A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach

Simon Whelan and Nick Goldman

Department of Zoology, University of Cambridge, Cambridge, England

Phylogenetic inference from amino acid sequence data uses mainly empirical models of amino acid replacement and is therefore dependent on those models. Two of the more widely used models, the Dayhoff and JTT models, are estimated using similar methods that can utilize large numbers of sequences from many unrelated protein families but are somewhat unsatisfactory because they rely on assumptions that may lead to systematic error and discard a large amount of the information within the sequences. The alternative method of maximum-likelihood estimation may utilize the information in the sequence data more efficiently and suffers from no systematic error, but it has previously been applicable to relatively few sequences related by a single phylogenetic tree. Here, we combine the best attributes of these two methods using an approximate maximum-likelihood method. We implemented this approach to estimate a new model of amino acid replacement from a database of globular protein sequences comprising 3,905 amino acid sequences split into 182 protein families. While the new model has an overall structure similar to those of other commonly used models, there are significant differences. The new model outperforms the Dayhoff and JTT models with respect to maximum-likelihood values for a large majority of the protein families in our database. This suggests that it provides a better overall fit to the evolutionary process in globular proteins and may lead to more accurate phylogenetic tree estimates. Potentially, this matrix, and the methods used to generate it, may also be useful in other areas of research, such as biological sequence database searching, sequence alignment, and protein structure prediction, for which an accurate description of amino acid replacement is required.

Introduction

The majority of likelihood methods used for reconstructing phylogenies from amino acid sequences rely on empirical models of protein evolution. These models need good replacement matrices, which represent the relative rates of amino acid replacement at homologous sites in a protein, to accurately estimate the true evolutionary distances and relationships among species. Unfortunately, none of the current methods used to estimate these replacement matrices are entirely satisfactory.

Dayhoff and colleagues (Dayhoff, Eck, and Park 1972; Dayhoff, Schwartz, and Orcutt 1978) used a parsimony-based counting method to generate accepted point mutation (PAM) matrices from the limited amount of protein sequence data available at the time. To achieve this, phylogenetic trees were estimated for multiple protein families, along with the ancestral sequences within those trees, using maximum parsimony (MP). This information was then used to estimate the relative rates of all amino acid replacements by simply counting both the inferred numbers of different amino acid replacements that occurred on all of the lineages of the trees and the numbers of occasions on which no change in amino acids was inferred.

Jones, Taylor, and Thornton (1992) applied a faster, automated procedure based on Dayhoff and colleagues' (Dayhoff, Eck, and Park 1972; Dayhoff, Schwartz, and Orcutt 1978) approach and used it to produce a replacement matrix from a much larger database. After estimating the phylogenetic tree for each protein family in the database, their method selected a pair of sequences from a phylogeny that were nearest-neighbors and were $>85\%$ identical and counted the differences between them. The pair of sequences was then discarded to avoid recounting changes on any given branch of a phylogeny. This process was repeated for all such pairs of sequences in all protein families from their database. The 85% identity rule was used to reduce the number of multiple changes recorded as single replacements. Both of these approaches, which we refer to as counting methods, produce matrices of counts that may be used to estimate Markov process models of amino acid replacement (Swofford et al. 1996; Liò and Goldman 1998). The two models described above, the most widely used for the phylogenetic analysis of amino acid sequences of globular proteins, are both estimated using these counting methods and are known as the Dayhoff model (Dayhoff, Schwartz, and Orcutt 1978) and the JTT model (Jones, Taylor, and Thornton 1992).

The counting methods effectively employ MP to estimate amino acid replacement matrices and are therefore susceptible to its inherent problems. In particular, MP intrinsically assumes that for any given site in an alignment, only one change takes place along any single branch in a tree. This can lead to a serious underestimation of the true number of replacements that have occurred in branches where multiple changes have occurred and may consequently lead to systematic error in any model estimated using counts of replacements. In addition, MP inferences of ancestral sequences may introduce still further inaccuracies (Yang, Kumar, and Nei 1995). The Dayhoff model may be affected by both of these problems. The 85% rule of the JTT method attempts to reduce the impact of these problems by reducing the expected number of multiple hits that are neglected. Without completely solving the problem, this also renders the method very wasteful because it discards all of the information available in the sequences

Key words: amino acid replacement, general reversible model, maximum likelihood, protein evolution.

Address for correspondence and reprints: Simon Whelan, Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, United Kingdom. E-mail: s.whelan@zoo.cam.ac.uk.

Mol. Biol. Evol. 18(5):691–699. 2001

that are <85% identical. The JTT method avoids making inferences of ancestral sequences, but at the further cost of using an inefficient method for avoiding the repeated counting on branches of phylogenetic trees, discarding many sequences which have >85% identity only with previously used sequences.

