Syst. Biol. 52(5):594–603, 2003
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150390235395

Estimation of Rates-Across-Sites Distributions in Phylogenetic Substitution Models

EDWARD SUSKO,[1] CHRIS FIELD,[1] CHRISTIAN BLOUIN,[2] AND ANDREW J. ROGER[2]

[1]*Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia B3H 3J5, Canada; E-mail: susko@mathstat.dal.ca (E.S.)*
[2]*Canadian Institute for Advanced Research, Program in Evolutionary Biology, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada*

Abstract.—Previous work has shown that it is often essential to account for the variation in rates at different sites in phylogenetic models in order to avoid phylogenetic artifacts such as long branch attraction. In most current models, the gamma distribution is used for the rates-across-sites distributions and is implemented as an equal-probability discrete gamma. In this article, we introduce discrete distribution estimates with large numbers of equally spaced rate categories allowing us to investigate the appropriateness of the gamma model. With large numbers of rate categories, these discrete estimates are flexible enough to approximate the shape of almost any distribution. Likelihood ratio statistical tests and a nonparametric bootstrap confidence-bound estimation procedure based on the discrete estimates are presented that can be used to test the fit of a parametric family. We applied the methodology to several different protein data sets, and found that although the gamma model often provides a good parametric model for this type of data, rate estimates from an equal-probability discrete gamma model with a small number of categories will tend to underestimate the largest rates. In cases when the gamma model assumption is in doubt, rate estimates coming from the discrete rate distribution estimate with a large number of rate categories provide a robust alternative to gamma estimates. An alternative implementation of the gamma distribution is proposed that, for equal numbers of rate categories, is computationally more efficient during optimization than the standard gamma implementation and can provide more accurate estimates of site rates. [Gamma model; Markov models; maximum likelihood; molecular evolution; phylogenetics; rate distribution.]

Many phylogenetic models used in a maximum likelihood analysis of sequence data assume independent Markov substitution processes across sites. Most of the parameters for these Markov processes are assumed constant across sites, but it has long been recognized that it is unreasonable to assume that the overall rate of evolution is constant across sites (Fitch and Markowitz, 1970; Uzzel and Corbin, 1971; Nei, 1987). Failure to account for variation in rates across sites can, under some conditions, lead to phylogenetic artifacts such as long-branch attraction, as has been shown in simulation studies (Huelsenbeck, 1995; Sullivan and Swofford, 2001) and in analyses of real data (Huelsenbeck, 1997; Silberman et al., 1999). Likelihood methods that adjust for the heterogeneity of rates across sites assume a probability distribution for these rates (Yang, 1994; Felsenstein and Churchill, 1996) and have been implemented in popular packages for phylogenetic estimation such as PHYLIP (Felsenstein, 1993), PAUP* (Swofford, 2000), TREE-PUZZLE (Strimmer and von Haeseler, 1996), and PAML (Yang, 2000). Most of these packages use a gamma model as the parametric family for the rate distribution and, for computational tractability, require some form of discretization of this distribution for estimation (Yang, 1994). It is possible that the family of gamma distributions may not be rich enough to model the actual rate distribution. This will occur, for instance, when relatively few rates are possible so that the rate distribution is discrete or when the rate distribution fits some other parametric family (e.g., log normal or inverse Gaussian). Another possibility would be that for the taxa under consideration, a gene is undergoing a rapid rate of evolution but, because of functional constraints, a certain proportion of sites have very low rates, giving rise to a bimodal distribution of rates across sites.

In addition to aiding more accurate phylogenetic estimation, rate distributions are used in obtaining rate estimates at sites; consequently, the use of incorrect rate distribution models can give rise to poor rate estimates. Rate estimates at a site have been used in studies relating protein function and structure to sites (Gaucher et al., 2001; Simon et al., 2002; Blouin et al., unpubl.). They have also been used in studies investigating "covarion behavior," i.e., rate variation across subtrees (Lopez et al., 2002; Susko et al., 2002). Clearly, the use of good rate estimates is important for any such study.

It is important to be able to check the validity of the gamma or any other parametric model for the rate distribution. To proceed we develop methods to compute nonparametric maximum-likelihood estimates of the distribution of rates-across-sites with no additional assumptions about the form of the rate distribution. If such a maximum-likelihood estimate is very different from a gamma distribution in shape and regions of mass, evidence is provided that the gamma distribution is inappropriate. We also present methods to construct bootstrap confidence intervals for the true rate distribution. With these intervals, we can verify whether a specific parametric family is appropriate for a given data set. We apply the methods to several amino acid data sets and use likelihood ratio tests and bootstrap confidence intervals to check whether the gamma distribution is appropriate. Although the gamma model provides a good fit for eight of our amino acid data sets, there is some evidence for lack of fit in the remaining five data sets. In cases where the gamma model does not fit well, rate estimates from the discrete model with a large number of rate categories can be very different than gamma model rates and provide a robust alternative because they do not require a gamma model assumption.

Rates-Across-Sites Models

The most common way of modeling rate variation is through a rate distribution. Rates at sites are treated as random (usually independent) variables that are drawn from a common distribution. To ensure that the branch lengths in the tree can be interpreted as the expected number of substitutions between the two nodes at the ends of the branch, the rate distribution must be chosen so that $E[r] = 1$.

The rate distribution in the rates-across-sites model can be any random distribution, including both continuous and discrete distributions. The most commonly used continuous model is the gamma model with density for rates:

$$g(r;\,\alpha) = \alpha^{\alpha} r^{\alpha-1} \exp(-\alpha r)/\,\Gamma(\alpha).$$

The gamma model allows for a variety of different shapes for the probability density function of rates but cannot model all possible probability densities. For instance, it cannot be used to model a bimodal density for the rates. When testing whether the gamma model is appropriate, to ensure that any distributional form can be approximated we use a discrete model with a large number of rates. The discrete model is specified by the set of rates, $r_1, \ldots, r_k$, that have positive probabilities, $\zeta_1, \ldots, \zeta_k$, of occurring and can approximate parametric densities well.

