RESEARCH ARTICLES

Different Versions of the Dayhoff Rate Matrix

Carolin Kosiol and Nick Goldman

EMBL-European Bioinformatics Institute, Hinxton, United Kingdom

Many phylogenetic inference methods are based on Markov models of sequence evolution. These are usually expressed in terms of a matrix (Q) of instantaneous rates of change but some models of amino acid replacement, most notably the PAM model of Dayhoff and colleagues, were originally published only in terms of time-dependent probability matrices ($P(t)$). Previously published methods for deriving Q have used eigen-decomposition of an approximation to $P(t)$. We show that the commonly used value of t is too large to ensure convergence of the estimates of elements of Q. We describe two simpler alternative methods for deriving Q from information such as that published by Dayhoff and colleagues. Neither of these methods requires approximation or eigen-decomposition. We identify the methods used to derive various different versions of the Dayhoff model in current software, perform a comparison of existing and new implementations, and, to facilitate agreement among scientists using supposedly identical models, recommend that one of the new methods be used as a standard.

Introduction

Dayhoff and colleagues (Dayhoff, Eck, and Park 1972; Dayhoff, Schwartz, and Orcutt 1978[1]) introduced a Markov model of protein evolution that resulted in the development of the widely used amino acid replacement matrices known as the PAM matrices. In these articles, the protein evolution model is only given in the form of probability matrices relating to specific levels of sequence divergence (e.g., PAM1 and PAM250 in DSO78). In many phylogenetic inference methods (including pairwise distance estimation, maximum likelihood phylogeny estimation and Bayesian analysis of phylogeny), however, it is necessary to be able to compute this probability matrix (which we denote $P(t)$) for any real evolutionary time (distance) $t \geq 0$ (Felsenstein 2003). This is achieved using an instantaneous rate matrix (IRM), often denoted $Q = (q_{ij})_{i,j=1,\ldots,20}$, which is related to $P(t)$ via

$$P(t) = \exp(tQ) = I + tQ + \frac{(tQ)^2}{2!} + \frac{(tQ)^3}{3!} + \cdots \qquad (1)$$

(Liò and Goldman 1998). The probability matrix $P(t)$ for any time $t > 0$ is fully determined by the IRM Q, which is itself independent of t.

In this article we summarize different methods that have been used to construct IRMs from the probability matrices of DSO78, identify their implementations in standard phylogenetic software packages, and indicate ways in which they lead to different and potentially inaccurate IRMs. We describe two simpler methods for deriving an IRM from information such as that published by Dayhoff and collaborators. We then compare the performance of all of these implementations using a test set of 200 protein domain families.

[1] For brevity, we generally refer to these publications as the work of "Dayhoff and colleagues." The latter, abbreviated to "DSO78," is better known and contains the data used for all implementations of the Dayhoff model, but we note that all important methodology was introduced in the former.

Key words: amino acid replacement, Dayhoff matrix, Markov models, phylogenetic inference, protein evolution.

E-mail: goldman@ebi.ac.uk.

Mol. Biol. Evol. 22(2):193–199. 2005
doi:10.1093/molbev/msi005
Advance Access publication October 13, 2004

The Dayhoff model and other similar models (e.g., the JTT model of Jones, Taylor, and Thornton 1992) have been very influential in molecular phylogenetics, database searching, and other fields, and they continue to be widely and regularly used. It is important to have a complete understanding of how these models can be accurately derived from raw data. In the interests of (1) remaining faithful to the information collected and published by Dayhoff and colleagues and others, (2) keeping computations as simple and accurate as possible, and (3) facilitating agreement among scientists using different implementations of supposedly identical methods, we propose a standardization of these IRMs in phylogenetic software.

Methods

Markov Models for Protein Evolution

Proteins are sequences of amino acids. The Markov model asserts that one protein sequence is derived from another during evolution by a series of independent substitutions, each changing one amino acid in the ancestral sequence to another in its descendant. We assume independence of evolution at different sites.

The continuous-time Markov process is a stochastic model in which $(P(t))_{ij}$ gives the probability that amino acid i will change to amino acid j at any single site after any time $t \geq 0$. Since there are 20 amino acids, i and j take the values 1, ... ,20. The 20×20 probability matrix is calculated as $P(t) = \exp(tQ)$ (eq. 1), where the matrix Q is independent of time in the Markov processes typically used in molecular phylogenetics. Q has off-diagonal entries $q_{ij,i\neq j}$ equal to the instantaneous rates of replacement of i by j, and diagonal entries defined by the mathematical requirement that each row sum is 0.

Associated with a matrix Q are equilibrium frequencies of the 20 amino acids, denoted f_i, and mutabilities, m_i, defined as the rate at which amino acid i changes to any other amino acid: $m_i = \sum_{j\neq i} q_{ij}$. Typically in phylogenetic applications, Q is normalized so that the mean rate of replacement at equilibrium ($\sum_i \sum_{j\neq i} f_i q_{ij}$ or $\sum_i f_i m_i$) is 1, meaning that times t are measured in units of expected numbers of changes per site. Here, we simplify our notation by omitting the trivial multiplicative constants

that achieve this normalization. More detailed descriptions of Markov models in molecular phylogenetics are given by Liò and Goldman (1998), Felsenstein (2003), and Thorne and Goldman (2003).