Adachi and Hasegawa (1996), Yang, Nielsen, and Hasegawa (1998), and Adachi et al. (2000) used maximum-likelihood (ML) methods to estimate models of amino acid replacement for vertebrate mitochondrial, mammalian mitochondrial, and chloroplast sequences, respectively. For an alignment of sequences related by a single phylogenetic tree, the amino acid replacement matrix that gave the highest likelihood was found simultaneously with the optimal phylogeny and branch lengths. This ML approach avoids the problems associated with the counting methods by using all of the information available in the sequences across all levels of homology and by having a model that explicitly allows multiple changes to occur on a single branch at any site in an alignment. Unfortunately ML, while providing a more reliable estimate of a model of replacement than the counting methods, has a much greater computational burden associated with it. The time each individual likelihood calculation takes and the number of calculations required to numerically maximize the likelihood increase significantly with each sequence added to an analysis. This has meant that relatively few sequences, each consisting of a number of concatenated genes available for all of the organisms studied, have been included in previous analyses: Adachi and Hasegawa (1996) analyzed 20 sequences, each of 3,357 residues; Yang, Nielsen, and Hasegawa (1998) used 23 sequences of similar lengths; and Adachi et al. (2000) used just 10 sequences, each of 9,957 residues. This may restrict the accuracy of the resulting models or the variety of proteins for which the models are subsequently found to be useful (see also P. Liò and N. Goldman, unpublished data).

Here, we combine the best attributes of the likelihood and counting methods to estimate a model of amino acid replacement from a large database of different globular protein families using an approximation to ML. This model should provide a better estimate of the evolutionary process than existing models estimated using counting methods and be applicable to phylogenetic studies of a much broader range of protein sequences than existing models estimated using the likelihood approach.

Models of Amino Acid Replacement

The Amino Acid Replacement Matrix

All the models discussed in this paper assume that all amino acid sites in an alignment evolve independently and according to the same Markov process. The Markov process is assumed to be both stationary and homogeneous, so that the amino acid frequencies and the model of evolution are assumed constant through time and across all sites in an alignment. Additionally, the Markov process is assumed to be reversible, implying that to an observer it would appear the same going backwards in time as it would going forward. The probability of amino acid i being replaced by amino acid j over time T is $P_{ij}(T)$, where i and j take the values 1, 2, . . . , 20, representing the 20 different amino acids. These probabilities can be written as a 20×20 matrix, $\mathbf{P}(T)$, which is calculated as $\mathbf{P}(T) = \exp(T\mathbf{Q})$, where $\mathbf{Q}$ is the rate matrix, with off-diagonal elements Q_{ij} being the instantaneous rates of change of amino acid i to amino acid j and with diagonal elements Q_{ii} being fixed so that the row sums of $\mathbf{Q}$ equal 0. The off-diagonal elements of the matrix $\mathbf{Q}$ can be described by the off-diagonal elements of the matrix product

$$\begin{pmatrix} — & s_{1,2} & s_{1,3} & \cdots & s_{1,20} \\ s_{1,2} & — & s_{2,3} & \cdots & s_{2,20} \\ s_{1,3} & s_{2,3} & — & \cdots & s_{3,20} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{1,20} & s_{2,20} & s_{3,20} & \cdots & — \end{pmatrix} \cdot \operatorname{diag}(\pi_1, \ldots, \pi_{20}), \qquad (1)$$

so $\mathbf{Q}$ can be defined by two sets of components, s_{ij} and π_i. The variables s_{ij} represent the exchangeabilities of amino acid pairs (i, j). Time reversibility is imposed by placing the restriction that $s_{ij} \equiv s_{ji}$ (as above), resulting in 190 such parameters. Empirically derived models of amino acid replacement describe the evolutionary process by fixing these exchangeabilities to values that have been estimated from a large amount of data. When performing likelihood calculations on a tree, the matrix $\mathbf{Q}$ is scaled to provide meaningful branch lengths (fixing $-\Sigma_i\ \pi_i q_{ii} = 1$ means that evolutionary distances T are measured in units of expected numbers of replacements per site), and this effectively removes one parameter, leaving 189 free parameters describing relative amino acid exchangeability.

The π_i values represent the equilibrium or stationary frequencies of the 20 amino acids. These frequencies may all be set to 1/20 or may be set to the values estimated from the original data used to estimate the s_{ij} values. These applications are now relatively rare in phylogenetics (and are not used in this paper), and the π_i are more typically estimated as being equal to the proportions of the amino acids as observed in a data set under phylogenetic analysis (Cao et al. 1994). When the frequencies are estimated from the data in this way, model names are generally given the suffix ''+F''; e.g., JTT+F would use s_{ij} as estimated by Jones, Taylor, and Thornton (1992) and the π_i observed in the data set under analysis. The 20 amino acid frequencies can be described by 19 free parameters because of the constraint $\Sigma_i \pi_i = 1$ and, in effect, weight s_{ij} according to sequences' amino acid compositions.