A discrete approximation will usually be required for continuous rate distribution models such as the gamma because of the computational difficulties in evaluating likelihoods. For the usual independent sites model, the likelihood is the product over sites of the unconditional probabilities, $f(x)$, of those data x observed at the sites. The unconditional probabilities $f(x)$ are calculated through $f(x\,|\,r)$, the conditional probability of x given the rate at a site, which can be calculated directly through a postorder tree traversal algorithm that requires summation of a number of terms that grows roughly linearly with the number of taxa. The unconditional probabilities are related to the conditional probabilities through

$$f(x) = \int f(x\,|\,r)g(r)\,dr \qquad (1)$$

for a continuous rate density $g(r)$ and through

$$f(x) = \sum_j f(x\,|\,r_j)\,\zeta_j \qquad (2)$$

for a discrete rate distribution. Explicit calculation of Equation 1 would prevent use of the postorder tree traversal algorithm and require summation over a number of terms that grows exponentially with the number of taxa. It is to avoid this computational difficulty that the discrete approximations are used in practice.

In the discussion below, we introduce the three main models that we will use. The first is the model that is most frequently implemented in present software, the second is a discrete alternative implemented here with large numbers of rate categories, and the third is a different implementation of the gamma model that allows one to more easily deal with large numbers of rate categories.

Discrete Gamma Estimate

The discrete gamma estimate (DGE) used by TREE-PUZZLE is based on the approximation of Yang (1994):

$$f(x) = \int f(x\,|\,r)g(r)\,dr \approx \sum_{j=1}^{k} \frac{1}{k} f(x\,|\,r_j), \qquad (3)$$

where $g(r)$ is the gamma density with parameter α. Here $r_1, \ldots, r_k$ are calculated as the $1/(2k), \ldots, (2k-1)/(2k)$ percentage points of the gamma distribution and then rescaled so that $\sum r_j/k = 1$. Because the gamma distribution depends on the parameter $\alpha, r_1, \ldots, r_k$ are functions of α. The parameter α is chosen so that $r_1, \ldots, r_k$ give the largest approximate likelihood $\prod_i f(x_i)$. Under this approach, each choice of α gives different r_i so that the $f(x\mid r_i)$ need to be recomputed during optimization, which is an expensive calculation.

Discrete Estimate

We construct a discrete estimate (DE) using a discrete model for the rate distribution. For a fixed grid $r_1 < \cdots < r_k$ of rates and corresponding probabilities $\zeta_1, \ldots, \zeta_k$, the probability of data x at a site is

$$f(x) = \sum_{j=1}^{k} \zeta_j\, f(x_i, r_j).$$

The parameters in the rate distribution that are estimated are $\zeta_1, \ldots, \zeta_k$. They are chosen to maximize the likelihood over all $\zeta_1, \ldots, \zeta_k$ that satisfy $\zeta_j \geq 0, \sum \zeta_j = 1$ and $E[r] = \sum \zeta_j r_j = 1$. In our examples, the maximization was done using the general constrained optimization algorithm VE11AD in the Harwell Subroutine Library (HSL). We vary the size of the grid for illustration but always choose the grid to be equally spaced and usually use a large grid of values ($k = 101$). This approach is in contrast to the grid used for the DGE, which is unequally spaced and usually based on a small number of intervals.

Even though the distribution we fit is discrete, with a large enough grid it can approximate any distribution reasonably well. Thus, the discrete estimate can be thought of as a nonparametric estimate of the rate distribution (it does not correspond to a family described by a few parameters). This approach is useful for comparison with parametric estimates and can be used to assess the appropriateness of a specific parametric family.

Discrete Gamma Probability Estimate

Alternative gamma model estimates can be obtained through different approximations to Equation 1. One estimate that will be of value for comparison to the discrete estimate is the discrete gamma probability estimate (DGPE).

Given a set of rates, $r_1, \ldots, r_k$, contained in intervals $I_1, \ldots, I_k$, let $\zeta_j(\alpha)$ be the probability of a rate in the interval I_j calculated under the gamma distribution with parameter α. The approximation to Equation 1 used to obtain the DGPE is

$$f(x) = \int f(x \mid r) g(r)\, dr \approx \sum_{j=1}^{k} \zeta_j(\alpha) f(x \mid r_j).$$

The parameter α is chosen so that $\zeta_1(\alpha), \ldots, \zeta_k(\alpha)$ give the largest approximate likelihood $\prod_i f(x_i)$. In our examples, given a set of rates, intervals were chosen as $I_1 = (0, b_1]$, $I_j = (b_{j-1}, b_{j+1}]$, $j = 2 \ldots, k-1$ and $I_k = (b_{k-1}, \infty)$, where $b_j = (r_j + r_{j+1})/2$, $j = 1, \ldots, k-1$.

Because the last rate category is always infinite in length, with a large number of categories, it should be chosen so that the gamma probabilities and site likelihoods for rates within it are small. One way to choose would be to check for rate estimates that are close to the lower boundary for the last rate category. If a significant number of rates are close, a refitting with a larger upper bound on the last rate category should be considered. In our examples, for simplicity, because of a lack of prior information, and to make comparison with the DE simpler, we have chosen the first $k-1$ intervals to be of equal width, but in principle they need not be.

Choosing the probabilities $\zeta_j(\alpha)$ through maximum-likelihood estimation for a fixed set of rates, which is how the DGPE is determined, has substantial computational advantages over choosing the rates $r_j(\alpha)$, which is what is done to obtain the DGE. For DGPE, which maximizes over the probabilities, the conditional probabilities of the data $f(x \mid r_i)$ are fixed throughout the computation. These conditional probabilities are relatively expensive to compute. In contrast for DGE, where rates $r_j(\alpha)$ change every time a new α is considered, the conditional probabilities must be recalculated.

