The Effect of Taxon Sampling on Estimating Rate Heterogeneity Parameters of Maximum-Likelihood Models

Jack Sullivan, David L. Swofford,† and Gavin J. P. Naylor‡*

*Department of Biological Sciences, University of Idaho; †Laboratory of Molecular Systematics, Smithsonian Institution, and ‡Department of Zoology and Genetics, Iowa State University

As maximum-likelihood approaches to the study of molecular systematics and evolution become both more flexible and more accessible, the importance of understanding the statistical properties of parameter estimation becomes critical. Using variation in NADH-2 sequences for 40 species of requiem sharks, we illustrate that estimates of rate heterogeneity parameters are highly sensitive to taxon sampling when the data are best explained by a mixed-distribution model of among-site rate variation (invariable sites plus gamma distribution [I+Γ]). Using computer simulation, we attempt to differentiate two possible causes of this sensitivity. While the possibility of nonstationarity cannot be definitively rejected, our results suggest that sampling error alone provides an adequate explanation for the pattern of uncertainty observed in estimates from the real data. Furthermore, we illustrate that two parameters estimated under the I+Γ model (the proportion of sites not free to change and the gamma distribution shape parameter) are highly correlated and that the likelihood surface across the rate heterogeneity parameter space can be poorly behaved when only a small number of sequences (taxa) are considered.

Introduction

The use of maximum-likelihood approaches in the study of molecular evolution and phylogeny has increased dramatically in recent years. This is attributable, at least in part, to the development of increasingly realistic models of sequence evolution and their implementation in widely available software for phylogenetic analysis. These models permit a great deal of flexibility in tailoring the assumptions made in phylogenetic inference to a particular data set, including the ability to allow for unequal base frequencies (e.g., Felsenstein 1981), different rates of change among pairs of nucleotides (e.g., Yang 1994*a*) or between nucleotide classes (such as transitions and transversions; Kimura 1980; Hasegawa, Kishino, and Yano 1985), and rate heterogeneity across nucleotide sites (e.g., Yang 1993, 1994*b*; Gu, Fu, and Li 1995; Waddell and Penny 1996).

Although the existence of rate heterogeneity across sites has been known for some time (Fitch and Margoliash 1967; Uzzell and Corbin 1971), its importance in evolutionary studies has received considerable attention recently (reviewed in Yang 1996*a*). The two most commonly used methods for explicitly dealing with among-site rate variation are the invariable-sites model (e.g., Fitch and Margoliash 1967; Hasegawa, Kishino, and Yano 1985), in which some proportion of sites (p_{inv}) is assumed to be completely resistant to change, with all variable sites assumed to evolve at the same rate, and the gamma-distributed-rates model, in which the distribution of relative rates over sites is assumed to follow a gamma distribution (Uzzell and Corbin 1971; Yang 1994*b*) whose shape parameter (α) determines the strength of rate heterogeneity. Recently, Gu, Fu, and Li (1995) and Waddell and Penny (1996) have suggested combining these models such that some sites are assumed to be invariable and rates at the remaining sites follow a gamma distribution. This has been called the invariable-sites-plus-gamma (I+Γ) model and is a mixture of a discrete distribution and a continuous distribution. In the example provided by Gu, Fu, and Li (1995), the application of this mixed-distribution model to real data did not lead to a significant improvement in fit relative to a gamma-distributed-rates model alone. Nevertheless, the I+Γ model is intuitively very appealing (e.g., Tourasse and Gouy 1997), and we have examined several data sets (see, e.g., Sullivan and Swofford 1997) in which use of this mixed-distribution model does significantly improve the fit (as assessed by a likelihood ratio test) relative to either invariable sites or gamma-distributed rates alone.

Several studies have demonstrated the importance of incorporating rate heterogeneity into phylogenetic models. The advantages of this approach include increased accuracy in estimation of branch lengths (e.g., Gu, Fu, and Li 1995), consistency of phylogenetic estimation under a greater range of conditions than is achievable if heterogeneity is ignored (e.g., Gaut and Lewis 1995; Huelsenbeck 1995*a*), and improved accuracy in the estimation of reliability of phylogenetic inferences (Yang 1996*b*; Sullivan, Markert, and Kilpatrick 1997). However, the use of realistic models is not without its cost, as complex models suffer from an inflation of variance relative to simpler models because more parameters must be estimated from the same amount of data. In addition, computational intensity rises with increasing model complexity. The high computational demand of maximum-likelihood analyses renders impractical the ideal strategy of simultaneously optimizing all substitution model parameters, in addition to branch lengths, for every tree examined during a tree search for data sets with many taxa. Thus, as for parsimony analyses, effective heuristic methods are required for model-

Abbreviations: α, gamma-distribution shape parameter; GTR, general time-reversible model of sequence evolution; p_{inv}, proportion of sites that are invariable.

Key words: rate heterogeneity, parameter estimation, maximum likelihood, molecular phylogeny, parametric bootstrap, simulation, stationarity.

Address for correspondence and reprints: Jack Sullivan, Department of Biological Sciences, University of Idaho, Moscow, Idaho 83844-3051. E-mail: sullivan@onyx.si.edu.

