_E(d)^2} = \frac{S''_E(d) - S_E(d)^2}{S'_E(d)^2} \quad (25)$$

we will compute E so that it minimizes $F(d)$. This will give us asymptotically (in n) optimal estimators. We have several choices for the minimization: we can derive the best E for

- a given distance d , e.g. $\min_{E_{ij}} F(d)$.
- a norm over a range of distances, e.g. for $d = 0..200$, i.e. $\min_{E_{ij}} \int_0^{200} F(t) dt$.
- the minimax over a range of distances, i.e for $d = 0..200$, i.e $\min_{E_{ij}} \max_{d=0..200} F(d)$.

There is no hope of finding a closed formula for this optimal E , but the first and second cases can be computed numerically without much difficulty.

3.1 Degrees of Freedom

Because our model of evolution is symmetrical, (equation 2) E must be symmetric. It is also easy to see that if we replace E_{ij} by $\alpha E_{ij} + \beta$ then $F(d)$ does not change (both numerator and denominator are multiplied by α^2). Before we do any minimization it is convenient to eliminate these two degrees of freedom. This can be done by arbitrarily setting two values of E , e.g. $E_{11} = -1$ and $E_{12} = 1$. By setting a diagonal element negative and an off diagonal element positive, the function $S_E(d)$ becomes monotonically increasing.

4 Example of optimal scores for mitochondrial DNA

We will compute the optimal values to estimate distances for human mitochondrial DNA sequences. We will follow in detail all steps of the computation. For compactness we use DNA instead of amino acids, as all the matrices and vectors are of dimension 4 instead of 20.

We use as source data the 89 complete mitochondrial human genomes used for Eve's tree. [?] At this point we will ignore the fact that some of the mitochondrial DNA is coding DNA and hence a more sophisticated method, which considers the individual codons, should be used. The source data can be summarized by a count matrix of selected alignments of the sequences. This count matrix is obtained by adding 1/2 at C_{ij} and C_{ji} every time that we align a base i against a base j . The count matrix is symmetric. The bases are numbered: A=1 C=2 G=3 T=4 for indexing in the matrices.

$$C = \begin{bmatrix} 6764069 & 1890.5 & 16503.5 & 1270.5 \\ 1890.5 & 6847226 & 360.5 & 20608 \\ 16503.5 & 360.5 & 2868325 & 120.5 \\ 1270.5 & 20608 & 120.5 & 5404882 \end{bmatrix}$$

Because there is little mutation among the mitochondrial DNA of humans, the count matrix is strongly diagonally dominant.

The first step is to compute the mutation matrix (at 1 PAM) from these counts. This is done by normalizing the count matrix so that each column adds to 1 and then powering the matrix to a suitable power so that the resulting matrix is 1-PAM. The M matrix satisfies the normal properties: the columns should add to 1 and the pseudo-symmetry condition $f_i M_{ji} = f_j M_{ij}$.

$$M = \begin{bmatrix} 0.9922 & 0.0007446 & 0.01540 & 0.0006339 \\ 0.0007541 & 0.9910 & 0.0003402 & 0.01023 \\ 0.006548 & 0.0001429 & 0.9841 & 0.00006149 \\ 0.0005071 & 0.008084 & 0.0001156 & 0.9891 \end{bmatrix}$$

The frequency vector can be computed from the eigenvector of M whose eigenvalue is 1 or from the first row/column of M ($f_i = \frac{M_{i1}}{M_{1i} \sum_j M_{j1}/M_{1j}}$) or from the initial count matrix C . All methods give the same results. Since the frequencies f can be derived from a mutation matrix M , only M is needed to do all these computations.

$$f_1 = .3088 \quad f_2 = .3128 \quad f_3 = .1314 \quad f_4 = .2471$$

The matrix M is 1-PAM, as can be checked by computing:

$$\sum_{i=1}^4 f_i (1 - M_{ii}) = 0.01$$

The eigenvalues of M are

$$\lambda_1 = 1.0000 \quad \lambda_2 = .9982 \quad \lambda_3 = .9809 \quad \lambda_4 = .9773$$