Most phylogenetic analyses with a rates-across-sites model require estimation of both rate distribution parameters and a tree. In principle, this estimation can be done by maximizing the likelihood over both by, for instance, alternating between estimation of the tree for fixed rate distribution parameters and estimation of the rate distribution parameters with a fixed tree. However, the estimation of a tree is very computationally expensive, and so, while this is the preferred approach in principle, we have not implemented it with any of the forms of estimation below. In all cases, the rate distribution parameters are estimated with a fixed tree; to further avoid computational difficulties, in most cases the tree was estimated by means other than maximum-likelihood estimation. In such implementations, if the initial distribution used in the estimation (usually a gamma model) is determined to be unacceptable as a rate distribution model, one should re-estimate the tree and check for differences. In the examples considered here, the gamma distribution was the initial distribution and was not found to be in gross error. For the eubacterial hsp-70 data set where the gamma model seemed in greatest doubt, re-estimation of the tree with a discrete estimate of the rate distribution gave the same topology as the gamma rate distribution and similar, but on average shorter, branch lengths (see Fig. 1).

Gene Families Examined

We fit rate distributions to a number of different data sets. The data sets are available at http://www.treebase.org/treebase under study accession number S910; the matrix accession numbers are indicated in Table 1. The data sets were β-tubulin; eukaryote, plant, archaebacterial, and eubacterial forms of hydroxymethylglutaryl-CoA (HMG-CoA) reductase; eubacterial plus mitochondrial-targeted chaperonin 60 sequences; eubacterial and mitochondrial heat shock protein (hsp) 70 sequences; cytosolic homologs of hsp-90 from eukaryotes; cytosolic-type malate dehydrogenase (cMDH) from eukaryotes, eubacteria, and chloroplasts; archaebacterial and eukaryotic elongation factor 1-α, (EF-1α); eukaryotic HBS1; and eukaryotic release factor 3 (eRF3).

All alignments were performed using the progressive alignment method implemented in ClustalW 1.8 (Thompson et al., 1994) with default settings. Alignments were inspected by eye, and regions of ambiguous alignment were removed. Alignments are available upon request from the authors. Phylogenies were inferred by first estimating a maximum-likelihood distance matrix using TREE-PUZZLE with an eight-category gamma distribution (DGE) model of rate variation and the PAM amino acid substitution matrix (Dayhoff and Eck, 1968; Dayhoff et al., 1979). The Fitch–Margoliash method (implemented in FITCH; Felsenstein, 1993) was used to infer trees from the distance matrices by multiple random stepwise addition replicates and global rearrangements. Branch lengths of the optimal topology were then re-estimated under maximum likelihood with model specifications as above. The β-tubulin data set was used as the main data set for illustration and was treated differently. Partly to illustrate the difficulties of rate estimation with small number of rate categories, four rate categories were used with this data set, and because it was the main illustrative example, full maximum-likelihood estimation was used to obtain the tree.

Rate Distribution Estimation

As an example of rate distribution estimation we consider in detail the aligned amino acid data set of β-tubulin composed of 431 sites for 22 taxa. The log likelihood for the model with a single rate was −3319.73246. By comparison the log likelihoods for all of the models that allow rates to vary according to a distribution were in the range of −3170 to −3150, suggesting that for these data some model of rate variation is needed.