Mol. Biol. Evol. 16(10):1347–1356. 1999

based analyses (e.g., Rogers and Swofford 1999). Yang (1994*b*) has suggested that the computational demands of maximum-likelihood analyses can be reduced by estimating model parameters with an initial analysis that includes only a small set of the taxa under study. However, during an examination of the effect of topology on estimates of rate heterogeneity parameters, Sullivan, Holsinger, and Simon (1996) observed that several estimates of the gamma distribution shape parameter derived from subsets of four taxa were significantly different from the value estimated from the entire (10-taxon) data set; a similar effect was also observed by Hershkovitz and Lewis (1996). In this paper, we explore the uncertainty of maximum-likelihood estimates of rate heterogeneity parameters associated with taxon sampling using analyses of both real and simulated data sets for a moderately large number (40) of taxa.

Analyses of Real Data

The data set we examine consists of 1,307 aligned nucleotides from the mitochondrial NADH-2 gene (1,047 nt) and flanking tRNA sequences (260 nt) from 40 species of requiem sharks (unpublished data). We restricted our attention to the single topology produced by a heuristic search under the minimum-evolution criterion using log determinant (LogDet, paralinear) distances (Lake 1994; Lockhart et al. 1994). This tree (see appendix) is characterized by many short internal branches. Our analyses of the real data, however, do not assume that this LogDet tree is the true tree. Work by Yang (1994*b*) and Sullivan, Holsinger, and Simon (1996), as well as our own experience with many other data sets, has demonstrated that for any given sample of taxa, the dependence of model parameter estimates on topology is minor as long as strongly supported groups are maintained. (These sequences, as well as the LogDet tree, are available on request from G.J.P.N.) All analyses were conducted using test versions of the PAUP* phylogenetic inference program (v4.0d45-d61, written by D.L.S.). For each combination of substitution model parameters (i.e., all parameters other than branch lengths) and tree topology, PAUP* searches for an optimal set of branch lengths by making multiple passes over the tree, adjusting one branch length at a time using standard Newton-Raphson iteration (e.g., Edwards 1972), until convergence is achieved. As noted by Steele (1994), this strategy may not be effective if there are multiple optima on the likelihood surface. However, Rogers and Swofford (1999) have provided evidence that multiple peaks in branch length space are unlikely to occur on trees that rank highly according to the likelihood criterion. To search for optimal substitution model parameter values, PAUP* uses Brent's (1973) modification of Powell's (1964) conjugate-direction-set method. Like other multidimensional optimization methods, Powell's algorithm does not guarantee that a final solution will be globally optimal. For many problems, this derivative-free method performs well in comparison with other available approaches (e.g., the variable-metric and conjugate-gradient methods) for which derivatives would need to be approximated numerically, although we have not performed a rigorous evaluation of other possible optimization strategies.

Table 1
Evaluation of Models of Sequence Evolution for the Mitochondrial NADH-2 Genes from 40 Species of Requiem Sharks

Model	ln Likelihood	df	χ^2
JC	−14,386.90334	7	7,324.52758
JC + Γ	−12,749.95859	6	4,050.63808
JC + I	−13,035.09740	6	4,620.91570
JC + I + Γ	−12,743.06606	5	4,036.85302
K2P	−12,888.96174	6	4,328.64438
K2P + Γ	−11,133.08083	5	816.88256
K2P + I	−11,469.28651	5	1,489.29329
K2P + I + Γ	−11,113.34975	4	777.42040
HKY85	−12,451.66573	6	3,454.05236
HKY85 + Γ	−10,822.65716	5	196.03522
HKY85 + I	−11,075.53078	5	701.76246
HKY85 + I + Γ	−10,794.92486	4	140.57062
GTR	−12,245.82493	2	3,042.37076
GTR + Γ	−10,765.24546	1	81.21182
GTR + I	−10,978.32540	1	507.37170
GTR + I + Γ	−10,724.63955	—	—

NOTE.—Models were evaluated on the minimum-evolution tree generated from LogDet distances. All of these are special cases of GTR + I + Γ, and all fit the data significantly worse than this model ($P \ll 0.001$). All relevant parameters for each model were simultaneously optimized, except that for models not assuming equal base frequencies (the HKY85 and GTR models), the base frequencies were fixed to mean values from the data ($\pi_A = 0.321$, $\pi_C = 0.288$, $\pi_G = 0.100$, $\pi_T = 0.291$).

In order to identify an appropriate model for these data, we calculated the likelihood scores for the LogDet tree under four substitution matrices: the Jukes-Cantor (JC; Jukes and Cantor 1969), Kimura two-parameter (K2P; Kimura 1980), Hasegawa-Kishino-Yano (HKY85; Hasegawa, Kishino, and Yano 1985), and general time-reversible (GTR; equals REV of Yang 1994*a*). In addition, four rate heterogeneity models were examined: (1) equal rates; 2) a proportion of sites assumed to be invariable (p_{inv}), with equal rates assumed at variable sites (I; Hasegawa, Kishino, and Yano 1985); (3) rates at all sites assumed to follow a gamma distribution (Γ; Yang 1994*b*); and (4) a mixture of invariable sites plus gamma-distributed rates (I+Γ; Gu, Fu, and Li 1995; Waddell and Penny 1996). Thus, 16 models were evaluated, each of which is a special case of the most general parameter-rich GTR+I+Γ model. (Likelihoods for models incorporating gamma-distributed rates were calculated using the discrete gamma approximation of Yang [1994*b*] with four rate categories.) All of the simpler models could be rejected for this data set using a likelihood ratio test with χ^2-approximation of the null distribution (table 1). Although the test statistic is only asymptotically χ^2-distributed, this approximation has been shown through simulation to be very useful in discriminating among nested models of nucleotide substitution (Yang, Goldman, and Friday 1995). The most appropriate substitution model for these data therefore appears to be the GTR+I+Γ model with the following parameters (cf. Yang 1994*a*):