All of the standard models of evolution used in this paper have previously been well documented (e.g., Swofford et al. 1996; Liò and Goldman 1998). The simplest model of amino acid evolution is the equiprobable (EQU) model, which assumes that all the exchangeability parameters s_{ij} are equal and sets all of the stationary frequencies to 1/20. The EQU+F form of this model allows the stationary frequencies to equal the proportions of the different amino acids observed in the data. The Dayhoff (Dayhoff, Schwartz, and Orcutt 1978) and JTT (Jones, Taylor, and Thornton 1992) models, which in their Dayhoff+F and JTT+F forms will be compared with our new models, have values of s_{ij} which have been estimated from large databases using counting methods. The mitochondrial- and chloroplast-specific models of Adachi and Hasegawa (1996), Yang, Nielsen, and Hasegawa (1998), and Adachi et al. (2000), which are not appropriate for direct comparison with our new model because of their sequence-specificity, have each had their 189 free s_{ij} parameters estimated by direct likelihood maximization for relatively few protein sequences and using a single evolutionary tree.

Assumptions Needed to Estimate a Model Using Multiple Phylogenies

The simultaneous use of many different protein families implies that an estimated model may be applicable to a wide range of proteins (as are the Dayhoff and JTT models), and the use of the likelihood approach to perform this estimation suggests that it may more accurately reflect the evolutionary process by avoiding systematic error and utilizing more of the available data. In order to estimate an empirical model of amino acid replacement simultaneously from many families of sequences, we have developed a new approximation to the likelihood approach, exploiting two observations about the ML estimation of parameters on phylogenetic trees. First, it has been shown that parameters describing the evolutionary process remain relatively constant across near-optimal tree topologies (e.g., Yang, Goldman, and Friday 1994, 1995; Sullivan, Holsinger, and Simon 1996; Yang, Nielsen, and Hasegawa 1998; Adachi et al. 2000). We exploit this by assuming it to be the case for the parameters used to describe amino acid replacement, in particular, assuming that the relative values of the amino acid exchangeability parameters s_{ij} stay approximately constant over near-optimal branch lengths and tree topologies. The implication of this assumption is that so long as branch lengths are close enough to optimal when estimating the new model, any changes in the branch lengths observed when they are reestimated under the new model would not influence the model estimated to any great extent.

The second observation relates to changes in individual branch lengths that occur when performing ML estimation under two different models of evolution for a single-tree topology. When the two models are quite different, the ML branch lengths they give can be quite different (demonstrated by comparing the statistics shown in table 3 for either the EQU or the EQU+F model of evolution with those for either the Dayhoff+F or the JTT+F model). When two models are alike in their abilities to describe the evolutionary process, however, there is often much less difference in the branch lengths (e.g., Yang, Nielsen, and Hasegawa 1998; also illustrated by comparing the statistics shown in table 3 for the EQU and EQU+F models or those for the Dayhoff+F and JTT+F models). We exploit the relatively small changes in branch lengths under ‘‘similarly good’’ models of evolution by assuming that the JTT+F model is capable of providing near-optimal branch lengths for the best general model of evolution (yet to be estimated).

Calculating a Likelihood Using Multiple Phylogenies

In order to calculate a likelihood for a complete database of aligned protein families, each protein family was taken in turn, and all pairwise phylogenetic distances between the sequences were estimated using the Dayhoff+F model. These distances were used to estimate phylogenetic tree topologies using neighbor-joining (Saitou and Nei 1987). These steps were performed using programs from the PHYLIP software package (Felsenstein 1995). Although perhaps some of these topologies will not be the ML topologies under the new model ultimately estimated, we assume they will be close enough to permit reasonably accurate estimation of this new model (see above). Next, the branch lengths for each family's phylogenetic tree were reestimated by ML under the JTT+F model (these and all subsequent analyses were performed using purpose-written software). These calculations are computationally slow, especially when using large numbers of sequences, and in order to reduce this burden, an arbitrary upper limit of 100 sequences was taken to be the maximum size for any single protein family. In the BRKALN database (see below), five families were larger than this limit and were split to form similarly sized subfamilies of fewer than 100 sequences. To try and minimize the amount of information lost from each family being split and to avoid the problem of overlapping trees where individual branches might be incorporated multiple times, families were split along single branches of their phylogenies as estimated in the neighbor-joining step.

Rather than completely fixing these estimates of the branch lengths during the estimation of the amino acid replacement model, only the ratios of branch lengths were fixed, and a scaling factor ρ was introduced which allowed all branch lengths to increase or decrease linearly. This parameter makes some allowance for any unforeseen changes in branch lengths between the JTT+F model and the new model being estimated, which could occur if the assumptions discussed above regarding branch lengths were invalid. As we assume that the families' topologies and relative branch lengths are now fixed at near-optimal values, the overall log-likelihood for the database can be calculated as

$$\log L = \log L(\mathbf{M}, \mathbf{T} | \mathbf{D}) \tag{2}$$

$$\approx \log L(\mathbf{M} | \mathbf{T}, \mathbf{D})$$

$$= \sum_{\text{families } i=1}^{n} \log L(\mathbf{M} | T_i, D_i), \tag{3}$$

where $\mathbf{D} = (D_1, \ldots, D_n)$ represents the database of n aligned protein families, $\mathbf{T} = (T_1, \ldots, T_n)$ represents the tree topologies and relative branch lengths for each family, and $\mathbf{M}$ represents the model of evolution consisting of the exchangeability parameters s_{ij}, the stationary fre-

Table 1
Relative Proportions of the Amino Acids in the BRKALN Database Used to Estimate the New Models