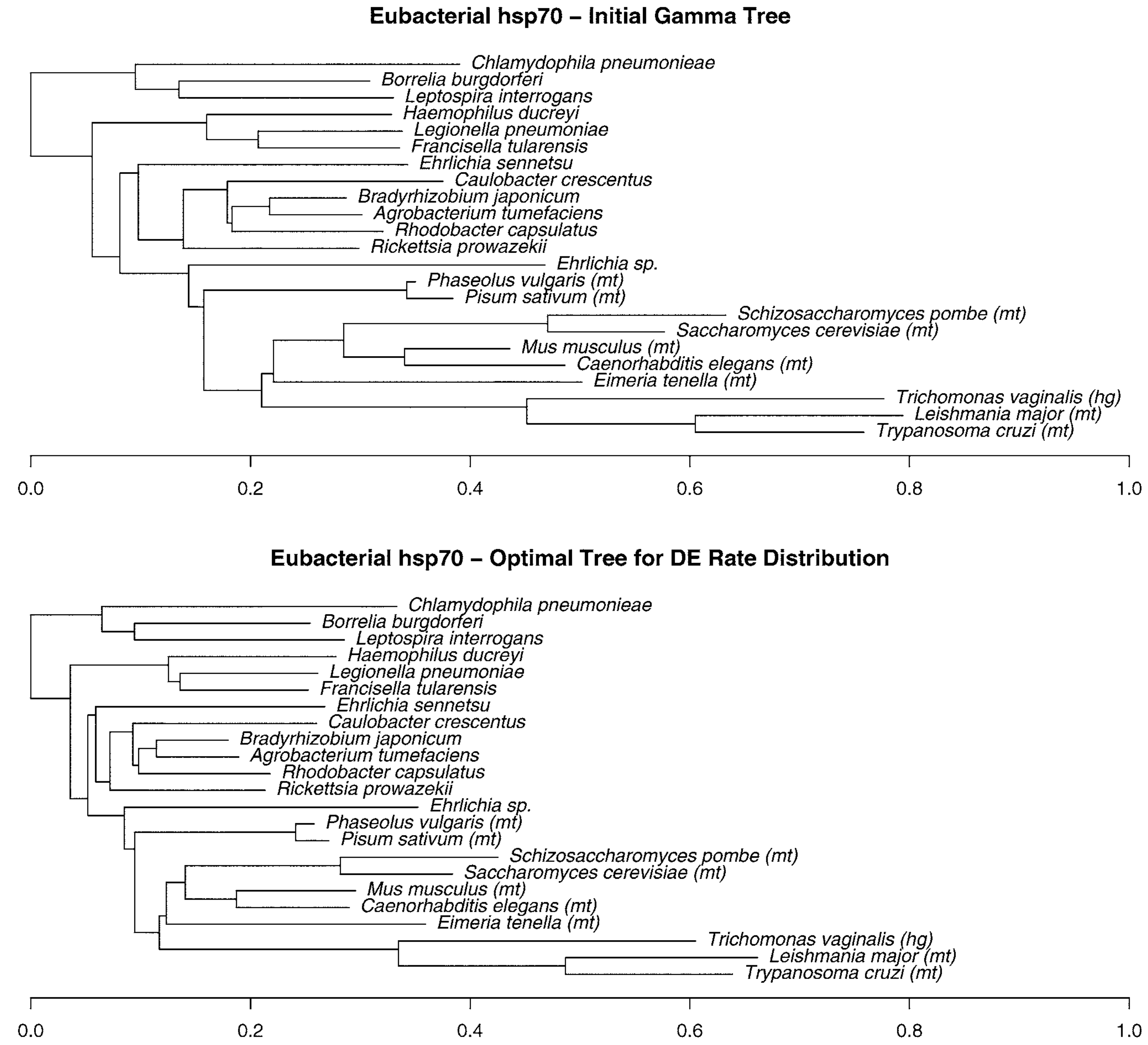

FIGURE 1. Plots of the estimated trees for the eubacterial hsp70 data set. Both trees were estimated using the Fitch–Margoliash method. Branch lengths were then re-estimated under maximum likelihood. The top panel gives the estimated tree with eight category DGE distances obtained using TREE-PUZZLE. The bottom panel gives the tree estimated using DE distances.

The plots of the estimated cumulative distribution function, $G(r) = \sum_{r_j \leq r} \zeta_j$, for the DE and the gamma cumulative distribution function for the $\alpha = 0.271$ resulting from DGPE are given in Figure 2; both the DE and DGPE were based on estimates with 101 rates between 0 and 10. The log likelihood for the discrete estimate was −3153.713 and that for the DGPE was −3154.249. The likelihoods for the models are similar, and the basic patterns in the distributions are similar, both having a rapid initial increase and then a leveling off (Gu et al., 1995; Waddell et al., 1997). The distributions are, however, quite different with the DE, exhibiting several large steps. This is not an optimization artifact. Estimation was done using the general constrained optimization algorithm VE11AD in the HSL and cross-checked with the E04UCF routine in the Numerical Algorithms Group FORTRAN libraries; convergence criteria were met.

With a large number of grid points, it was quite common for the DE to assign probability to only a few rates. This happens because the probabilities $f(x\,|\,r)$ of data given the rate are almost the same for rates r that are close together. Because of this, a rate distribution that assigns relatively large probability to a few rates, say $r_1, \ldots, r_k$, will give almost the same likelihood as a rate distribution that, for instance, assigns approximately the same probability as was assigned to each of the r_i spread

TABLE 1. Differences in log likelihoods for the DE and DGPE models for several data sets. The P value is given for a test of the null hypothesis that the gamma model is appropriate as it is the critical value for a 0.05-level test; both were determined by parametric bootstrapping.

Accession no.	Data set	No. sites	No. taxa	Difference	Cutoff	P
M1508	β-tubulin	431	22	0.54	4.12	0.932
M1500	eukaryotic, EF-1α	269	27	3.00	7.33	0.652
M1499	eubacterial, chaperonin 60	513	22	3.58	7.77	0.532
M1496	archaebacterial, HMG-CoA	251	11	2.40	5.18	0.440
M1506	plant, HMG-CoA	251	15	1.58	3.05	0.290
M1501	eukaryotic, HMG-CoA	251	17	2.84	5.27	0.270
M1497	archaebacterial, EF-1α	269	13	3.14	4.69	0.175
M1505	eukaryotic, cMDH	319	16	3.33	3.94	0.072
M1502	eukaryotic, HBS1	269	13	6.02	5.24	0.020
M1504	eukaryotic, hsp-90	547	37	10.69	7.90	0.008
M1498	eacterial, HMG-CoA	251	14	8.15	5.23	0.002
M1507	eukaryotic, eRF3	269	17	11.24	5.54	0.000
M1503	eubacterial, hsp-70	479	23	15.38	6.48	0.000

out over a large number of rates that are very close to r_i. Nevertheless, because the $f(x \mid r)$ are usually quite different for rates that are further apart, estimates that involve summing over rates, such as the cumulative distribution function and the conditional mean rate at a site, tend to be quite stable.

There was only a small difference in log likelihoods between the gamma and discrete models for this data set. Because the DE is a maximizer of the log likelihood over a very flexible family of distributions, one can conclude that estimation with any other family of rate distributions would similarly show a small difference in likelihood; otherwise, a discrete approximation to it would have given the DE. This similarity is illustrated in the present case by the log likelihood for the gamma + invariable sites model, which gave a log likelihood of −3154.242, only marginally larger than the log likelihood of −3154.29 for the gamma model.

To check whether the estimate of the tree would change significantly with a different rate distribution, we reestimated the tree using the DE as the rate distribution. The resulting log likelihood was −3153.35. The topology was the same, and there were only small changes in some of the branch lengths. This result reinforces our point that more iterations of estimating the rates and then the tree may not be necessary.

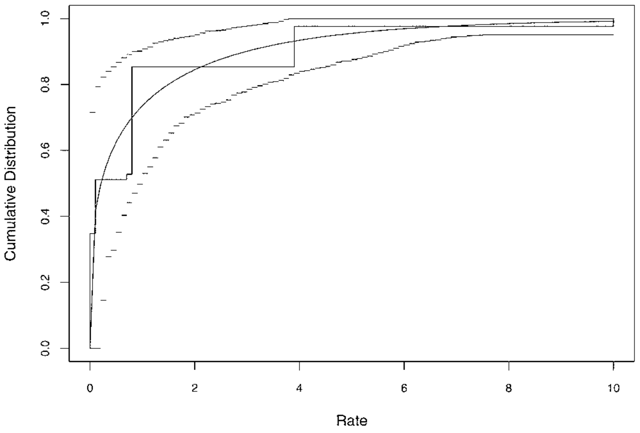

FIGURE 2. The DE and DGPE rate cumulative distribution estimates and 95% bootstrap confidence bounds for the β-tubulin data. The DE and DGPE rate distribution estimates were estimated with 101 rates equally spaced between 0 and 10.