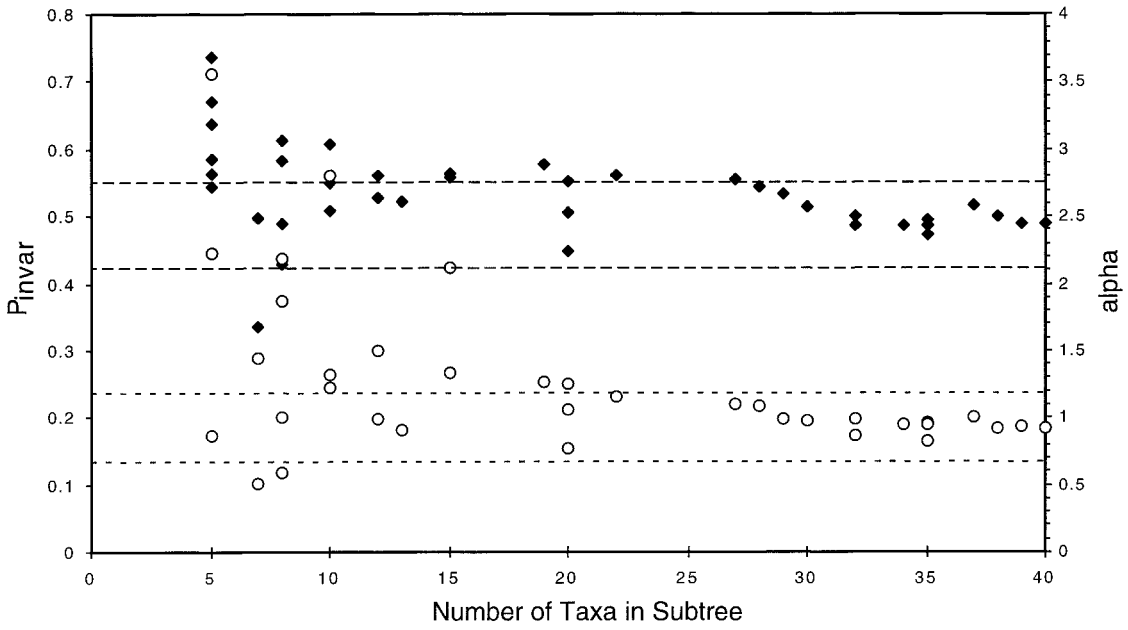

FIG. 1.—The sensitivity of rate heterogeneity parameters to taxon sampling in the shark NADH-2 data set. Rate heterogeneity parameters were simultaneously estimated for each of the subtrees listed in the appendix. Diamonds represent estimates of p_{inv}, and circles represent estimates of α. The dashed and dotted lines represent 95% confidence intervals for p_{inv} and α, respectively, estimated by the parametric bootstrap for the 40-taxon LogDet tree. For one five-taxon subtree, the estimate of α diverged toward infinity; this point was omitted for graphical purposes.

base frequency parameters:

$$\pi_A = 0.321$$

$$\pi_C = 0.288$$

$$\pi_G = 0.100$$

$$\pi_T = 0.291$$

relative-substitution-rate parameters:

$$r_{AC} = 5.852$$

$$r_{AG} = 42.154$$

$$r_{AT} = 4.967$$

$$r_{CG} = 3.864$$

$$r_{CT} = 145.473$$

$$r_{GT} = 1.000$$

rate-heterogeneity parameters:

$$p_{inv} = 0.49$$

$$\alpha = 0.924.$$

All subsequent analyses of the shark data use this model.

To examine the reliability of the estimates of rate heterogeneity parameters derived from the entire set of taxa, we used the parametric bootstrap (Efron and Tibshirani 1993; Huelsenbeck, Hillis, and Jones 1996) to estimate confidence intervals. One hundred 40-taxon data sets with the same sequence length as the original data (1,307 bp) were generated using LogDet topology (with branch lengths estimated from the original data via maximum likelihood) under the above model. We then optimized all parameters for each simulated data set on the model tree. The resulting 95% confidence intervals for rate heterogeneity parameters are rather small ($0.4314 < p_{inv} < 0.5380$; $0.6668 < \alpha < 1.1666$) and indicate that these parameters can be estimated quite reliably for this moderately large number of taxa with moderately long sequences (1,307 bp).

To examine the sensitivity of estimates of rate heterogeneity to taxon sampling, we estimated α and p_{inv} on 40 subtrees of the original LogDet tree containing 5–39 of the original 40 taxa. These subtrees are listed in the appendix; they were chosen to represent both clumped and stratified sampling of taxa, such that in some instances only the most recently diverged taxa were included, and in other instances the included taxa covered the deepest divergence in the tree. The distribution of the rate heterogeneity parameters estimated across the subtrees is illustrated in figure 1. For the most part, subtrees containing >20 taxa yield reliable estimates of the rate heterogeneity parameters. However, subtrees containing 20 or fewer taxa yield highly variable estimates of α and p_{inv}. Even with as few as five taxa, the estimates from some of the subtrees were quite good (within the confidence interval estimated from the entire data set), but even with as many as 22 taxa in the analysis, the estimate of p_{inv} may be outside the confidence interval estimated for the full data set. Furthermore, with five taxa, p_{inv} is sometimes estimated as the limiting value equal to the proportion of sites that are observed to be constant. The effect of taxon sampling on estimates of α are quite similar. Stable estimates (within the 40-taxon confidence interval derived above) are obtained in analyses of >20 taxa, and highly variable estimates are obtained in analyses involving 20 or