Dayhoff and colleagues devised a method to estimate $P(t)$ that relied on the comparison of closely related pairs of aligned sequences. They selected pairs of sequences that were 85% identical and counted the differences between them. In general these differences underestimate the actual numbers of changes because when multiple changes occur at a single site they are observed as at most single replacements. However, if sufficiently closely related sequences are used, then the probability of these "multiple hits" is reduced. Dayhoff and colleagues assumed that at 85% sequence identity there are no multiple hits (see also Jones, Taylor, and Thornton 1992; Goldman, Thorne, and Jones 1996, 1998). The cost of this 85% rule is some loss in accuracy and inefficient use of data, because divergent sequence pairs have to be discarded.

Advances in methodology and computer speed have made it possible to estimate Markov models of amino acid replacement without any constraints on the levels of divergence between the sequences (e.g., Adachi and Hasegawa 1996; Yang, Nielsen, and Hasegawa 1998; Adachi et al. 2000; Müller and Vingron 2000; Whelan and Goldman 2001; Devauchelle et al. 2001; Dimmic et al. 2002; Veerassamy, Smith, and Tillier 2003). While these methods may now give superior results, less sophisticated methods based on simple counts of changes under the 85% rule are still used (e.g., Goldman, Thorne, and Jones 1996, 1998) and were the basis for two of the most widely used evolutionary models in molecular phylogenetics (DSO78; Jones, Taylor, and Thornton 1992). The rest of our analysis concentrates solely on cases such as these, where differences between sequence pairs can be counted and assumed to accurately represent actual evolutionary events.

We write the observed number of occurrences of sites in all aligned sequence pairs with amino acids i in one sequence and j in the other as n_{ij}. The data collection technique of Dayhoff and colleagues leads to the relationships $n_{ij} = n_{ji}$, because the direction of an evolutionary change cannot be determined from the observation of two contemporary sequences. (As a consequence, equilibrium is assumed and resulting models must be reversible; see Liò and Goldman 1998.) Because we assume the n_{ij} accurately represent all evolutionary events, an intuitive estimate for the rate of change of i to j ($j \neq i$) is given by the number of such events as a proportion of all observations of i:

$$q_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}. \tag{2}$$

Corresponding estimators for the m_i and f_i are given by

$$m_i = \frac{\sum_{j \neq i} n_{ij}}{\sum_k n_{ik}} \tag{3}$$

and

$$f_i = \frac{\sum_j n_{ij}}{\sum_k \sum_l n_{kl}} \tag{4}$$

(see also Goldman, Thorne, and Jones 1996).

Equation (2) represents a simple way to estimate the IRM Q directly from counts of observed differences between closely related sequence pairs and was used by Goldman, Thorne, and Jones (1996, 1998). Applications of the models of Dayhoff and colleagues and Jones, Taylor, and Thornton (1992) in molecular phylogentics have, however, used a different approach, estimating Q not directly but via corresponding probability matrices $P(t)$. We proceed by describing these methods.

The Dayhoff Model

Dayhoff and colleagues published probability matrices based on counts of sequence differences, frequencies, and mutabilities. These articles include the values of the counts n_{ij}, but in an incomplete manner: the numbers of positions containing amino acid i in both sequences, n_{ii}, are omitted.

The mutabilities m_i were calculated slightly differently from equation (3) above. The data from individual sequence pairs $s = 1, \ldots, S$ were considered separately, and the estimator

$$\begin{aligned} m_i &= \frac{\sum_{s=1}^{S} \left(\sum_{j \neq i} n_{ij}^{(s)} \right)}{\sum_{s=1}^{S} \left(\sum_g \sum_{h \neq g} n_{gh}^{(s)} / N_s \right) \left(\sum_k n_{ik}^{(s)} \right)} \\ &= \frac{\sum_{j \neq i} n_{ij}}{\sum_{s=1}^{S} \left(\sum_g \sum_{h \neq g} n_{gh}^{(s)} / N_s \right) \left(\sum_k n_{ik}^{(s)} \right)} \end{aligned} \tag{5}$$

used. Here, the observed counts from each sequence pair s are denoted $n_{ij}^{(s)}$ and N_s is the number of sequence positions in pair s (so $\sum_s n_{ij}^{(s)} = n_{ij}$ and $\sum_{i,j} n_{ij}^{(s)} = N_s$). Note that the numerators of the estimators given by equations (3) and (5) are the same. The denominators differ only by the use of the weighting factor $\sum_g \sum_{h \neq g} n_{gh}^{(s)} / N_s$ for each sequence pair s. Dayhoff and colleagues described this as an estimator of the "exposure to evolution" or evolutionary separation of each sequence pair. There is no reason why this factor is necessary, as the evolutionary separation of pair s does not make the changes $n_{ij}^{(s)}$ any more or less indicative of typical protein evolution than those observed in any other pair. Nevertheless, the weighting factor introduces no systematic bias (it will not be correlated with the relative values of the $\sum_k n_{ik}^{(s)}$ term it is applied to) and analysis (not shown) of the expectations of the numerators and denominators of equations (3) and (5) indicates that both ratios give reasonable estimators of mutability.