LIKELIHOOD RATIO TESTS FOR THE FIT OF THE GAMMA MODEL

The gamma distribution seems to provide a reasonable model for the β-tubulin data. To check whether this property was more widely applicable, we obtained the likelihood differences between DE and DGPE for a number of other data sets. These were then used in likelihood ratio tests of the null hypothesis that the gamma model is appropriate. The results are given in Table 1, sorted with respect to P value in decreasing order. The estimated P values range from 0.000 to 0.932. Five of the data sets have P values <0.05, with one other having a relatively small estimated P value of 0.072. For the other seven data sets, there is no significant evidence of a departure from the gamma model.

Two data sets have P values between 0.01 and 0.10, and four data sets have estimated P values <0.01, with two of these having P values estimated as 0.000. Nevertheless, none of the log-likelihood differences were extremely large, suggesting that although there is significant evidence of a departure from the gamma model, the departure may not be an indication of a very different rate distribution. To investigate this possibility further, we constructed confidence bounds for the true rate distribution for the data set (eubacterial hsp-70) that gave the largest difference in log likelihoods. Approximate 95% confidence bounds for the rate distribution were constructed using the bootstrap methodology, discussed in more detail below. The result is given in Figure 3. One can see that the estimated gamma cumulative rate distribution is well within the bounds in the region where the rate distribution increases most quickly. The distribution falls outside of the bounds only for large rates. In this case, the data suggest that it is unlikely that the rate distribution has mass at large rates yet the gamma model requires positive probability everywhere.

The results reported in Table 1 are based on parametric bootstrapping; because many of the estimated

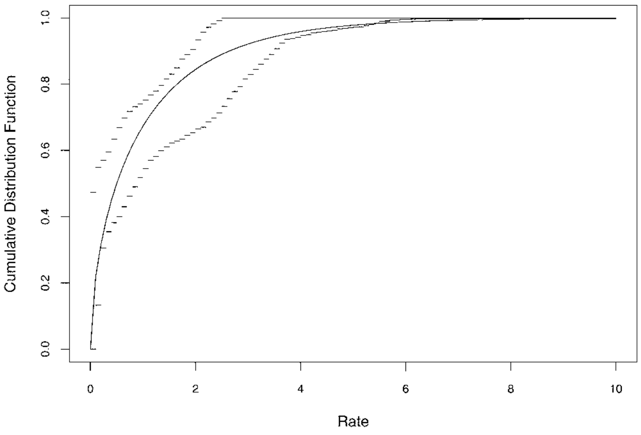

FIGURE 3. The gamma rate distribution estimates and 95% bootstrap confidence bounds for the eubacterial hsp-70 data set. The dashed line gives the cumulative distribution function for a gamma distribution with $\alpha = 0.563$, the DGPE of α.

probabilities in the DE are on the boundary of the parameter space it is unlikely that usual likelihood theory is applicable here. To approximate the distribution of the differences in log likelihoods under the null hypothesis, we repeatedly simulated data from a gamma model and obtained the differences in log likelihoods between the DE and the DGE for each of the simulated data sets. The critical value for an α-level test can then be estimated as the $(1 - \alpha) \times 100$th percentile of the differences in log likelihoods from the simulated data sets and the P value as the proportion of simulated samples with larger log-likelihood differences than the difference observed for the actual data. We made these estimations separately for each of the data sets that were analyzed. Table 1 reports the critical values (cutoff) for a 0.05-level test and P values for each of the data sets. Simulated data were generated using Pseq-Gen (Rambaut and Grassley, 1997) with the estimated tree and estimated DGPE α value. A total of 1,000 simulated samples were generated for each data set. The DE and DGPE were estimated with 101 equally spaced rates.

BOOTSTRAP ESTIMATION OF CONFIDENCE BOUNDS

The discrete rate distribution estimates can be used to obtain bootstrap confidence bounds for the actual rate distribution. We used this method to obtain confidence bounds for the cumulative rate distribution at each of the rates. The procedure is described as follows:

1. Sample *with replacement* sites from the n sites available and record the data at these sites as $x_1^*, \ldots, x_n^*$.
2. Estimate the discrete rate distribution with the selected sites.
3. Store the values of the estimated cumulative distribution function at the grid points $G(r_1)^*, \ldots, G(r_k)^*$.

Steps 1–3 are repeated a large number of times. Upon conclusion, the lower and upper bounds for a $(1 - \alpha) \times 100\%$ confidence interval for $G(r_j)$ are taken as the $\alpha/2$th and $(1 - \alpha/2)$th percentiles of the generated $G(r_j)^*$. This method of bootstrapping is sometimes referred to as the percentile method (Davison and Hinkley, 1997: sect. 5.3) and has been shown to give appropriate coverage properties in a number of cases.

The bootstrap bounds calculated through the above method are marginal bounds: the probability is approximately $1 - \alpha$ that the actual rate distribution at r_j will be within the bounds given. The bounds are not however simultaneous: the probability is less than $1 - \alpha$ that the actual distribution will be contained within the bounds simultaneously at each r_j. To change the bounds to simultaneous bounds we used the above strategy to obtain the shape of the initial bounds. Let l_r and u_r denote the resulting lower and upper bounds at rate r. Logistic transformations were applied to these bounds, giving new bounds $\log(l_r/[1 - l_r])$ and $\log(u_r/[1 - u_r])$ on the logistic scale. A bisection algorithm was then used to determine the smallest constant $a > 0$ so that $(1 - \alpha) \times 100\%$ of the generated rate distributions G^* satisfied that

$$\begin{aligned}\log(l_r/[1 - l_r]) - a &\leq \log\{G^*(r)/[1 - G^*(r)]\} \\ &\leq \log(u_r/[1 - u_r]) + a\end{aligned}$$

simultaneously for all a. This process gives simultaneous bounds on the logit scale. To transform back to the cumulative distribution scale, a bound b was transformed to $\exp(b)/[1 + \exp(b)]$. Note that this same additive adjustment could have been made on the original scale; however, it can lead to a violation of the restriction that the bounds be between 0 and 1.

The bootstrap bounds based on 1,000 bootstrap simulations for the β-tubulin data are included in Figure 2. The gamma estimate is within the bootstrap bounds, and the bounds are tight: No distribution with a single mass point (which corresponds to not requiring a rate distribution) would fall within the bounds, and any reasonable rate distribution would have to have a rapid rate of increase.