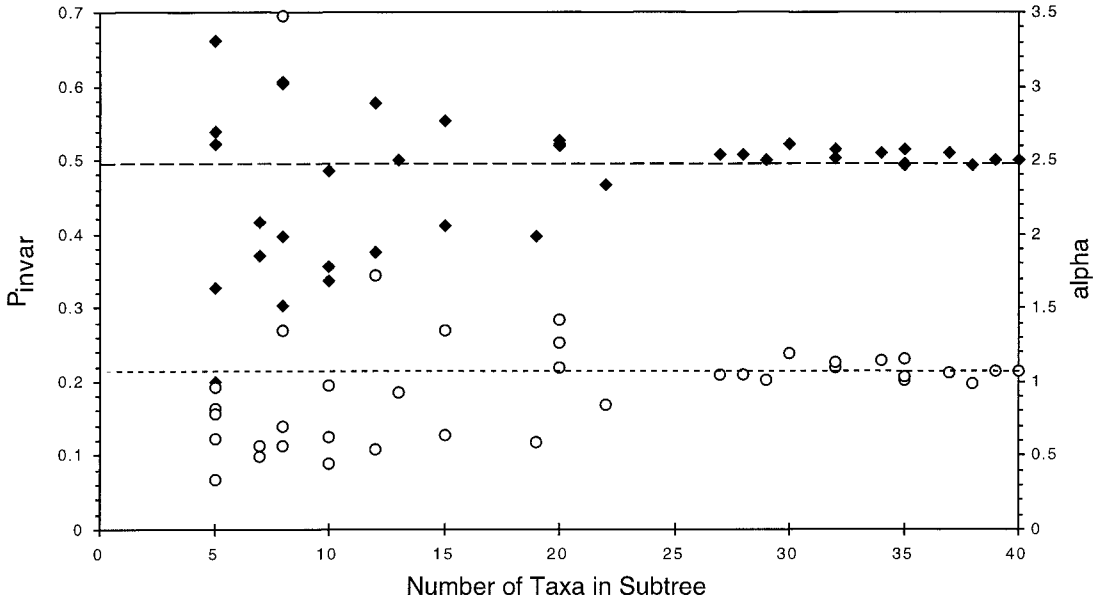

FIG. 2.—The sensitivity of rate heterogeneity parameters to taxon sampling in a representative replicate of the simulated data. Diamonds represent estimates of p_{inv}, and circles represent estimates of α. Sequences were simulated on the LogDet tree under the model with the best fit to the original data. Stationarity of parameters has been enforced for these data so that the observed scatter can only be attributed to sampling error. The same set of subtrees as in figure 1 was used. The dashed and dotted lines represent the true values of p_{inv} and α across the tree. Estimates of α that diverged toward infinity were omitted.

fewer taxa; the estimate of α goes to infinity (identical to equal rates at all variable sites) for one of the five-taxon analyses (point not plotted).

Two possible explanations can account for the observed sensitivity of estimates of rate heterogeneity parameters to taxon sampling. The first is that this sensitivity simply reflects increased susceptibility to stochastic error due to small sample effects. The second is that the data violate the assumption that the rate heterogeneity parameters are stationary, as is expected under a time-reversible homogeneous Markov substitution process. That is, the true value may actually vary in different parts of the tree, as might be the case, for example, if the covarion hypothesis (Fitch and Markowitz 1970) is operating in the evolution of these sequences. We used simulations to attempt to discriminate between these possibilities.

Taxon-Sampling Simulations

These simulations also follow the protocol for the parametric bootstrap (Efron and Tibshirani 1993; Huelsenbeck, Hillis, and Jones 1996), in that the model used for simulation was estimated from the original data (via maximum likelihood). Again, because none of the simpler models could be accepted as an adequate approximation of the most general model for the real data (table 1), sequences were simulated on the LogDet topology with maximum-likelihood branch length estimates under the GTR+I+Γ model defined as above. For each of 10 replicate data sets, the same set of subtrees used in the analyses of the real data set (see appendix) was used to assess the sensitivity of rate heterogeneity parameters to taxon sampling when those parameters were known to be stationary.

If the sensitivity of rate heterogeneity parameters observed in the real data (fig. 1) were the result of nonstationarity and not simply an effect of sampling error, we would expect much less scatter in the distribution of parameter estimates across the subtrees in the simulations in which stationarity is imposed. However, this was not the case; the patterns in sensitivity of estimates of rate heterogeneity parameters to taxon sampling in each replicate of the simulation analyses are quite similar to those of the analyses of real data (compare figs. 1 and 2). Estimates close to the true values of the generating model were obtained consistently only with relatively large numbers of taxa ($>$20). In addition, estimates of p_{inv} range up to the observed proportion of constant sites, and in almost all replicates, some subtrees containing fewer than 10 taxa yield estimates of infinity for α (not plotted in fig. 2 for graphical convenience). Therefore, it seems unlikely that the observed sensitivity to taxon sampling in rate heterogeneity parameter values estimated from the real data (fig. 1) is due primarily to nonstationarity, because very similar patterns of sensitivity to taxon sampling are observed in the simulations (e.g., fig. 2), in which stationarity has been imposed.