In this case we want to find the optimal scoring matrix for d in the range $.1 \leq d \leq 10$. This means that instead of minimizing the function $F(d)$ (eq 25) for one particular distance, we will

minimize its average (integral) over the given interval. After setting $E_{11} = -1$ and $E_{12} = 1$ and numerically minimizing $\int_{.1}^{10} F(t)dt$ on the rest of the E_{ij} unknowns, we obtain:

$$E = \begin{bmatrix} -1.0 & 1.0 & 0.9618 & 1.004 \\ 1.0 & -1.001 & 1.044 & 0.9701 \\ 0.9618 & 1.044 & -1.006 & 1.155 \\ 1.004 & 0.9701 & 1.155 & -1.002 \end{bmatrix}$$

This minimization is the most difficult computational step. In this case it was computed using Maple [3]. From these values, using equation 10, we can compute

$$S(d) = -.5529\lambda_2^d - .3672\lambda_4^d - .5452\lambda_3^d + .4634$$

The derivative of $S(d)$ with respect to d is:

$$S'(d) = .001002\lambda_2^d + .008411\lambda_4^d + .01051\lambda_3^d$$

and we can see that all the terms are strictly positive, so that the function is strictly monotonic for all d . This makes computing the inverse of $S(d)$ quite easy, even though it has to be done with a numerical method. The second moment of E (needed to compute the variance of the estimator) is:

$$S_E^2(d) = -.04637\lambda_2^d + .01677\lambda_4^d + .0170\lambda_3^d + 1.016$$

The scoring matrix E and the function $S(d)$ are all what we need to estimate PAM distances between sequences in an optimal way.

Suppose we have aligned these two sequences:

```
G C A G T G T C T
G T A G A A C C T
```

The score for this alignment using E divided by it's length is

$$w/n = (E_{33} + E_{24} + E_{11} + E_{33} + E_{41} + E_{31} + E_{42} + E_{22} + E_{44})/9 = -.1233$$

From this value, and using any numerical method, we can find the unbiased estimate of the distance, d^* which satisfies the equation $S(d^*) = -.1233$. The solution is $d^* = S^{-1}(-.1233) = 97.26$. The variance of the estimator (eq 24) is $\frac{S^2(d) - S(d)^2}{nS'(d)^2} \approx 9525.7$. The variance is very large, and so is the standard deviation (97.60), which is not unexpected for such a short alignment.

References

- [1] Pierre Baldi, Yves Chauvin, Tim Hunkapiller, and Marcella A. McClure. Hidden markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA*, 91:1059–1063, 1994.
- [2] L. Cavalli-Sforza and A. Edwards. Phylogenetic analysis: models and estimation procedures. *Evolution*, 32:233–57, 1967.
- [3] Bruce W. Char, Keith O. Geddes, Gaston H. Gonnet, Benton L. Leong, Michael B. Monagan, and Stephen M. Watt. *Maple V Language Reference Manual*. Springer-Verlag, 1991.

- [4] Margaret Oakley Dayhoff, Robert M. Schwartz, and Bruce C. Orcutt. A model for evolutionary change in proteins. In Margaret Oakley Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, pages 345–352. National Biomedical Research Foundation, 1978.
- [5] J. Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Amer. J. Human Genetics*, 25:471–492, 1973.
- [6] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27:27 – 33, 1978.
- [7] J. Felsenstein. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.*, 266:418–427, 1996.
- [8] W.M. Fitch and E. Margoliash. The construction of phylogenetic trees. *Science*, 155:279 – 284, 1967.
- [9] Gaston H. Gonnet and Steven A. Benner. Probabilistic ancestral sequences and multiple alignments. In *Fifth Scandinavian Workshop on Algorithm Theory*, pages 380–391, Reykjavik, Iceland, July 1996.
- [10] Gaston. H. Gonnet, Mark A. Cohen, and Steven A. Benner. Exhaustive matching of the entire protein sequence database. *Science*, 256:1443–1445, June 1992.
- [11] J. Hein. An optimal algorithm to reconstruct trees from additive distance data. *Bull. Math. Biol.*, 51:597 – 603, 1989.
- [12] P. Hogeweg and P. Hesper. The alignment of sets of sequences and the construction of phylogenetic trees: an integrated method. *J. Mol. Evol.*, 20:175 –86, 1988.