Dayhoff and colleagues used much the same idea to estimate the frequencies f_i of the amino acids, combining the data for each comparison s to derive the following estimator:

$$\begin{aligned} f_i &= \frac{\sum_{s=1}^{S} \left(\sum_g \sum_{h \neq g} n_{gh}^{(s)} / N_s \right) \left(\sum_j n_{ij}^{(s)} \right)}{\sum_{s=1}^{S} \left(\sum_k \sum_l n_{kl}^{(s)} \right)} \\ &= \frac{\sum_{s=1}^{S} \left(\sum_g \sum_{h \neq g} n_{gh}^{(s)} / N_s \right) \left(\sum_j n_{ij}^{(s)} \right)}{\sum_k \sum_l n_{kl}}. \end{aligned} \tag{6}$$

This agrees with equation (4) above, again apart from the unneeded weighting factor $\sum_g \sum_{h \neq g} n_{gh}^{(s)} / N_s$. Dayhoff and colleagues published the values they computed for the

m_i and f_i. It is not possible to work out Dayhoff and colleagues' missing n_{ii} values from their published $n_{ij,i\neq j}$, m_i and f_i.

Dayhoff and colleagues described an estimator $P_D = (p_D)_{ij}$ for probability matrices as follows:

$$(p_D)_{ij} = \begin{cases} \dfrac{cm_i n_{ij}}{\sum_{k\neq i} n_{ik}} & \text{if } i \neq j \\ 1 - cm_i & \text{if } i = j \end{cases} \quad (7)$$

where c is a constant. The quantity $\delta = 1 - \sum_i f_i (p_D)_{ii}$ then gives the proportion of amino acids that are observed to differ after the evolutionary interval represented by the matrix P_D and depends on the choice of c:

$$\delta = 1 - \sum_i f_i (p_D)_{ii} = 1 - \sum_i f_i (1 - cm_i)$$
$$= 1 - \sum_i f_i + c \sum_i f_i m_i = c \sum_i f_i m_i. \quad (8)$$

To create the PAM1 matrix, Dayhoff and collaborators chose c so that δ is 1 observed mutation per 100 sites. Thus

$$c = \frac{\delta}{\sum_i f_i m_i} = \frac{0.01}{\sum_i f_i m_i}, \quad (9)$$

and we may write PAM1 $= P_D(\delta = 0.01)$. Dayhoff and colleagues also defined matrices PAMn, equal to (PAM1)^n (*not* $P_D(\delta = n/100)$).

The "time" at PAM1 is defined by $\delta = 1$ observed amino acid change per 100 amino acids, i.e., excluding changes that are unobservable because of multiple hits. Therefore, PAM1 does not correspond to a probability matrix $P(0.01)$ calculated according to equation (1), but to some $P(0.01 + \epsilon)$, where ϵ is the extra time needed to account for unobserved substitutions. For small times t, the difference is negligible (e.g., PAM1 approximates $P(0.01 + 6.85 \times 10^{-5})$ when using the IRM derived with the DCMut method described below); even for larger times, the difference is still small (PAM250 is close to $P(2.52)$ under the DCMut model).

Dayhoff and colleagues' estimator embodied in equations (7)–(9) makes two assumptions: that the data from which the observed counts n_{ij} are derived are taken to be sufficiently closely related to exclude multiple hits (the 85% rule), and that there will be no multiple hits in the probability model defined by P_D. The latter assumption is approximately true for $\delta = 0.01$, and becomes more accurate as δ decreases. For larger values of δ it is less accurate; if δ is large enough, the resulting matrix P_D will not even be a valid probability matrix (see fig. 1).