We examined the issue of stationarity further by an additional set of simulations. In analyses of both real (fig. 1) and simulated data (e.g., fig. 2), the widest variation in parameter estimates occurs when five-taxon subtrees are used. Thus, we would expect the effect of any nonstationarity to be manifest most strongly in the five-taxon analyses. Therefore, we again used a parametric bootstrap to estimate confidence intervals of rate heterogeneity parameters for each of the six subtrees containing just five taxa. In all six cases, the estimates from the real data are inside the 95% confidence interval

Table 2
Comparison of Parameter Estimates from Five-taxon Subtrees of the Real Data, with Confidence Intervals Derived via the Parametric Bootstrap Under the Null Model of Stationarity

Five-Taxon Subtree	p_{inv} (confidence interval)	c.v.[a] (confidence interval)
14	0.5445 (0.1391–0.8968)	0.0000 (0–2.0802)
32	0.5623 (0.1497–0.8435)	1.0733 (0–2.2828)
33	0.7363 (0.1167–0.9061)	0.0801 (0–3.4434)
34	0.6690 (0.1536–0.8576)	1.0041 (0–2.0350)
35	0.5861 (0.0441–0.7955)	0.5296 (0–2.3873)
36	0.6387 (0.1361–0.8142)	0.6896 (0–2.1380)

[a] Because for many of the replicates, estimates of alpha diverged toward infinity, this value was converted to the coefficient of variation (c.v. = $1/\alpha^{1/2}$).

generated under the null model of stationarity (table 2). We therefore cannot reject the null hypothesis of stationarity of rate heterogeneity parameters; sampling error is a sufficient explanation for the pattern observed in the real data (fig. 1).

Correlation of Error in Parameter Estimates

An interesting pattern emerges if the results of the taxon sampling simulations are plotted another way. In figure 3, the paired rate heterogeneity parameters (α and p_{inv}) from each subtree are kept separate; estimates from analyses with the same number of taxa are not pooled into one category as in figures 1 and 2. A strong correlation is apparent in the estimates of the two rate heterogeneity parameters, even though the true values remain constant across all subtrees. Whenever p_{inv} is underestimated, α is also underestimated. Likewise, when either parameter is overestimated, the other is also overestimated. Statistical testing of this correlation is complicated because the data points are not independent due to shared lineages in the subtrees used to estimate the parameters; however, in all pairs of estimates derived from 6 of the 40 subtrees with nonoverlapping branches (60 pairs of parameters, 6 subtrees for each of 10 replicates), the directions of the errors were identical.

This pattern of errors in the rate heterogeneity parameter estimates is easily explained by examining the behavior of the rate heterogeneity parameters in each of their respective single-distribution models. Under both of the single-distribution models (Γ alone or I alone), some sites are not expected to have experienced any substitutions. Under the Γ model, all sites are potentially variable, but some have a sufficiently slow rate of evolution (probability of substitution) to be essentially invariable; as the shape parameter (α) gets more extreme (smaller), the proportion of such sites increases. Under the I model alone, some sites are not free to vary (presumably as a result of structural/functional constraints), and the size of this class of sites is determined by p_{inv}. Therefore, in the I+Γ model, a high proportion of the sites observed to be constant could be accommodated by a large value of p_{inv}, with the gamma distribution left to account primarily for rate heterogeneity at sites observed to vary; α will then assume an artificially high value. Thus, when p_{inv} is overestimated, α is also overestimated because p_{inv} accounts for the low-rate sites of the gamma distribution. Similarly, a large proportion of sites observed to be constant could be accounted for by an underestimation of α, such that a preponderance of the truly invariable sites are assigned a very low rate (but still are potentially variable under the model), and p_{inv} will be underestimated accordingly. Thus, the observed correlation for the two rate heterogeneity parameters of the I+Γ model is attributable to difficulty in appropriately differentiating between truly invariable sites and extremely slowly evolving sites, many of which are expected (under a gamma distribution) to have remained constant in the sequences. The conflation

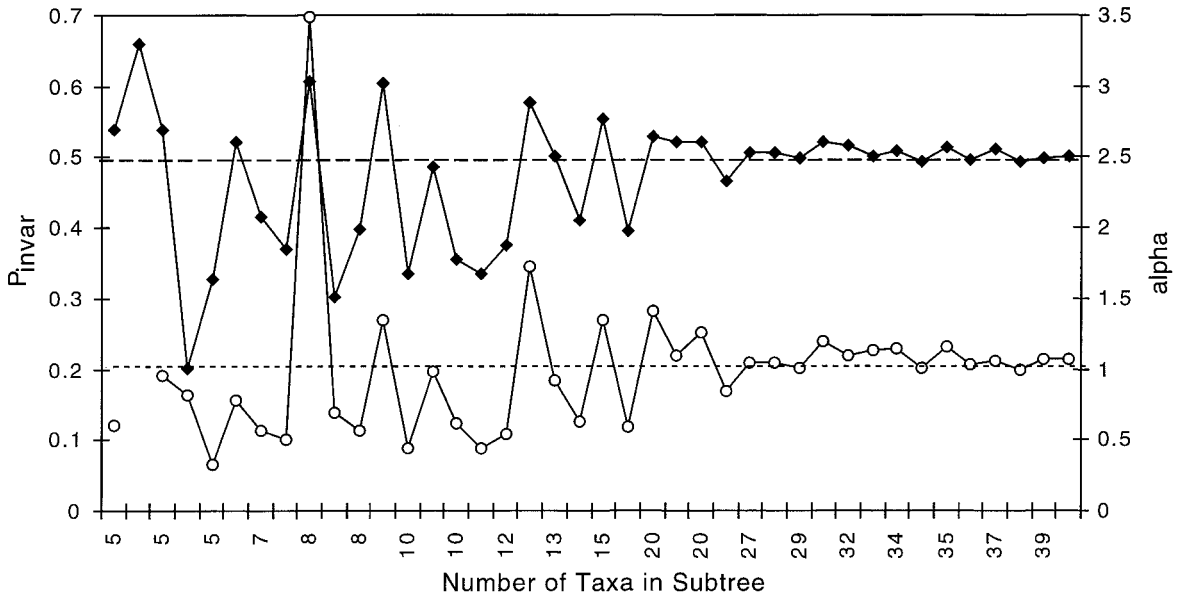

FIG. 3.—The same set of data as in figure 2, plotted to illustrate the correlation in the paired rate heterogeneity parameters. In almost all cases, the error in p_{inv} is in the same direction as the error in α.

