Optimal Scoring Matrices for Estimating Distances Between Aligned Sequences

Gaston Gonnet and Chantal Korostensky
e-mail: {gonnet, korosten}@inf.ethz.ch

July 14, 1999

Abstract

Sequence alignment is typically the first step in many research areas of bioinformatics, where some form of score or distance is derived. Those scores and distances are often used for evolutionary tree construction, multiple sequence alignments, all-against-all comparisons of whole genomes and many other tasks. Since those scores and distances are the basis for further studies, it is important that they can be estimated as well as possible. In this paper, we prove that the scores obtained from Dayhoff matrices (or from any other matrix) are not consistent for tree construction. Then we show how this can be corrected and how to create an optimal scoring matrix to estimate distances. This scoring matrix is optimal within a large class of estimators. Finally we present a complete example.

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 with mutation matrices: a mutation matrix, denoted by M, describes the probabilities of amino acid mutations for a given period of evolution.

$$Pr\{\text{amino acid } i \longrightarrow \text{amino acid } j\} = M_{ij}$$

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^∞ (M multiplied by itself an infinite number of times) or the eigenvector of M whose corresponding eigenvalue is 1 ($Mf = f$). When we find a mutation in aligned sequences, we cannot distinguish which one mutated into which, or whether a third amino acid mutated into them. This implies a simple relation for the entries of M:

$$f_j \cdot M_{ij} = f_i \cdot M_{ji}$$

M describes mutations over a given period of evolution. We must quantify this amount of change in a mathematically meaningful way. Dayhoff et. al.[3] 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 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 $M^k p$.

There are $8000 = 20 \times 20 \times 20$ possible events where an amino acid X mutates in one branch of evolution to amino acid A and to B in another branch. The sum of all these probabilities must be 1.

$$\sum_{X=1}^{20} f_X \sum_{A=1}^{20} \sum_{B=1}^{20} M_{AX}^{d_{XA}} M_{BX}^{d_{XB}} = \sum_{A=1}^{20} f_A \sum_{B=1}^{20} M_{BA}^{d} = 1 \tag{1}$$

where $d = d_{XA} + d_{XB}$. The above left hand side can be summed in X and we can express the probabilities of each event in terms of the children A and B (present day) and the total amount of evolution between A and B.

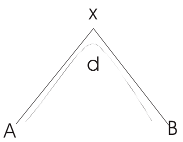

Figure 1: Amino acids A and B are d PAM units apart and originate from a common ancestor X.

2 Dayhoff Scores and Evolutionary Trees

There are many tree construction methods based on PAM distances [7, 8, 10, 11, 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 [4, 6]. A question usually posed is: can 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

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

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) the two sequences are independent of each other. 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

$$Pr\{\text{independent alignment of } A \text{ and } B\} = f_A f_B$$

b) the two sequences have evolved from some common ancestral sequence after some amount t of evolution.

$$\begin{aligned} Pr\{A \text{ and } B \text{ from a common ancestor } X\} &= \sum_X f_X Pr\{X \to A\} Pr\{X \to B\} \\ &= \sum_X f_X (M^t)_{AX} (M^t)_{BX} \\ &= \sum_X f_B (M^t)_{AX} (M^t)_{XB} \\ &= f_B (M^{2t})_{AB} = f_A (M^{2t})_{BA} \end{aligned}$$

$$D_{AB} = 10 \log_{10} \left(\frac{Pr\{A \text{ and } B \text{ descended from some } X\}}{Pr\{A \text{ and } B \text{ are independent}\}} \right)$$

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. 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. To do this we look at sequences with length n. The expected score $S_D(d)$ for each aligned amino acid is

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

where $D_{AB,d}$ is the score of amino acids A and B at PAM distance d and the probabilities are from equation 1. In our case, D_d is a Dayhoff matrix for PAM distance d. 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, we have to analyze the derivative of $S_D(d)$ with respect to d. To compute this derivative, M^d can be rewritten as

$$M^d = U \Lambda^d U^{-1}$$

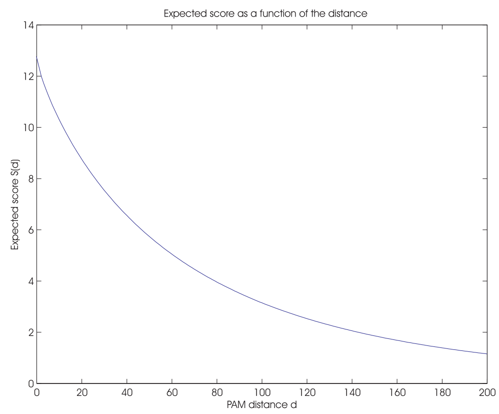

Figure 2: The expected score $S(d)$ as a function of the PAM distance d.