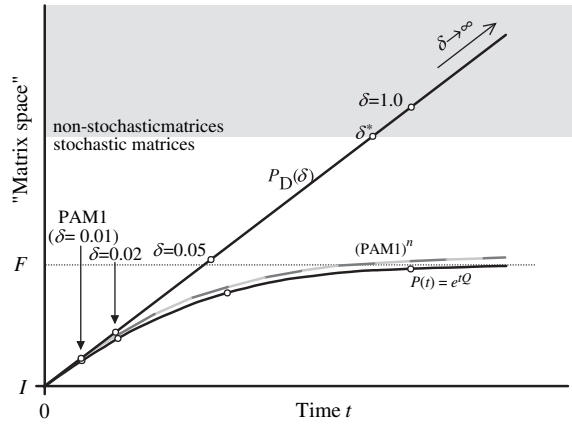

FIG. 1.—Schematic diagram showing relationships of matrices described in the text. The y-axis represents the unbounded 400-dimensional space of 20 × 20 matrices, with the region of valid stochastic matrices (all elements $\in$ [0,1] and all row sums = 1; a bounded 380-dimensional subspace) and its complement (matrices that are not stochastic) indicated. Time t measures evolutionary distance. The lowest (smooth) curve represents matrices $P(t)$ describing amino acid replacement probabilities generated from a particular instantaneous rate matrix Q according to $P(t) = \exp(tQ)$. When $t = 0$, no replacements have occurred and $P(t)$ equals the identity matrix I. As $t \to \infty$, $P(t)$ converges to F, the matrix with each row equal to the equilibrium distribution of Q (i.e., $F_{ij} = f_j$). Q can be recovered from $P(t)$ for any time t, using the procedure described by Kishino, Miyata, and Hasegawa (1990; equation (10) in this article). The uppermost (straight) line indicates the matrices $P_D(\delta)$ generated by the procedure of Dayhoff and colleagues, embodied in equations (7)–(9), which is "linear" in the sense that it assumes that no multiple hits occur. For sufficiently small values of δ, $P_D(\delta)$ is a reasonable approximation to $P(t)$ and the method of Kishino, Miyata, and Hasegawa (1990) applied to $P_D(\delta)$ can give a close approximation to Q (see *From Probability to Rate Matrix via Eigen-Decomposition*). Alternatively, *any* matrix on the uppermost line (i.e., $P_D(\delta)$ for any δ) can be used to recover Q using the method embodied in equation (14), including matrices $P_D(\delta)$, which are not stochastic matrices (there is a value $\delta^* < 1$ such that at least one diagonal element of $P_D(\delta)$ is < 0 for all $\delta \geq \delta^*$). Dayhoff and colleagues' PAM1 matrix lies on the uppermost line and corresponds to $\delta = 0.01$. Matrices denoted PAMn ($n > 1$) by Dayhoff and colleagues are defined to equal (PAM1)^n; consequently, each remains close to $P(n/100)$. This is illustrated by the middle line, which shows (PAM1)^n for increasing powers of n (note the piecewise linearity of this line, shown in alternating lighter and darker shades, which represents n increasing in integer steps). A color version of this graph is available in the Supplementary Material online.

From Probability to Rate Matrix via Eigen-Decomposition

Kishino, Miyata, and Hasegawa (1990) described a method to obtain an IRM matrix from the probability matrix PAM1. The key idea of this method is the relationship of the eigenvalues and eigenvectors of matrices $P(t)$ and Q related by equation (1). If P has eigenvalues ρ_i and eigenvectors u_i ($i = 1, \ldots, 20$), and defining also $\lambda_i = \log \rho_i$, $U = (u_1, \ldots, u_{20})$ and $\Lambda = \text{diag}\ (\lambda_1, \ldots, \lambda_{20})$ (the diagonal matrix with entries λ_i), then the IRM Q can be calculated as

$$Q = U \Lambda U^{-1}. \quad (10)$$

This method is appropriate for recovering Q if P is generated according to $P(t) = \exp(tQ)$ for any specific t. However, Dayhoff and colleagues' approach (equations 7–9) generates a matrix $P_D(\delta)$ which is only an approximation to $P(t)$ because multiple hits occurring over the time period corresponding to δ are neglected. (Indeed, the PAM1 matrix cannot be generated as $\exp(t^*Q^*)$ for any valid IRM Q^* and time $t^* \geq 0$—proof not shown.) This approximation is increasingly poor as δ increases (fig. 1). Kishino, Miyata, and Hasegawa (1990) adopted $\delta = 0.01$, to generate PAM1 as in DSO78. It turns out that this value of δ is not small enough. Using this eigen-decomposition method, we have calculated the IRMs Q for δ in the range $[3 \times 10^{-7}, 3 \times 10^{-1}]$, and figure 2 shows that these can

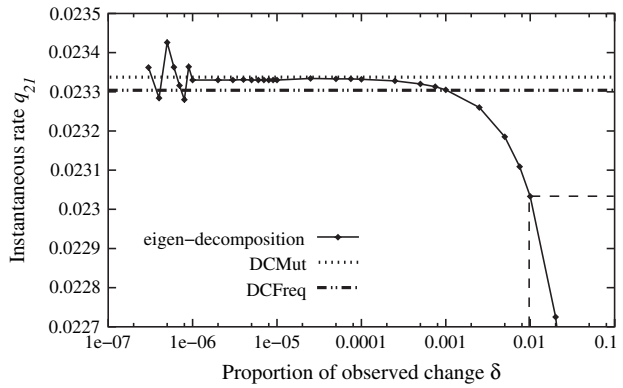

FIG. 2.—Element q_{21} (Arg → Asn) of the IRM computed by the eigen-decomposition method of Kishino, Miyata, and Hasegawa (1990) applied to the PAM1 matrix, and by the DCMut and the DCFreq methods. For the eigen-decomposition method, the value of q_{21} depends on δ; note the numerical instability we encountered for $\delta < 10^{-6}$. Also indicated is the value of q_{21} derived using $\delta = 0.01$; clearly, this does not attain the convergent behavior observed for $10^{-6} < \delta < 10^{-4}$. Note that in general the values of q_{ij} from the converged eigen-decomposition and DCMut methods do not agree as closely as observed in this case. A color version of this graph is available in the Supplementary Material online.