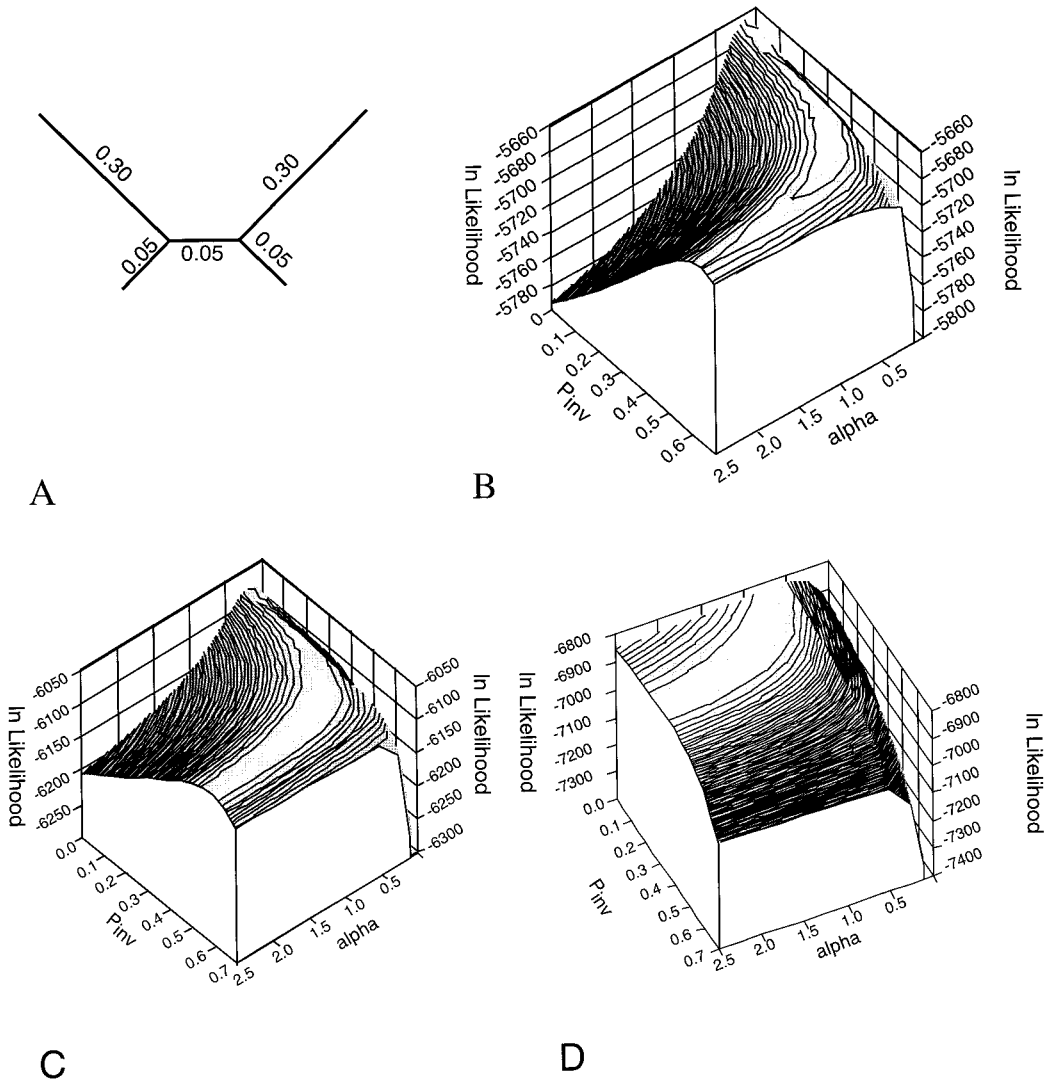

FIG. 4.—*A*, Tree used to simulate sequence evolution under the HKY+I+Γ model of evolution for examination of the likelihood surface across the rate heterogeneity parameter space. *B–D*, The likelihood surface when rate heterogeneity is (*B*) extreme ($p_{inv} = 0.5$, $\alpha = 0.5$), (*C*) strong ($p_{inv} = 0.5$, $\alpha = 1.0$), and (*D*) moderate ($p_{inv} = 0.2$, $\alpha = 1.0$).

of these parameters is difficult to overcome, but inclusion of many taxa (>20 for this set of 1,307-bp sequences) leads to reliable estimates.