Amino Acid	Frequency
A.	0.0866
R.	0.0440
N.	0.0391
D.	0.0570
C.	0.0193
Q.	0.0367
E.	0.0581
G.	0.0833
H.	0.0244
I	0.0485
L.	0.0862
K.	0.0620
M	0.0195
F	0.0384
P	0.0458
S	0.0695
T.	0.0610
W	0.0144
Y.	0.0353
V.	0.0709

quencies π_i, and the scaling factor ρ. To find the ML model of evolution, we need only maximize log L over $\mathbf{M}$ in equation (3) while fixing the parameters associated with $\mathbf{T}$, as our assumptions mean that the resulting model will be close to that obtained by maximizing equation (2) over both $\mathbf{M}$ and $\mathbf{T}$. This dramatically reduces the computational time required for large amounts of data because most of the parameters that would normally require optimization are located within $\mathbf{T}$.

Table 2
Log-Likelihood Values of the Entire Database of Protein Families Under the WAG and WAG* Models and Other Commonly Used Models

Model	Log-Likelihood (+F)[a] (improvement over JTT)[b]	Log-Likelihood (+mF)[c] (improvement over JTT)[b]	Improvement of +mF over +F
EQU	−770,812.7[d] (—)	−767,010.7 (—)	3,802.0*
Dayhoff . . .	−732,442.1 (—)	−730,132.4 (—)	2,309.7*
JTT.	−728,611.8 (—)	−726,166.2 (—)	2,445.6*
WAG*	−722,008.4 (6,603.4)	−719,332.0 (6,834.2)	2,676.4*
WAG	−721,930.9 (6,680.9)	−719,428.3 (6,737.9)	2,502.6*

NOTE.—An asterisk in the right-hand column indicates that the increase in likelihood between the +F and +mF models is statistically significant ($P < 0.01$). Significance is tested using a standard likelihood ratio test between the two models, comparing twice the difference in log-likelihoods with a $\chi^2_{3,439}$ distribution, where 3,439 is the number of degrees of freedom by which the two models differ. A normal approximation to the $\chi^2_{3,439}$ distribution (Lindgren 1976) gives these log-likelihood differences standard z-scores from 12.9 to 19.8; all P values are too small to calculate reliably.

[a] Model applied with +F option: one set of amino acid frequencies applied to whole database.

[b] Difference recorded only for models with log-likelihoods exceeding that of JTT.

[c] Model applied with +mF option: different sets of amino acid frequencies applied to each family in the database.

[d] We note that this is probably the lowest log-likelihood value ever recorded in phylogenetics.

Application to Real Data

This method was used to estimate a general model of amino acid replacement from the BRKALN database of aligned protein sequence families (D. Jones, unpublished data). This database has previously been used to estimate amino acid replacement models specific to different protein secondary structures (e.g., Goldman, Thorne, and Jones 1996), and we used 3,905 sequences split into 182 protein families, each containing no more than 100 sequences. The amino acid frequencies for the entire database are shown in table 1.

As described so far, the methods above assume that a single set of stationary frequencies is sufficient to describe the evolution of all of the protein families in a database. Different protein families may, however, be expected to have different amino acid compositions due to a variety of biochemical factors, such as differing cellular environments or variable proportions of protein secondary-structure elements. Two different estimation methods were used to address this question. The first method used a single set of stationary frequencies (19 free parameters) for all protein families, estimated by counting the amino acids observed in the database (as in table 1), with the remaining parameters of $\mathbf{M}$ (189 free exchangeability parameters s_{ij} and the scaling factor ρ) estimated by ML. The resulting replacement model is called the WAG model, after the authors of this paper. The second method used a different set of stationary frequencies for each protein family ($19 \times 182 = 3,458$ free parameters, estimated by counting the amino acids observed in each family), with the remaining parameters of $\mathbf{M}$ again estimated by ML. In this case, the resulting model is called the WAG* model.

Both the WAG and the WAG* models required only 190 parameters to be numerically optimized, compared with >7,500 (before even considering optimization over topologies) under the traditional likelihood approach. This optimization was still computationally slow: estimation of the WAG model took approximately 18 h on a Digital 600au Personal Workstation. To avoid local maxima, estimations of the WAG and WAG* models were each performed using two sets of starting values, those of the EQU and the JTT models; the same estimates were recovered in each case. We found it computationally impractical to estimate the stationary frequency parameters π_i by ML (Yang and Roberts 1995) simultaneously with the estimation of the s_{ij}. Given the large size of the BRKALN database, we expect that any differences in the estimated π_i would be small and that any consequent differences in the estimated s_{ij} would be insignificant.