suffer from two convergence problems. On the one-hand, if δ is too large, elements q_{ij} may not yet be converged. On the other hand, choosing δ too small may cause the numerical eigen-analysis that finds the ρ_i and u_i to become unstable. Thus it is necessary to check convergence for each of the $19 \times 20 = 380$ values of $q_{ij,i\neq j}$. We find that many of the q_{ij} computed by the method of Kishino, Miyata, and Hasegawa (1990) using $\delta = 0.01$ do not attain converged values. Therefore, the IRM derived this way does not give a fully accurate representation of the amino acid replacement data collected by DSO78.

Direct Derivation of the Rate Matrix Using Mutabilities or Frequencies

It would be preferable to compute the elements of the IRM Q directly, and so avoid both relatively complex eigen-decomposition and lengthy convergence analysis to check that a suitable value (or values) of δ were used. The method of equation (2), as used by Goldman, Thorne, and Jones (1996, 1998), would be suitable, but this requires knowledge of the n_{ii} and these were not published by Dayhoff and colleagues. We now present two direct ways to estimate Q using only the information given by Dayhoff and colleagues, i.e., the observed changes $n_{ij,i\neq j}$, the mutabilities m_i and the frequencies f_i.

We call the first method *Direct Computation with Mutabilities* (DCMut), because it uses only the observed changes and the mutabilities. Rearranging equations (2) and (3):

$$q_{ij} = \frac{m_i n_{ij}}{\sum_{k\neq i} n_{ik}} \tag{11}$$

for $i \neq j$ (as usual, $q_{ii} = -\sum_j q_{ij}$).

The second method, *Direct Computation with Frequencies* (DCFreq), relies only on the observed changes and the frequencies and is defined by rearranging Equations (2) and (4):

$$q_{ij} = \frac{n_{ij}}{f_i \sum_k \sum_l n_{kl}}. \tag{12}$$

Since $\sum_k \sum_l n_{kl}$, the total number of amino acid sites in the sequence data, is a constant, it may be ignored because of the subsequent normalization of Q to have mean rate 1 (see above). Therefore, we may simply write:

$$q_{ij} = \frac{n_{ij}}{f_i}. \tag{13}$$

Note that both the DCMut and DCFreq derivations of IRMs require neither the consideration of any limit $\delta \to 0$ nor matrix eigen-analysis. They do, however, still require that the sequence pairs from which the n_{ij} are derived should be closely related (e.g., satisfying the 85% rule). Although there are many ways to calculate $P(t) = \exp(tQ)$ (Moler and Van Loan 2003), the most popular method in molecular phylogenetics uses eigen-decomposition (Liò and Goldman 1998) and thus even with the DCMut and DCFreq approaches to deriving Q we do not avoid eigen-analysis in its use.

Had Dayhoff and colleagues calculated mutabilities and frequencies according to equations (3) and (4), the IRMs given by equations (2), (11), and (13) would be identical. Likewise, if δ approaches zero then the Kishino, Miyata, and Hasegawa (1990) method applied to P_D (eq. 10) would also converge to this rate matrix. But because Dayhoff and colleagues estimated the mutabilities and frequencies incorporating the weighting factors described above (eq. 5 and eq. 6), we expect slightly different IRMs. Figure 2 shows the element q_{21} (Arg → Asn) calculated according to equations (10), (11), and (13). Although very close, the values are not identical. To facilitate agreement among scientists wishing to use supposedly identical models, we would like to be able to suggest one standard implementation of the model of DSO78 for molecular phylogenetics. In the *Results and Discussion* we perform a comparison of the different IRMs' performance in practice and propose a new standard.

Results and Discussion

We have identified different versions of the Dayhoff model described in the literature and used in current phylogenetic software packages, and implemented all of these variants in the *codeml* program of the PAML package (Yang 1997).[2] We distinguish between six different implementations, KMH, Paml, Proml, Molphy, DCMut, and DCFreq, and have assessed their impact on phylogenetic analysis by maximum likelihood.

KMH refers to the IRM calculated according to the eigen-decomposition method described by Kishino, Miyata, and Hasegawa (1990) applied to PAM1 of DSO78. No widely used software packages currently use exactly this implementation. We find that a number of off-diagonal

[2] Files prepared following the pattern of the distributed files (e.g.) dayhoff.dat and jones.dat are available from http://www.ebi.ac.uk/goldman/dayhoff.

elements of Q are < 0, meaning that this is not strictly a valid IRM (see the earlier section *From Probability to Rate Matrix via Eigen-Decomposition*). However, this implementation ran without problems in *codeml*, and so we made no further alterations.