where Λ is a diagonal matrix containing the eigenvalues of M. In Λ^d each diagonal element is $\Lambda_{ii} = \lambda_i^d$. For any matrix E, $S_E(d)$ can be rewritten as

$$S_E(d) = \sum_{B=1}^{20} f_B \sum_{A=1}^{20} (U\Lambda^d U^{-1})_{AB} \cdot E_{AB} \tag{2}$$

If we multiply everything out we can rewrite $S_E(d)$ as

$$S_E(d) = \sum_{B=1}^{20} f_B \sum_{A=1}^{20} E_{AB} \sum_{i=1}^{20} U_{Bi} U_{iB}^{-1} \lambda_i^d = \sum_{i=1}^{20} T_{ii} \lambda_i^d \tag{3}$$

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

$$S_E'(d) = \sum_{i=1}^{20} T_{ii} \lambda_i^d \ln \lambda_i$$

and can be computed and verified to be negative for all d in the range 0 to 250, for the Dayhoff matrix in [9].

2.2 Scores are inconsistent to build phylogenetic trees

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

$$d_{AB} + d_{CD} < d_{AC} + d_{BD} \tag{4}$$

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):

$$S_{AB} + S_{CD} > S_{AC} + S_{BD} \tag{5}$$

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

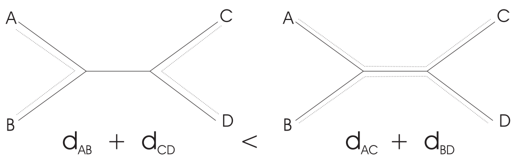

Figure 3: If Dayhoff scores are consistent, then $S_{AB} + S_{CD} > S_{AC} + S_{BD}$

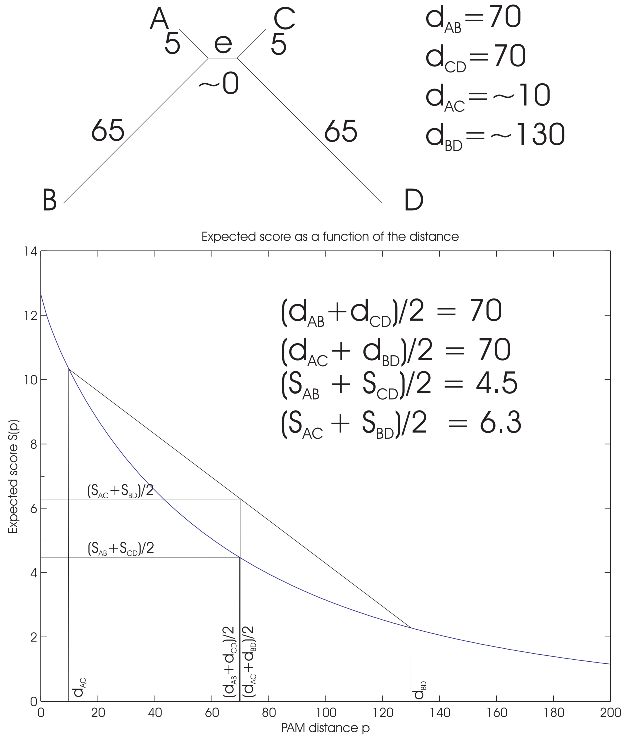

Figure 4: For certain types of trees equation 5 does not hold when equation 4 is true

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

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 length of the middle edge e to be close to zero. 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 $(d_{AC} + d_{BD})/2$ is the midpoint of the line between S_{AC} and S_{BD}. The intersection with the y axis gives us the score, which is about 6.3. We do the same for $(S_{AB} + S_{CD})/2$. This case is simpler, as both values are the same, so we already have the midpoint. The score is around 4.5. If the graph is not a straight line, there will be points where the sum of the distances are the same, and sum of the scores are different.