This correlation of α and p_{inv} can be further illustrated by examining the likelihood surface across the rate heterogeneity parameter space. In this analysis, the four-taxon tree shown in figure 4*A* was used to generate sequences of 2,000 bp under the HKY85 model with the I+Γ model of rate heterogeneity, with the following parameters: $\pi_A = 0.2$, $\pi_C = 0.3$, $\pi_G = 0.3$, $\pi_T = 0.2$, and $\kappa = 8$; these parameter settings are similar to those used by Gu, Fu, and Li (1995). Three combinations of rate heterogeneity parameters were simulated: extreme ($p_{inv} = 0.5$; $\alpha = 0.5$; fig. 4*B*), strong ($p_{inv} = 0.5$; $\alpha = 1.0$; fig. 4*C*), and moderate ($p_{inv} = 0.2$; $\alpha = 1.0$; fig. 4*D*). Likelihood scores were calculated across the rate heterogeneity parameter space by fixing the parameters to 165 different combinations of α and p_{inv}. For each of the three rate heterogeneity conditions simulated, there is a plateau on the likelihood surface. These plots illustrate graphically that error in one parameter (p_{inv}, for example) can be compensated by a change in the other (α) such that the likelihood score changes very little. Interestingly, in figure 4*B* and *C*, there appear to be multiple peaks in the likelihood surface, and in figure 4*B*, the higher peak does not correspond to the parameter values used for the simulation. However, when 40-taxon trees are simulated using the best model for the real data, the likelihood surface is much better behaved, with a single well-defined peak (fig. 5). An examination of the nature of the likelihood surface for the rate heterogeneity parameter space under a wider variety of simulated conditions would be useful.

Discussion

Our taxon-sampling simulations (e.g., fig. 2), as well as examination of confidence intervals for five-taxon subtrees (table 2), suggest that the sensitivity of rate heterogeneity parameter estimates to taxon sampling ob-

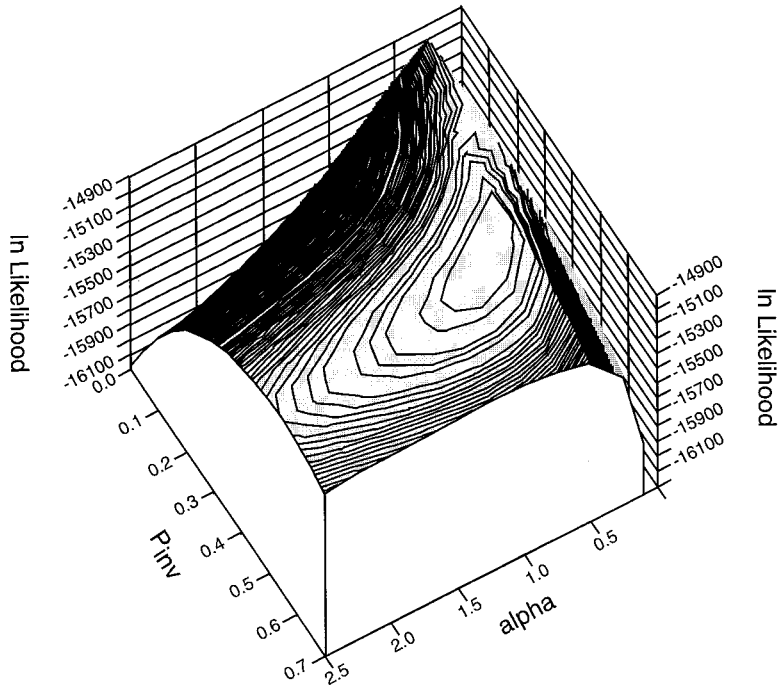

FIG. 5.—The likelihood surface across rate heterogeneity parameter space for a simulation with 40 taxa. The data were generated using the LogDet tree and model parameters estimated from the shark NADH-2 data.

served in the real data (fig. 1) can be attributed to sampling error rather than nonstationarity of those parameters. Similar patterns of sensitivity to taxon sampling were exhibited both by the real data, where nonstationarity may apply, and simulated data, where stationarity was imposed. This does not guarantee that the parameters have remained constant throughout the evolution of these shark species; it is possible that the effect of any deviation from stationarity may simply be masked by the large sampling error. Differentiating between complex stochastic models of sequence evolution, such as a stationary GTR+I+Γ versus a covarion-like model (in which rate heterogeneity parameters may change in the tree) is extremely difficult even for relatively large data sets.

Gu, Fu, and Li (1995) reported large sampling error in the estimation of α and p_{inv} (their θ) in simulations of four taxa under a restricted HKY+I+Γ model. Our results corroborate and extend theirs to a wider array of sample sizes (number of sequences) and to a more general model of sequence evolution. For the conditions we examined, this large sampling error persists when many more than four taxa are included in the analyses. Although reliable estimates of the rate heterogeneity parameters were consistently attainable only when >20 sequences were included in our analyses, there is no assurance that this represents a general threshold above which parameters will always be estimated accurately and below which estimates will always be highly variable. It is clear, however, that when a model as complex as the GTR+I+Γ model is required (as for the shark NADH-2 data), relatively large numbers of sequences will be necessary to estimate rate heterogeneity parameters accurately, at least with sequence lengths commonly obtained.