The IRM for the Dayhoff model distributed with the PAML package, version 3.13d, (Yang 1997) is calculated using the Kishino, Miyata, and Hasegawa (1990) eigen-decomposition method applied to PAM1 (as for KMH above). The rate matrix $Q = (q_{ij})$ is then uniquely decomposed into $q_{ij}=f_j \times s_{ij}$ (see, e.g., Whelan and Goldman 2001), where the f_j are the equilibrium frequencies derived from this Q (and which thus differ minutely from those published by DSO78). The exchangeabilities s_{ij} are multiplied by approximately 100, rounded, and stored as integers. All except one of the $s_{ij} < 0$ become equal to 0 after this rounding. Following Yang (1997), we replace the rounded value of -1 for $s_{\mathrm{Glu} \rightarrow \mathrm{Arg}}$ with $+1$. We refer to this version of the Dayhoff model as the Paml implemention; it is also used in the *MrBayes* v. 3.0 (Ronquist and Huelsenbeck 2003) and *Phyml* v. 2.4 (Guindon and Gascuel 2003) programs.

The *proml* program distributed with version 3.6 of Felsenstein's PHYLIP package (Felsenstein 2002) includes an implementation of the Dayhoff model based on an IRM Q_{F} defined by

$$Q_{\mathrm{F}} = U \cdot \mathrm{diag}(\rho_{\mathrm{i}} - 1) \cdot \mathrm{U}^{-1}, \qquad (14)$$

where the ρ_i and u_i are the eigenvalues and eigenvectors, respectively, of a probability matrix P_{F} (Joe Felsenstein, pers. comm.). This means that $Q_{\mathrm{F}} = P_{\mathrm{F}} - I$, where I is the identity matrix. Felsenstein (2002) uses a matrix P_{F} based on the published DSO78 PAM1 matrix (not the original counts n_{ij}), modified to ensure reversibility (Felsenstein 1996; reversibility is violated by the originally published PAM1 matrix because of rounding errors). This implementation is thus an approximation to the DCMut method since $Q_{\mathrm{F}} = P_{\mathrm{F}} - I \approx P_{\mathrm{D}} - I$, which is proportional to the matrix defined by equation (11).

Early versions of the *protml* program of the MOLPHY package (Adachi and Hasegawa 1992) used the Kishino, Miyata, and Hasegawa (1990) eigen-decomposition of PAM1; later versions (e.g., v. 2.3b3) use the DCFreq approach, suggested to the authors of MOLPHY by Korbinian Strimmer and Arndt von Haeseler (Korbinian Strimmer, pers. comm.). However, the counts n_{ij} used were changed from those in DSO78: where zeros occurred, they were substituted by scaled values taken from Jones, Taylor, and Thornton (1992). This makes the later MOLPHY implementation, also used in the TREE-PUZZLE v. 5.2 package (Schmidt et al. 2002) and in the sequence simulation program *PSeq-Gen* v. 1.1 (Grassly, Adachi, and Rambaut 1997), a hybrid between the Dayhoff and JTT models.

The DCMut method has never been used before. We implemented the Dayhoff model according to this method, using data from DSO78. We have also calculated a DCFreq implementation, again using exactly the data from DSO78.

To get an idea of the impact of the different versions of the Dayhoff model on phylogenetic analysis, we performed a small test. We calculated the maximum likelihoods (Felsenstein 2003) for 200 protein families and their associated tree topologies taken from release 7.6 of the PANDIT database (Whelan, de Bakker, and Goldman 2003) using the implementations described above. Equilibrium frequencies were either taken from the appropriate implementation of the Dayhoff model (analyses labeled KMH, Paml, Proml, Molphy, DCMut, DCFreq), or were estimated separately for each protein family and incorporated using the "+F" method of Cao et al. (1994; see also Thorne and Goldman 2003; analyses KMH+F, Paml+F, Proml+F, Molphy+F, DCMut+F, DCFreq+F). (Note that it is not currently possible to use the Proml+F model in the *proml* program of Felsenstein's PHYLIP package. Note also the importance of re-normalizing Q so that the mean rate of change at equilibrium equals 1 when using the +F method, particularly if different software is used for simulating sequence data and then analyzing that simulated data.) Table 1 summarizes the results.

Table 1
Relative Success of 12 Implementations of the Dayhoff Matrix over a Test Set of 200 Protein Domain Families

Implementation	Median Rank	Interquartile Range of Ranks
Molphy+F	1	[1, 2]
DCMut+F	3	[3, 4]
DCFreq+F	3	[2, 5]
Paml+F	4	[4, 5]
Proml+F	6	[4, 6]
KMH+F	6	[5, 6]
Molphy	7	[7, 7]
DCMut	8	[8, 9]
DCFreq	8	[8, 10]
Paml	10	[9, 10]
KMH	11	[10, 11]
Proml	12	[8, 12]

NOTE.—The 200 amino acid sequence alignments used and log-likelihoods attained under all 12 models are listed in the Supplementary Material online.