RATE ESTIMATION

Estimating the rate of evolution of sites in proteins has recently been recognized as important to the undestanding and/or prediction of their functional or structural properties (Gaucher et al., 2001; Simon et al., 2002; Blouin et al., 2003; Blouin et al., unpubl.). Estimates of rates are most naturally constructed from the conditional distribution of rates given the data at a site, which can by obtained through Bayes's formula:

$$p(r_j \mid x) = f(x \mid r_j)\zeta_j \bigg/ \left(\sum_j f(x \mid r_j)\zeta_j \right).$$

The most common rate estimate is the conditional mode: the rate giving the largest conditional probability $p(r_j \mid x)$. An alternative rate estimate that is sometimes

TABLE 2. The rate distribution estimates for the DE (maximum likelihood, ML) and the DGE from TREE-PUZZLE for the β-tubulin data set.

	DGE		
ML rate estimate	0	0.551	3.393
0	298	0	0
0.8	0	51	35
3.9	0	0	42
10	0	0	5

used is the conditional mean estimate:

$$E[r \mid x] = \sum_{j} r_j \, p(r_j \mid x),$$

which has the optimality property of giving minimal expected mean square error.

Different rate estimates are obtained from different methods of rate distribution estimation. Our interest here is in contrasting a "conventional" rate estimates (DGE) with more flexible estimates. We consider the differences that result using the DGE and DE of rate distributions for the β-tubulin amino acid data set.

Table 2 gives the conditional mode rate estimates using the DE, estimated with 101 rates, of the rate distribution cross-tabulated against the conditional mode rate estimates from the DGE obtained from TREE-PUZZLE with four rate categories. Many of the rates are similar: 51 of the sites have rates that are estimated to be 0.551 by TREE-PUZZLE, and these are estimated as 0.8 by using the discrete rate distribution estimate. Similarly 42 of the rates are estimated as 3.9 using the discrete distribution and as 3.933 by TREE-PUZZLE. One difference that is worth noting is the five sites that were estimated to have a large rate of 10 using the discrete distribution but that were estimated to have a rate of 3.393 by the discrete gamma. The estimated conditional probability $p(10 \mid x)$ for these sites provides an indication of how likely it is that the rates at these sites are much larger than 4. For the five sites with estimated rates of 10, these estimated conditional probabilities were 0.89, 0.77, 0.56, 0.99, and 0.93, giving strong evidence for an estimated rate of 10. The difficulty for the gamma estimates is the discrete nature of the approximation used by the discrete gamma model. The largest rate category is always of the form (b, ∞), with b chosen so that the probability of the interval under the gamma model is $1/k$. For many α parameters, this value will be small, and so any rate that is very large will be assigned the relatively small rate b.

Figure 4 includes scatter plots comparing the conditional mean and mode estimates corresponding to the different forms of rate distribution estimation for the β-tubulin data set. The DE and DGPE rate distribution estimates had mass on 101 values between 0 and 10. The DGE is for a discrete gamma model obtained from TREE-PUZZLE with four equal-probability rate categories. As with the conditional mode estimates, a major feature is that a number of sites are estimated to have large rates using the DE but relatively small estimated rates based on the DGE. This is true when comparisons are made between the DGE and DGPE estimates as well. The discrepancy between the DGE and DGPE, which are both based on a gamma model, indicates that the DGE are in error largely because of the large last rate category. This error can be avoided by using a larger number of categories when fitting the models. In practice, however, a small number of categories is often used. Moreover, while the use of a small number of rate categories may not appreciably affect the estimation of topology and will make computation more manageable, the example indicates that the rate estimates may be poor. One possible adjustment would be to use a DGE for estimation of the topology and then use DGPE to estimate the rates afterwards.

The DGPE mean rate estimates for the β-tubulin data set are highly correlated with but different from the DGPE mode rate estimates. The majority of mode rate estimates are smaller than the mean estimates, with a few large rate estimates providing exceptions. Plots of conditional distributions or rates for the sites where the mode and mean estimates differed most indicated that skewness in these distributions caused these estimates to differ.

One of the other features of note in Figure 4 is the agreement between the DE mean rates and DGPE mean rates. For the β-tubulin data set, the likelihood ratio test for the null hypothesis that the gamma model is appropriate gave a P value of 0.932. Because there is no significant evidence that the gamma model is inappropriate, the rate estimates can be expected to be similar and this turned out to be the case. In contrast, the likelihood ratio test for the eubacterial hsp-70 data set gave a P value that was estimated as 0.000. Although the result of the test indicates a significant departure from the gamma model, the plot of the gamma distribution function with bootstrap bounds (Fig. 3), suggests that the departure is not large in magnitude. Because the gamma distribution falls outside of the bootstrap bounds only for large rates, and in this case falls below the bootstrap bounds, the plot suggests that the gamma distribution places a little too much mass at larger rates. The implication of this bias for rate estimation should be that large DGPE rate estimates will be larger than the corresponding DE rate estimates. This indeed turns out to be the case, as is illustrated in Figure 5, which plots the DGPE means against the DE means. Some of the smaller rate estimates also appear to be underestimated. For this data set, if an analysis were being conducted where rate estimation was important, the DE rate estimates would be preferred because they do not assume a gamma model, which in this case can be rejected.