Now we change the length of edge e slightly by adding a very small amount. $(d_{AC} + d_{BD})/2$ 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 the example, we would decide to connect leaves AC and BD, not AB and CD. Felsenstein [5] 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 (equation 2), 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, this is straightforward. For any scoring matrix E for which $S_E(d)$ is strictly monotonic we can invert $S_E(d)$. Our new score is computed using the inverse, i.e.

$$d^* = S_E^{-1}(w/n) \tag{6}$$

where w is the actual score obtained from a pair of aligned sequences $\langle a, b\rangle$.

$$w = \sum_{i=1}^{n} E_{a_i b_i} \tag{7}$$

For n sufficiently large, w/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 will 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 w/n are:

$$E[w/n] = S_E(d)$$

$$\sigma^2(w/n) = E[(w/n - S_E(d))^2] = \frac{S_E^2(d) - S_E(d)^2}{n} = \frac{\mu_2}{n} \tag{8}$$

$$E[(w/n - S_E(d))^3] = \frac{S_E^3(d) - 3S_E^2(d)S_E(d) + 2S_E(d)^3}{n^2} = \frac{\mu_3}{n^2} \tag{9}$$

$$E[(w/n - S_E(d))^k] = O(n^{\lfloor -k/2 \rfloor}), \quad (k \ge 1) \tag{10}$$

2.4 How general are these scoring functions

A completely generic scoring function is

$$S(< a, b >) = G(E_{a_1b_1}, E_{a_2b_2}, ...)$$

where E is a 20×20 arbitrary symmetric matrix and G is an arbitrary function which is symmetric in all its arguments. The symmetry of E is needed because when we inspect a mutation we cannot distinguish which mutated from which. Hence assigning different values to E_{ij} and E_{ji} is biologically incorrect. The symmetry of G is required because each mutation is unrelated to any other and does not depend on its position in the alignment. Therefore G depends only on the number of values E_{ij} for each i and j.

A brute force proof shows that for any symmetric polynomial of degree at most 3, the most efficient estimator is the same as for the linear case. At this time we are unable to prove that this holds for an arbitrary degree or for an arbitrary function, but we conjecture that the linear estimator (equation 6) is optimal for any symmetric function. Notice that it should be possible to approximate any reasonable estimator by an arbitrary degree polynomial, hence we conjecture that the estimator described here is optimal among all *reasonable* functions G.

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 $d^* = S_E^{-1}(w/n)$ be our estimate of the distance using w. To compute the expected value of d^* and its variance, we use the Taylor expansion of $S_E^{-1}(w/n)$ around $S_E(d)$. These computations were done with a computer algebra system (Maple), they are too tedious and error prone to be done by hand.

3.1 Expected value of the estimator

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

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

Simplifying S_E^{-1} and its derivatives we can rewrite the above equation as

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

Taking expected values we find that

$$E[d^*] = d + E[w/n - S_E(d)]\frac{1}{S_E'(d)} - E[(w/n - S_E(d))^2]\frac{S_E''(d)}{2S_E'(d)^3} + O(\frac{\mu_2}{n^2})$$

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

since $E[w/n - S_E(d)] = E[w/n] - S_E(d) = 0$. When $n \to \infty$ then $E[d^*] \to d$, and hence $E[d^*]$ is a consistent estimator (we could correct it to any order of n if desired).

3.2 Variance of the estimator

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 $(w/n - S_E(d))$. Then we take expected values, replacing the powers of $(w/n - S_E(d))^k$ by the central moments described in equations 8, 9 and 10. Truncating after two terms we obtain

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

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

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

where $F(d) = \frac{\mu_2}{S_E'(d)^2}$. We can 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 for a range of distances, e.g. for $d = 0$ to 200, i.e. $\min_{E_{ij}} \int_0^{200} F(t)dt$.
- the minimax for a range of distances, e.g. for $d = 0$ to 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.

3.3 Degrees of Freedom

Because our model of evolution is symmetrical, i.e. it does not distinguish a particular sequence, 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 we have to eliminate these two degrees of freedom. Most of the time this can be done by arbitrarily setting two values of E, e.g. $E_{11} = -1$ and $E_{12} = 1$.