For each of the 200 protein families, the 12 model versions were ranked according to their maximum likelihood values. This led to each model version being assigned 200 ranks—potentially from 1 to 12—according to its relative performance for the 200 protein families, and we report the medians and interquartile ranges of these ranks. First, as would be expected, we note that every implementation performs better in its "+F" version. Second, irrespective of whether we consider +F versions or not, we find the six basic implementations are ordered (best to worst): Molphy, DCMut, DCFreq, Paml, KMH $\approx$ Proml (the KMH and Proml models fared approximately equally poorly). Differences in actual maximum likelihoods among the six +F implementations were up to 102.82 log-likelihood units (median 2.28; inter-quartile range 0.77–6.36) over the 200 alignments analyzed; these differences may be of the order of the differences between competing topologies. For the six implementations without the +F modification, the corresponding differences were up to 87.67 log-likelihood units (median 2.28; interquartile range 0.79–6.82). Full results of this experiment are available in the Supplementary Material online.

By modern standards not much data was available to Dayhoff and colleagues, and their published values of n_{ij}

include a number of zeros. It seems likely that these give significant underestimates of the frequencies of replacement between certain amino acids, and we attribute the Molphy implementation's success to its hybrid nature, incorporating information from the counts published by Jones, Taylor, and Thornton (1992) to redress this underestimation. Among versions based solely on the information published by DSO78, the DCMut and DCFreq versions have advantages in computational ease and seem to give a small advantage in terms of maximum likelihood scores over the test data sets studied. The versions based on eigen-decomposition of PAM1 (i.e., with $\delta = 0.01$, which does not guarantee convergence) give the worst performance and seem to have no advantages. It is interesting to note that the Paml implementation fares better than KMH; evidently the variations introduced by the rounding procedure in the Paml version, in combination with the usage of a non-converged eigen-decomposition method and the sparse data published by DSO78, give a model that is no worse, and may even be better, than the version without rounding.

Conclusions

Methodological advances, increased database sizes, and faster computers now permit the inference of potentially superior Markov process models for protein sequence evolution. Even so, it is still valuable to consider simpler methods which are appropriate for making inferences from pairwise comparisons of closely related proteins. These methods may become more important as sequencing projects complete the proteomes of closely related species such as human and chimpanzee (International Human Genome Sequencing Consortium 2001; International Chimpanzee Chromosome 22 Consortium 2004), mouse and rat (Mouse Genome Sequencing Consortium 2002; Rat Genome Sequencing Project Consortium 2004), and many others. Indeed, even within-species studies will become possible, for example using SNPs (International SNP Map Working Group 2001).

We have summarized different approaches (including two new ones) to calculating an instantaneous rate matrix from the (incomplete) information given by Dayhoff and colleagues. In practice, the differences are small but may be non-trivial. All the implementations studied are valid models of protein sequence evolution, and they may be applied to any protein sequence alignments. Nevertheless, while this and similar models remain of interest in molecular phylogenetics and other fields and for the sake of consistency, particularly among investigators developing models and software implementations, we suggest that it is of value to identify a "standard" implementation for the model of Dayhoff, Schwarz, and Orcutt (1978).

Although the Molphy implementation (Adachi and Hasegawa 1992) performs well, it uses a hybrid data set based on both Dayhoff, Schwarz, and Orcutt (1978) and Jones, Taylor, and Thornton (1992). The Proml implementation also uses the Dayhoff, Schwarz, and Orcutt (1978) data in a modified form. The Proml, Paml, and KMH implementations perform least well in our maximum likelihood implementation experiment, and the generation of their instantaneous rate matrices is based on relatively complex eigen-decomposition and, in the cases of Paml and KMH, exhibits convergence problems. Therefore, looking for standardization based only on the data published by Dayhoff, Schwarz, and Orcutt (1978), we propose the adoption of the DCMut matrix, which is straightforward to calculate and performed well in our test. Adopting DCMut also means that no decision has to be made regarding the fact that, because of rounding errors, the equilibrium frequencies published by Dayhoff, Schwarz, and Orcutt (1978) do not sum to 1. We do not suggest that DCMut is a better model than any other implementation described here, but only that it is a reasonable choice when a single model is required for comparing results across different analyses, computer programs, or data sets.

Jones, Taylor, and Thornton (1992) used much the same methodology as Dayhoff and colleagues, but based on a larger sequence database. Again, only a probability matrix, mutabilities, frequencies and incomplete counts are provided in their original article. The JTT probability matrix has also been used to compute an IRM for use in phylogenetics and has been implemented (again, in a variety of ways) in the MOLPHY, *MrBayes*, PAML, PHYLIP, *Phyml, PSeq-Gen*, and TREE-PUZZLE software. We again suggest that an implementation based on the DCMut method be adopted as standard.[3]

We hope to persuade software developers to agree to include our DCMut implementations of the Dayhoff and JTT models into their programs, and we are happy to discuss providing the necessary data in whatever forms required.[4]

Acknowledgments

Thanks go to Jeff Thorne, for assistance with understanding Dayhoff and colleagues' calculation of relative mutabilities, and to Joe Felsenstein, Korbinian Strimmer, and Elisabeth Tillier for helpful discussions of their implementations of the Dayhoff and JTT models. C.K. is supported by a Wellcome Trust Prize Studentship and is a member of Wolfson College, University of Cambridge. N.G. is supported by a Wellcome Trust Fellowship in Basic Biomedical Research.