Results

Comparison of the Overall Performance of the New Models with Other Commonly Used Models

Log-likelihood values (eq. 3) for the two newly estimated models and other commonly used models of

Table 3
Tree-Related Statistics and Log-Likelihood Values for the Phosphoenolpyruvate Carboxykinase Alignment

	Model of Evolution					
	EQU	EQU+F	Dayhoff+F	JTT+F	WAG+F	WAG*+F
Tree length	1.889	1.893	1.977	1.951	2.032	1.923
Longest branch. . . .	0.144	0.144	0.158	0.156	0.150	0.150
Shortest branch. . . .	0.0096	0.0091	0.0087	0.0080	0.0108	0.0106
Median branch	0.059	0.060	0.059	0.058	0.060	0.060
Average branch . . .	0.057	0.057	0.060	0.059	0.062	0.058
Log-likelihood	−2,386.66	−2,347.66	−2,210.38	−2,216.74	−2,168.13	−2,167.28

NOTE.—This protein family consists of 18 sequences of 163 unambiguously aligned residues, not including gaps. Branch lengths are in units of expected numbers of replacements per site.

amino acid replacement were calculated for the complete database and are shown in table 2. Each model was applied in both the +F form, with one set of amino acid frequencies estimated from the entire database and applied to the analysis of all protein families, and in a form denoted +mF (multiple frequencies), with a different set of amino acid frequencies estimated for each family. In all cases, models that used multiple sets of stationary frequencies were significantly better than the equivalent model using only a single set of stationary frequencies (likelihood ratio test; twice the log-likelihood difference compared with a χ^2 distribution with $3{,}458 - 19 = 3{,}439$ df—see Yang, Goldman, and Friday 1994). All models of amino acid replacement that allow for unequal amino acid exchangeabilities have much higher likelihood values than the EQU model.

Most importantly, both of the new models (WAG and WAG*) have higher likelihoods than any of the other models. Statistical comparisons of these models against the JTT model can be made by comparing twice the log likelihood differences given in table 2 with a χ^2_{190} distribution, with the 190 df being derived from the 190 parameters s_{ij} and ρ estimated during the generation of the WAG* and WAG models. In all cases, (JTT+F vs. WAG*+F or WAG+F; JTT+mF vs. WAG*+mF or WAG+mF), the WAG* and WAG models give a much better fit to the data: using a normal approximation to the χ^2_{190} distribution (Lindgren 1976), z-scores for these four tests are all >90 (cf. a standard normal distribution with mean 0 and variance 1), and all P values are too small to calculate reliably.

We also note that, even after making allowance for the estimation of 190 additional parameters, the improvement in likelihood of the WAG* or WAG model over the JTT model is greater than the improvement of the JTT model over the Dayhoff model. This suggests that the improvement achieved by estimating a model of evolution using our new method may be at least as great as the improvement obtained by using a larger database from which to estimate a model of evolution, which is the main detail in which the Dayhoff and JTT models differ.

When the WAG* and the WAG models are compared, neither appears clearly better than the other in examining the whole database of families. As expected, each performs best for the analysis conditions it was optimized for, with WAG giving a better likelihood when using a single set of stationary frequencies (+F option) and WAG* performing better for multiple sets of stationary frequencies (+mF). When the two models are compared using equivalent methods (both +F or both +mF) for calculating the stationary frequencies, their log-likelihood values are very similar (changing by approximately 0.01%), suggesting that there is little difference between the two models.

The branch length scaling factor ρ was estimated as 1.027 during the generation of the WAG model. This suggests that the likelihood maximization procedure was not trying to change the branch lengths dramatically and that the approximations used were valid. While we note that this scaling factor may not detect nonlinear changes in branch lengths (i.e., changes not proportional to the original lengths), results obtained by the reestimation of the branch lengths and a subsequently reestimated model give no indication of this occurring (see below).

Performance of New Models on Specific Phylogenies

The new models' performance when estimating individual phylogenies is of more practical relevance than their performance when estimating a likelihood for an entire database. To give an example of the improvement in fit to the data that might be achieved with our new models, an alignment of 18 Lepidopteran sequences of the phosphoenolpyruvate carboxykinase protein (Friedlander et al. 1996; Goldman, Thorne, and Jones 1998) was chosen as a typical example of data used to perform a phylogenetic analysis. Some statistics of the ML trees under different replacement models are shown in table 3. From these statistics, it is clear that the use of the WAG+F or WAG*+F model results in a considerably higher likelihood value than any of the other models. The difference between the two new models is very small. There is some change in the branch length statistics between JTT and the new models, although it does not appear large enough to invalidate the assumptions used to estimate the WAG and WAG* models. Note that the phosphoenolpyruvate carboxykinase family is not represented in the BRKALN database, and so there is no possibility of the WAG and WAG* models having any unfair advantage. In this example, the WAG and WAG* models of sequence evolution are superior, and, in general, we expect the use of the models giving the best fit to the observed data to lead to more accurate phylogenetic estimation. Even small changes in branch lengths may lead to changes in optimal ML tree topology, resulting in an estimated tree being closer to the true evolutionary tree.

In order to demonstrate this improvement in performance for the whole BRKALN database, ML values were calculated for each protein family under the JTT+F, WAG+F, and WAG*+F models, this time fixing the evolutionary models and tree topologies but reestimating the branch lengths for each family's phylogeny. Figure 1*A* and *B* shows that the majority of protein families (146 out of 182, or 80%) had higher likelihoods when analyzed under either the WAG+F or the WAG*+F model than under the JTT+F model. The protein families whose likelihoods were higher under the JTT+F model were examined in more detail and compared with the families whose likelihoods were higher under the WAG+F and WAG*+F models to see if there was any common feature defining which model was preferred. It was found that families for which the JTT+F model gave a higher likelihood had relatively shorter average branch lengths than those families for which the new models gave higher likelihoods (data not shown). This may be the result of Jones, Taylor, and Thornton's (1992) counting method of replacement matrix estimation using only sequences that are >85% identical to estimate the JTT model and perhaps consequently overfitting the model to relatively short evolutionary distances.