4 Example

We use a random matrix M as an example. For simplicity we will assume that we have only 4 amino acids, i.e. M will be a 4×4 matrix. M is generated to satisfy the normal properties: the columns should add to 1 and the pseudo-symmetry condition $f_i M_{ji} = f_j M_{ij}$ must hold.

$$M = \begin{bmatrix} 0.9920 & 0.007612 & 0.001144 & 0.004677 \\ 0.003761 & 0.9819 & 0.002670 & 0.003118 \\ 0.001410 & 0.006660 & 0.9931 & 0.006235 \\ 0.002821 & 0.003806 & 0.003051 & 0.9860 \end{bmatrix}$$

The frequency vector f can be computed from the eigenvector of M whose eigenvalue is 1 or from the first row/column of M [1].

$$f_i = \frac{M_{i1}}{M_{1i} \sum_j M_{j1}/M_{1j}}$$

$$f_1 = .3003 \quad f_2 = .1484 \quad f_3 = .3702 \quad f_4 = .1811$$

The eigenvalues of M are

$$\lambda_1 = 1.0000 \quad \lambda_2 = .9914 \quad \lambda_3 = .9830 \quad \lambda_4 = .9786$$

In this case we want to find the optimal scoring matrix for d in the range $100 \leq d \leq 200$. After setting $E_{11} = -1$ and $E_{12} = 1$ and numerically minimizing $\int_{100}^{200} F(t)dt$ on the rest of the E_{ij} unknowns, we obtain:

$$E = \begin{bmatrix} -1.0 & 1.0 & 3.522 & 1.720 \\ 1.0 & -1.379 & 1.604 & 1.331 \\ 3.522 & 1.604 & -0.6688 & 1.601 \\ 1.720 & 1.331 & 1.601 & -1.596 \end{bmatrix}$$

From these values, using equation 3, we can compute

$$S(d) = 1.258 - .3307(.9786)^d - 1.459(.9914)^d - .5098(.9830)^d$$

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

$$S'(d) = .007136(.9786)^d + .01264(.9914)^d + .008734(.9830)^d$$

and we can see that all the terms are strictly positive, so $S(d)$ is strictly monotonic for all d.

Suppose we have aligned these two sequences:

$$\begin{array}{c} A_3 A_2 A_1 A_2 A_4 A_3 A_4 A_2 \\ A_3 A_4 A_1 A_3 A_1 A_1 A_2 A_2 \end{array}$$

The score for this alignment is

$$w/n = (E_{33} + E_{24} + E_{11} + E_{23} + E_{41} + E_{31} + E_{42} + E_{22})/8 = .8076$$

From this value, and using any numerical method, we can find the d^* which satisfies the equation

$$S(d^*) = .8076$$

The solution is

$$d^* = 149.99$$

The variance of the estimator is $\frac{S^2(d) - S(d)^2}{nS'(d)^2} \approx 16215.0$. The variance is very large, and so is the standard deviation (127.34), 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] Margaret O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model for evolutionary change in proteins. In Margaret O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, pages 345–352. 1978.

[4] J. Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Amer. J. Human Genetics*, 25:471–492, 1973.

[5] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27:27 – 33, 1978.

[6] J. Felsenstein. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods in Enzymology*, 266:418 – 427, 1996.

[7] W.M. Fitch and E. Margoliash. The construction of phylogenetic trees. *Science*, 155:279 – 84, 1967.

[8] Gaston H. Gonnet and Steven A. Benner. Probabilistic ancestral sequences and multiple alignments. In *Fifth Scandinavian Workshop on Algorithm Theory, Reykjevik July 1996*, 1996.

[9] Gaston H. Gonnet, Mark A. Cohen, and Steven A. Benner. Exhaustive matching of the entire protein sequence database. *Science*, 256:1443–1445, 1992.

[10] J. Hein. An optimal algorithm to reconstruct trees from additive distance data. *Bull. Math. Biol.*, 51:597 – 603, 1989.

[11] P. Hogeweg and P. Hesper. The alignmen of sets of sequences and the construction of phylogenetic trees: an integrated method. *J. Mol. Evol.*, 20:175 –86, 1988.