Literature Cited

Adachi, J., and M. Hasegawa. 1992. MOLPHY version 2.3: Programs for Molecular Phylogenetics Based on Maximum Likelihood. Computer Science Monographs 28, Institute of Statistical Mathematics, Tokyo. http://www.ism.ac.jp/software/ismlib/softother.e.html#molphy

———. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. **42**:459–468.

[3] A file containing this matrix, suitable for use in the *codeml* program of the PAML package, is available from http://www.ebi.ac.uk/goldman/dayhoff.

[4] The authors of *MrBayes* (Ronquist and Huelsenbeck 2003), PAML (Yang 1997), PAUP* (Swofford 2002), *Phyml* (Guindon and Gascuel 2003), *Seq-Gen* (Rambaut and Grassly 1997; the successor to *PSeqGen*) and TREE-PUZZLE (Schmidt et al. 2002) have agreed to incorporate the DCMut versions of the Dayhoff and JTT IRMs in future releases of their software.

Adachi, J., P. J. Waddell, W. Martin, and M. Hasegawa. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. **50**:348–358.

Cao, Y., J. Adachi, A. Janke, S. Pääbo, and M. Hasegawa. 1994. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J. Mol. Evol. **39**: 519–527.

Dayhoff, M. O., R. V. Eck, and C. M. Park. 1972. A model of evolutionary change in proteins. Pp. 89–99 *in* M. O. Dayhoff, ed., Atlas of Protein Sequence and Structure Vol. 5. National Biomedical Research Foundation, Washington, D.C.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. Dayhoff, ed., Atlas of Protein Sequence and Structure Vol. 5, suppl. 3. National Biomedical Research Foundation, Washington, D.C.

Devauchelle, C., A. Grossmann, A. Hénaut, M. Holschneider, M. Monnerot, J. L. Riesler, and B. Torrésani. 2001. Rate matrices for analyzing large families of protein sequences. J. Comp. Biol. **8**:381–399.

Dimmic, M. W., J. S. Rest, D. P. Mindell, and R. A. Goldstein. 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J. Mol. Evol. **55**:65–73.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. **266**:418–427.

———. 2002. PHYLIP (Phylogeny Inference Package) Version 3.6a. Department of Genome Sciences, University of Washington, Seattle, Wash. http://evolution.genetics.washington.edu/phylip.html

———. 2003. Inferring phylogenies. Sinauer Associates, Sunderland, Mass.

Goldman, N., J. L. Thorne, and D. T. Jones. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analysis. J. Mol. Biol. **263**:196–208.

———. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics **149**: 445–458.

Grassly, N. C., J. Adachi, and A. Rambaut. 1997. PSeq-Gen: an application for the Monte Carlo simulation of protein sequence evolution along phylogenetic trees. CABIOS **13**:559–560. http://evolve.zoo.ox.ac.uk/software.html?id=pseqgen

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**:696–704. http://atgc.lirmm.fr/phyml

International Chimpanzee Chromosome 22 Consortium. 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. Nature **429**:382–388.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409**:928–933.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS **8**:275–282.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. **31**:151–160.

Liò, P., and N. Goldman. 1998. Models of molecular evolution and phylogeny. Genome Res. **8**:1233–1244.

Moler, C., and C. Van Loan. 2003. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev. **45**:3–49.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520–562.

Müller, T., and M. Vingron. 2000. Modeling amino acid replacement. J. Comp. Biol. **7**:761–776.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. CABIOS **13**:235–238. http://evolve.zoo.ox.ac.uk/software.html?id=seqgen

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428**:493–521.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572–1574. http://morphbank.ebc.uu.se/mrbayes3

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. Tree-Puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18**:502–504. http://www.tree-puzzle.de

Swofford, D. L. 2002. PAUP*. *Phylogenetic analysis using parsimony (and other methods) version 4. Sinauer Associates, Sunderland, Mass. http://paup.csit.fsu.edu

Thorne, J. L., and N. Goldman. 2003. Probabilistic models for the study of protein evolution. Pp. 209–226 *in* D. J. Balding, M. Bishop, and C. Cannings, eds. Handbook of Statistical Genetics, 2nd Ed. Wiley, Chichester.

Veerassamy, S., A. Smith, and E. R. M. Tillier. 2003. A transition probability model for amino acid substitutions from Blocks. J. Comp. Biol. **10**:997–1010.

Whelan, S., P. I. W. de Bakker, and N. Goldman. 2003. Pandit: a database of protein and associated nucleotide domains with inferred trees. Bioinformatics **19**:1556–1563. http://www.ebi.ac.uk/goldman-srv/pandit

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. **18**:691–699.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13**:555–556. http://abacus.gene.ucl.ac.uk/software/paml.html

Yang, Z., R. Nielsen, and M. Hasegawa. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. **15**:1600–1611.

Arndt von Haeseler, Associate Editor

Accepted October 4, 2004