Figure 1*A* and *B* shows that the increases in performance of the WAG+F and WAG*+F models compared with the JTT+F model are very similar. To demonstrate the relative performance of the two new models, the increase in likelihood of the WAG*+F model over the WAG+F model for each individual protein family is shown in figure 1*C*. It is apparent that the difference in likelihood between the two models is minimal for the majority of the families: log-likelihood differences between the WAG+F and WAG*+F models are <1 for 102 of the 182 families; 95 families lie above the *x*-axis in figure 1*C*, and 87 lie below. The few cases in which the likelihoods were clearly different between the two models are located toward the highest log-likelihood values (i.e., the left-hand side of fig. 1*C*), with the WAG model clearly outperforming the WAG* model. These cases were investigated in more detail, and it was found that these families consisted of only two very similar sequences. This suggests that the largest differences in likelihood between the WAG and the WAG* models were caused by one or two differences in the amino acid replacement matrices of the two models coinciding with differences between two closely related sequences and can thus be attributed to chance effects. We conclude that the overall difference between the two models' performances is negligible, and the additional parameters (and computation time) used when estimating the WAG* model are not required for model estimation from these data.

Comparison of the Structure of the New Models with Those of Other Commonly Used Models

A comparison of the differences in the patterns of amino acid replacement between the empirically derived Dayhoff, JTT, WAG, and WAG* models of evolution is shown in figure 2. From these graphs, it is clear that the overall structures of the four models are similar, which suggests that they are all modeling the same process. Closer examination shows no discernible pattern to the differences in the values of the parameters s_{ij} of the amino acid replacement matrices of the JTT and WAG models; this is illustrated in figure 3. There is almost no difference between the values in the replacement matrices of the WAG and WAG* models. The exchangeability parameters s_{ij} defining the WAG and WAG* models are available via http://www.zoo.cam.ac.uk/zoostaff/goldman/WAG.

Testing the Adequacy of Approximations

The methodology presented here may be considered similar to a single round of optimization in the algorithms often used to maximize a likelihood for a single given phylogenetic tree, which involve alternating cycles of branch length optimization and model optimization. In our methodology, we first optimize branch lengths for multiple families under a fixed model and then find the optimal model using those branch lengths. It is therefore of interest to see whether further rounds of optimization in our methodology would provide any substantial increase in likelihood and, consequently, a better fit to the data. We performed a second round of optimization, involving a single reestimation of all of the branch lengths for each individual protein family using the WAG+F model followed by the reestimation of the replacement model using these branch lengths. This reestimated model was then used to examine the individual families in the BRKALN database. This resulted in only trivial changes to branch lengths, estimated parameter values, and likelihood values (e.g., an average increase in log-likelihood of only 0.026 per family), and from this we conclude that second and subsequent rounds of iteration are unnecessary to get a good estimate of the optimal evolutionary model.

Discussion

The methodology presented here allows the estimation of a model of amino acid replacement from large numbers of sequences from many different families. By doing so, it combines the best attributes of the counting methods of Dayhoff, Schwartz, and Orcutt (1978) and Jones, Taylor, and Thornton (1992), used to estimate the Dayhoff and JTT amino acid replacement models, and the true ML methods of Adachi and Hasegawa (1996), Yang, Nielsen, and Hasegawa (1998), and Adachi et al. (2000).

The newly estimated WAG and WAG* models both gave significantly higher likelihoods than any other commonly used models when used to assess phylogenies for all 182 protein families of the BRKALN data-