DISCUSSION

The rate distribution cannot be expected to be fully recovered from character data alone. Nevertheless significant information can be obtained from data. The bounds for the rate distribution given in Figure 2 are tight and

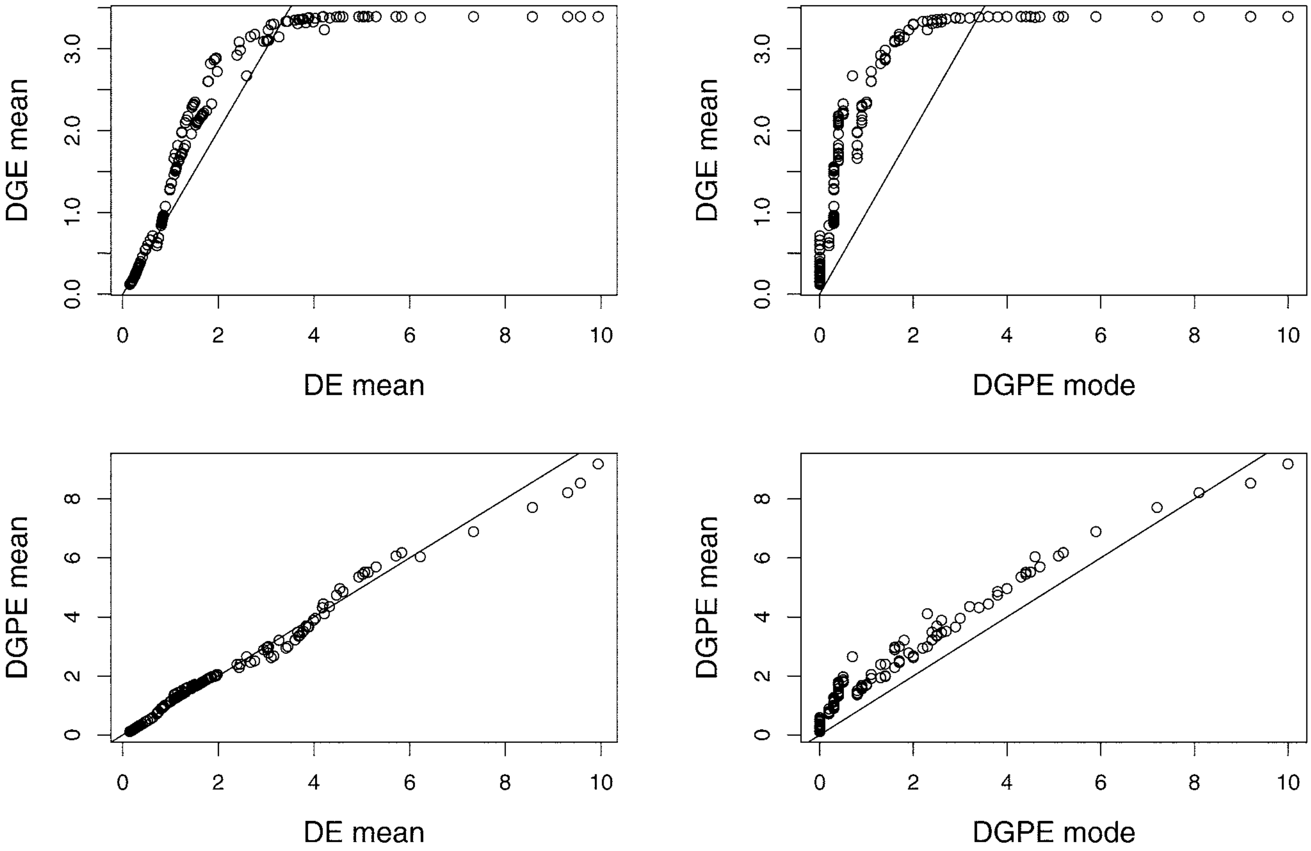

FIGURE 4. Scatter plots comparing the different conditional mean and mode estimates corresponding to the different forms of rate distribution estimation for the β-tubulin data set. The DE and DGPE rate distribution estimates had mass on 101 values between 0 and 10. The DGE is for a discrete gamma with four equal-probability rate categories.

do not allow for many choices of rate distribution. Moreover, at least for the types of data sets that have been considered here, the gamma distribution often provides a reasonable model for the rate distribution (similar conclusions were arrived at by Waddell et al., 1997). The appropriateness of the gamma model was inferred by comparison with the discrete model, which provides a very flexible family of rates distributions that can be used for comparison with a parametric family and to construct bootstrap confidence bounds. In constrast, likelihood ratio comparisons between the gamma model and other parametric models usually suffer from the difficulty that the models are nonnested and always suffer from the fact that the alternative parametric families imply restrictions on the alternate form of the rate distribution.

One of the other conclusions that comes out of the analysis is that for estimation of the rate at a site, it may be valuable to consider a DGPE with a larger number of rate categories rather than those for the DGE. The heavy right tail of the gamma distribution creates a very large last rate category so that a few sites with rates that should be inferred as much larger than the rest will be grouped into this largest rate category using the DGE. Sites with large rates are important because they are often suspected of being in locations in the sequence that are not functionally or spatially constrained for the organisms under study. Accurate identification of unconstrained versus constrained sites is required if inferences regarding function are to be made from analyses of site rates. Finally, in cases where rate estimation is important and a gamma model can be rejected, the DE mean rate estimates, which do not require a gamma assumption, provide a robust alternative to gamma estimates. Software to obtain the DE, DGPE rate distributions, and rate estimates is available at http://www.mathstat.dal.ca/~tsusko.

Although we did not find much evidence for changed tree topologies, the rate distribution can in principle make a big difference to estimation, especially if there are long branches present (Huelsenbeck, 1995; Sullivan and Swofford, 2001). Ideally, the shape parameter of the gamma rate distribution should be re-estimated for every tree examined during the course of maximum-likelihood estimation. Computational limitations associated with the recalculation of site likelihoods when optimizing α using the standard DGE discretization have made this re-estimation all but impossible. However, implementation of the DGPE rate estimation with the number of rate categories comparable to that of the traditional DGE

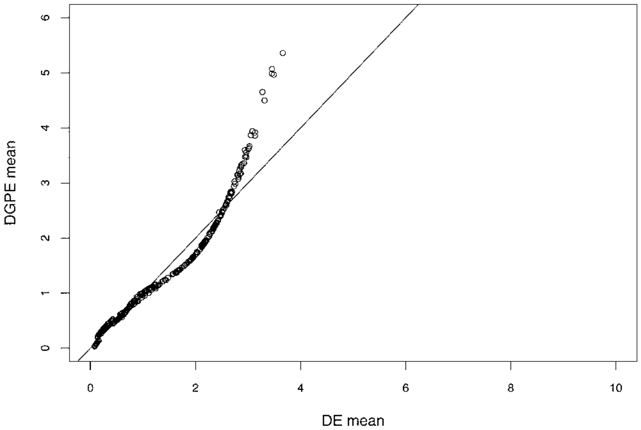

FIGURE 5. A scatter plot comparing the DGPE conditional means with the DE conditional means for the eubacterial hsp-70 data set.

method should allow for a much greater speed of computation and could permit the re-estimation of the rate distribution for all trees examined, potentially improving the estimation of topologies. Furthermore, use of DE and DGPE distributions with a large number of rates and a fixed tree should allow for much more accurate estimation of the rates of evolution at sites. This will be extremely useful to those developing more realistic models of protein evolution.