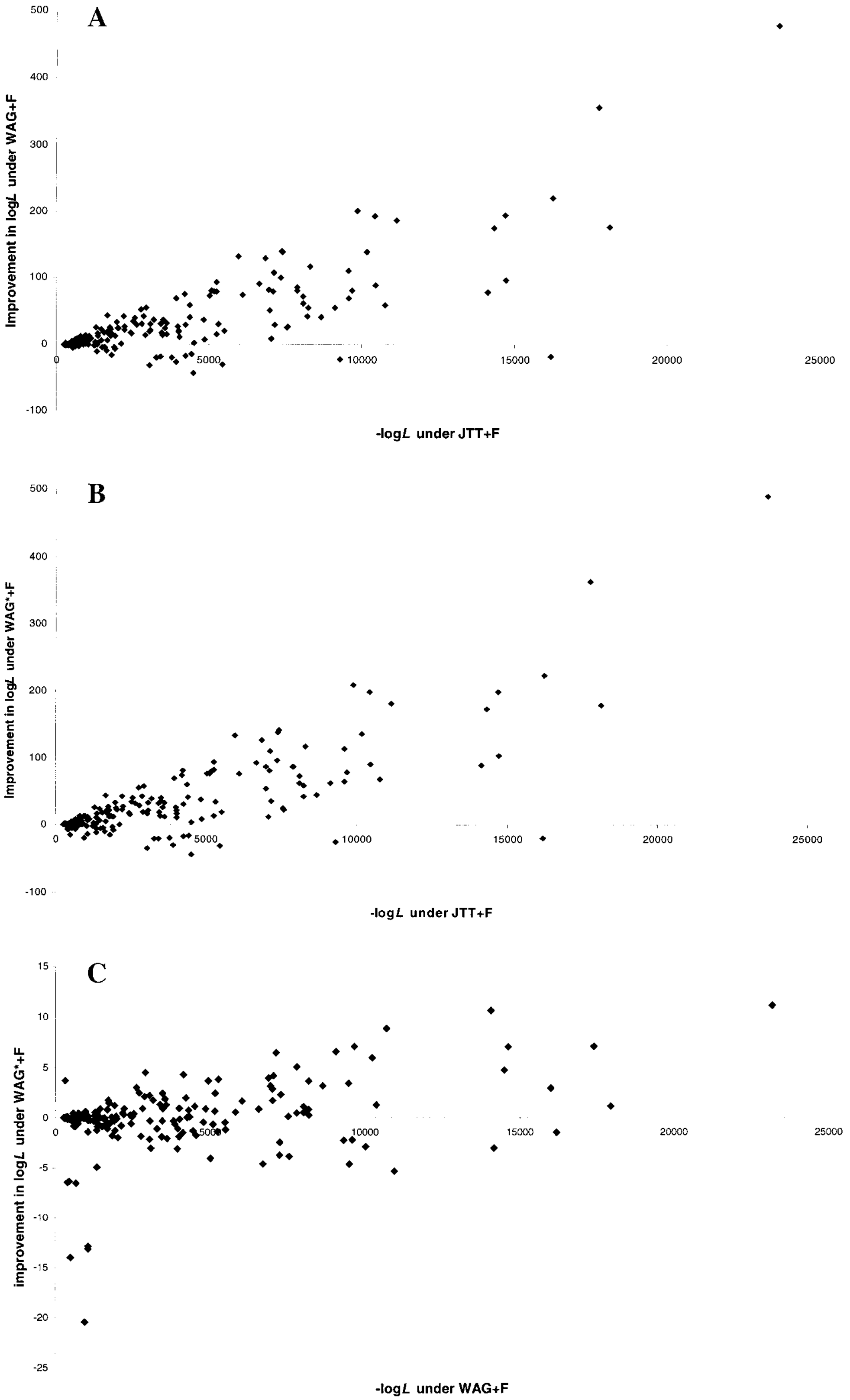

FIG. 1.—Likelihood improvements under the new models. *A,* Improvement in log-likelihood (log *L*) obtained in phylogenetic inference with the WAG+F model relative to the JTT+F model for each protein family in the BRKALN database. *B,* Improvement in log-likelihood obtained in phylogenetic inference with the WAG*+F model relative to the JTT+F model for each protein family in the database. *C,* Improvement in log-likelihood obtained in phylogenetic inference with the WAG*+F model relative to the WAG+F model for each protein family in the database. Note the different scale on the *y*-axis.

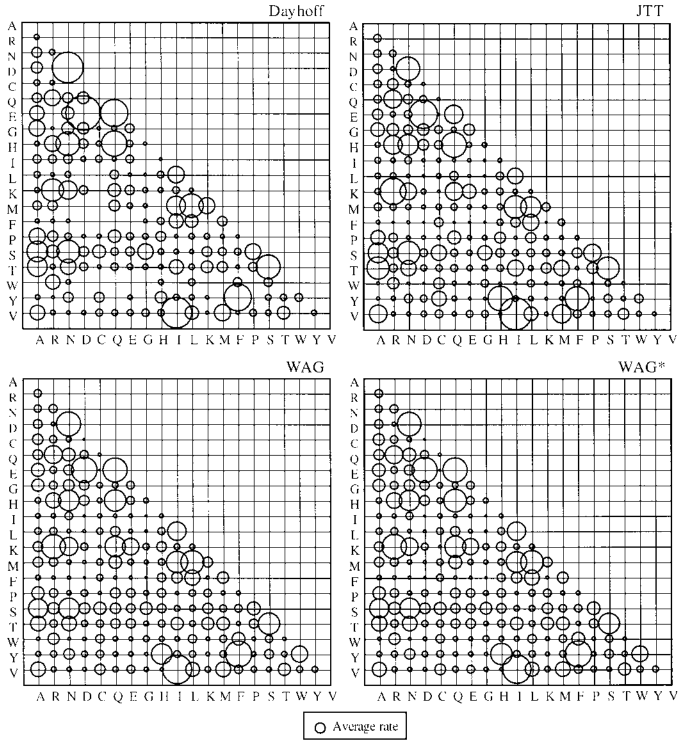

FIG. 2.—Schematic representations of amino acid replacement matrices. The area of each bubble represents the amino acid exchangeability parameter (s_{ij}) for the replacement of amino acid *i* by amino acid *j* or vice versa. The models depicted are those of Dayhoff, Schwartz, and Orcutt (1978) and Jones, Taylor, and Thornton (1992) and the new WAG and WAG* models. All s_{ij} values are scaled so that the mean rate of evolution is 1, assuming equal frequency parameters.