Yang et al. (2000) considered a similar problem in which the rate of nonsynomymous to synonymous substitutions is allowed to vary across sites. They considered 10 parametric families as models for the distribution of these rates. Although these are a different type of rate than what is being considered here, for 7 of the 10 genes they considered the difference between the log likelihood for a gamma model and the best other parametric model was less than 4. For one of the data sets, the difference was 20, and for two others it was larger than 5. A discrete model, such as the DE considered here but with a small number of categories, fit well for all of the data sets.

The discussion here assumed a model where rates vary across sites but are constant at a given site. Recent work by Lockhart et al. (1998), Galtier (2001), and Susko et al. (2002), among others, has suggested that the rates of molecular evolution often vary across subtrees of the larger evolutionary tree as well as across sites.

ACKNOWLEDGMENTS

E.S. and C.F are supported by the Natural Sciences and Engineering Research Council of Canada. A.J.R. thanks the Canadian Institute for Advanced Research Program in Evolutionary Biology for fellowship support. A.J.R. and C.B. were supported by NSERC Operating Grant 227085-00 and NSERC Genomics Grant 228263-99. This collaboration is part of a Genome Atlantic/Genome Canada Large-Scale Project.

REFERENCES

BLOUIN, C., Y. BOUCHER, AND A. ROGER. 2003. Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. Nucleic Acids Res. 31:790–797.

BLOUIN, C., E. SUSKO, AND A. ROGER. 2002. Structural determinants of evolutionary constraints in enolase. Preprint.

DAVISON, A. C., AND D. V. HINKLEY. 1997. Bootstrap methods and their application. Cambridge Univ. Press, New York.

DAYHOFF, M. O., AND R. V. ECK. 1968. Atlas of protein sequence and structure 1967–1968. National Biomedical Research Foundation, Silver Spring, Maryland.

DAYHOFF, M. O., R. M. SCHWARTZ, AND B. C. ORCUTT. 1979. A model of evolutionary change in proteins. Pages 345–352 *in* Atlas of protein sequence and structure, Volume 5, supplement 3 (M. O. Dayhoff, ed.). National Biomedical Research Foundation, Silver Spring, Maryland.

FELSENSTEIN, J. 1993. PHYLIP (phylogeny inference package), version 3.5c. Distributed by the author, Department of Genetics, Univ. Washington, Seattle.

FELSENSTEIN, J., AND G. A. CHURCHILL. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13:93–104.

FITCH, W. M., AND E. MARKOWITZ. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4:579–593.

GALTIER, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18:866–873.

GAUCHER, E. A., M. M. MIYAMOTO, AND S. A. BENNER. 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. Proc. Natl. Acad. Sci. USA 98:548–552.

GU, X., Y. FU, AND W. LI. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12:546–557.

HUELSENBECK, J. P. 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

HUELSENBECK, J. P. 1997. Is the Felsenstein zone a fly trap? Syst. Biol. 46:69–74.

LOCKHART, P. J., D. H. HUSON, U. MAIRER, M. J. FRAUNHOLZ, Y. VAN DE PEER, A. C. BARBROOK, C. J. HOWE, AND M. A. STEEL. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. 15:1183–1188.

LOPEZ, P., D. CASANE, AND H. PHILIPPE. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19:1–7.

NEI, M. 1987. Molecular evolutionary genetics. Columbia Univ. Press, New York.

RAMBAUT, A., AND N. C. GRASSLY. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

SILBERMAN, J. D., C. G. CLARK, L. S. DIAMOND, AND M. L. SOGIN. 1999. Phylogeny of the genera *Entamoeba* and *Endolimax* as deduced from small-subunit ribosomal RNA sequences. Mol. Biol. Evol. 16:1740–1751.

SIMON, A. L., E. A. STONE, AND A. SIDOW. 2002. Inference of functional regions in proteins by quantification of evolutionary constraints. Proc. Natl. Acad. Sci. USA 99:2912–2917.

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., AND D. L. SWOFFORD. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst. Biol. 50:723–729.

SUSKO, E., Y. INAGAKI, C. FIELD, M. E. HOLDER, AND A. J. ROGER. 2002. Testing for differences in rates across sites distributions in phylogenetic subtrees. Mol. Biol. Evol. 19:1514–1523.

SWOFFORD, D. L. 2000. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer, Sunderland, Massachusetts.

THOMPSON, J. D., D. G. HIGGINS, AND T. J. GIBSON. 1994. Clustalw: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

UZZEL, T., AND K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. Science 173:1089–1096.

WADDELL, P. J., D. PENNY, AND T. MOORE. 1997. Hadamard conjugations and modeling sequence evolution with unequal rates across sites. Mol. Phylogenet. Evol. 8:33–50

YANG, Z. 1994. Maximum likelihood phylogentic estimation from DNA sequences when substitution rates differ over sites: Approximate methods. J. Mol. Evol. 39:306–314.

YANG, Z. 2000. Phylogenetic analysis by maximum likelihood (PAML), version 3.0. Univ. College, London, U.K. (http://abacus.gene.ucl.ac.uk/software/paml.html).

YANG, Z., R. NIELSEN, N. GOLDMAN, AND A. K. PEDERSEN. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.

First submitted 5 June 2002; reviews returned 19 August 2002; final acceptance 30 March 2003

Associate Editor: Bruce Rannala