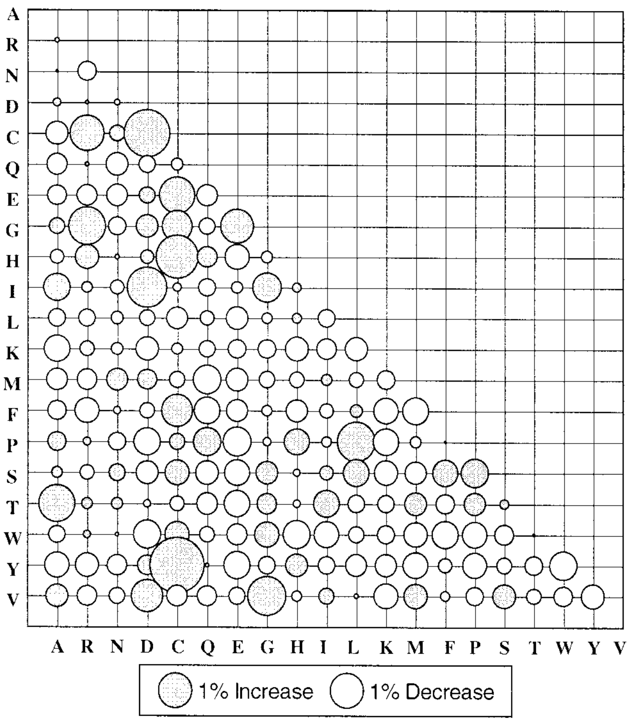

FIG. 3.—Schematic representation of the differences between the JTT and WAG amino acid replacement matrices. The area of each bubble is calculated as (s_{ij}(JTT) − s_{ij}(WAG))/s_{ij}(JTT). In total, there are 66 increases (gray bubbles) and 124 decreases (white bubbles) in the WAG matrix relative to the JTT matrix. The s_{ij} values are scaled as in figure 2.

base simultaneously and for a large majority of the individual phylogenies examined. It was unclear which of the two new models performed better, but we would tentatively suggest that in this case the methodology used to estimate the WAG model was preferable because it involved fewer parameters being estimated from the amino acid sequence database. We note some change in the estimated branch lengths under the new models of evolution. The better statistical fit of our new models to the data suggests that in many cases they may provide more accurate estimates of phylogenetic trees than existing models, although differences in branch length estimates do not appear so great as to invalidate the assumptions used in our method for estimating models.

We hope that the WAG and WAG* models of amino acid replacement will be of value in phylogenetic analyses of amino acid sequences as potentially superior alternatives to the Dayhoff and JTT models. Both our new methodology and models produced using it may have further applications outside of phylogenetics, in fields that rely on accurate descriptions of amino acid replacement, such as protein structure prediction, the detection of sequence homology (including database searching), and sequence alignment.

Acknowledgments

S.W. was supported by a BBSRC studentship. N.G. was supported by a Wellcome Trust Biodiversity Fellowship. Further details of algorithms used and source code are available on request from the authors, and the WAG and

WAG* matrices are available in electronic form via http://www.zoo.cam.ac.uk/zoostaff/goldman/WAG. The WAG model is implemented in the PAML (Yang 1997—see http://abacus.gene.ucl.ac.uk/software/paml) and TREE-PUZZLE (Strimmer and von Haeseler 1996—see http://www.tree-puzzle.de) software packages.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. **42**:459–468.

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.

FELSENSTEIN, J. 1995. PHYLIP (phylogenetic inference package). Version 3.57. Distributed by the author, Department of Genetics, University of Washington, Seattle.

FRIEDLANDER, T. P., J. C. REGIER, C. MITTER, and D. L. WAGNER. 1996. A nuclear gene for higher-level phylogenetics—phosphoenolpyruvate carboxykinase tracks Mesozoic-age divergences within *Lepidoptera* (Insecta). Mol. Biol. Evol. **13**:594–604.

GOLDMAN, N., J. L. THORNE, and D. T. JONES. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J. Mol. Biol. **263**: 196–208.

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

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.

LINDGREN, B. W. 1976. Statistical theory. 3rd edition. Macmillan, New York.

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

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

STRIMMER, K., and A. VON HAESELER. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

SULLIVAN, J., K. E. HOLSINGER, and C. SIMON. 1996. The effect of topology on estimation of among site rate variation. J. Mol. Evol. **42**:308–312.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. Sinauer, Sunderland, Mass.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13**:555–556.

YANG, Z., N. GOLDMAN, and A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. **11**:316–324.

———. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. **44**: 384–399.

YANG, Z., S. KUMAR, and M. NEI. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641–1650.

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.

YANG, Z., and D. ROBERTS. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. **12**:451–458.

WILLIAM MARTIN, reviewing editor

Accepted January 4, 2001