Further, we have been able to explore the nature of this error by considering the correlation in the pairs of rate heterogeneity parameter estimates. When one parameter (e.g., p_{inv}) is estimated with error, a simultaneous error occurs in the other parameter estimate (α), and together these erroneous estimates may fit the data rather well as judged by the likelihood score. This effect is manifested as plateaus, or even multiple peaks, in the likelihood surface across the rate heterogeneity parameter space (fig. 3) and results from the difficulty in distinguishing between truly invariable sites and very slowly evolving but potentially variable sites in the gamma distribution. Tourasse and Gouy (1997) recently attempted to circumvent this difficulty with a minimum-evolution (parsimony-based) approach to a mixed-distribution model of rate heterogeneity across sites. These authors proposed fitting the parsimony-inferred distribution of observed number of substitutions per site to an invariable-sites-plus-truncated-negative-binomial distribution. In order to avoid the difficulty caused by very slowly evolving, yet potentially variable, sites, Tourasse and Gouy (1997) fixed p_{inv} to the proportion of sites observed to be constant in the data and used a truncated negative-binomial distribution in which no sites are allowed to be so slowly evolving that they have a high probability of stasis. Estimates of both rate heterogeneity parameters, however, are expected to be biased for this method; the proportion of sites observed to be constant is usually an overestimate of p_{inv} (the proportion of sites that are truly invariable), whereas estimates of the gamma distribution shape parameter obtained by fit-

ting the parsimony-inferred distribution of observed substitutions per site to a negative-binomial distribution have repeatedly been shown to underestimate the strength of rate heterogeneity (overestimate α; Wakeley 1993; Sullivan, Holsinger, and Simon 1995; Yang and Kumar 1996). Because these biases are expected to decrease with an increasing number of sequences, the method of Tourasse and Gouy (1997) may provide an adequate approximate description of rate heterogeneity when the number of sequences under examination exceeds the practical limitations of full maximum-likelihood estimation. Currently, however, maximum-likelihood estimation of model parameters is feasible for around 100 sequences, and this limit is likely to expand rapidly in the near future as parallel processing strategies are incorporated into phylogenetic analyses.

Finally, we emphasize that the difficulty observed in estimating the rate heterogeneity parameters under the mixed-distribution model should not be taken as an indictment of the use of maximum likelihood for inferring evolutionary trees. Estimation of substitution model parameters and estimation of the tree topology are different, albeit related, problems (Yang 1997). Our study highlights the difficulty in obtaining accurate estimates of the true rate heterogeneity parameters, but accurate estimation of these parameters is not necessarily prerequisite to reliable estimation of phylogeny. In particular, for choosing a tree topology, it may not be important to differentiate between a high proportion of invariable sites, a gamma distribution with a small shape parameter, or a mixture of these, as long as the low-rate sites are accounted for in some way (unpublished data). These considerations, coupled with the general robustness of maximum-likelihood tree inference to model violations (e.g., Huelsenbeck 1995*b*), suggest that the problems outlined in this paper are more relevant to investigations of the effect of structural and functional constraints on sequence evolution than to the estimation of the tree topology itself.

Acknowledgments

We would like to thank Mark Hershkovitz, John Huelsenbeck, Paul Joyce, Paul Lewis, and Peter Waddell for helpful discussions and two anonymous reviewers for constructive comments. While much of this work was being conducted, J.S. was supported by the Smithsonian Institution's Molecular Evolution Fellowship.

Appendix

Taxon Labels:

1	*Carcharhinus acronotus*	11	*Carcharhinus longimanus*	21	*Carcharhinus perezi*	31	*Rhizoprionodon acutus*
2	*Carcharhinus limbatus*	12	*Carcharhinus obscurus*	22	*Galeocerdo cuvier*	32	*Triakis megalopterus*
3	*Carcharhinus sealei*	13	*Carcharhinus melanopterus*	23	*Sphyrna mokarran*	33	*Carcharhinus macloti*
4	*Carcharhinus sorrah*	14	*Carcharhinus brachyurus*	24	*Rhizoprionodon terranovae*	34	*Sphyrna tiburo*
5	*Carcharhinus leucas*	15	*Carcharhinus brevipinna*	25	*Mustelus canis*	35	*Hemigaleus microstoma*
6	*Carcharhinus amboinensis*	16	*Carcharhinus falciformis*	26	*Triaenodon obesus*	36	*Negaprion brevirostris*
7	*Carcharhinus porosus*	17	*Carcharhinus amblyrhynchos*	27	*Carcharhinus altimus*	37	*Sphyrna tudes*
8	*Prionace glauca*	18	*Carcharhinus albimarginatus*	28	*Sphyrna lewini*	38	*Carcharhinus wheeleri*
9	*Carcharhinus plumbeus*	19	*Carcharhinus signatus*	29	*Rhizoprionodon lalandei*	39	*Nasolamia velox*
10	*Carcharhinus isodon*	20	*Carcharhinus galapagensis*	30	*Sphyrna zygaena*	40	*Carcharhinus dussumieri*

Log Det Tree:

(1,(((((((((2,13),4),(((((8,16),((17,38),18)),((22,(((23,((28,(34,37)),30)),((25,32)35)),((24,29),31))),36)),((9,27),19)),((11,(12,20)),21))),(5,6)),((14,15),26)),(7,33)),(3,40)),10),39))

Subtrees:

1 (1,(((((((((2,13),4),(((((8,16),((17,38),18)),((22,(((23,((28,(34,37)),30)),35),((24,29),31))),36)),((9,27),19)),((11,(12,20)),21))),(5,6)),((14,15),26)),(7,33)),(3,40)),10),39));
2 (1,(((((((((2,13),4),(((((8,16),((17,38),18)),((22,((23,((28,(34,37)),30)),((24,29),31))),36)),((9,27),19)),((11,(12,20)),21))),(5,6)),((14,15),26)),(7,33)),(3,40)),10),39));
3 (1,(((((((((2,13),4),(((((8,16),((17,38),18)),((22,((24,29),31)),36)),((9,27),19)),((11,(12,20)),21))),(5,6)),((14,15),26)),(7,33)),(3,40)),10),39));
4 (1,(((((((((2,13),4),(((((8,16),((17,38),18)),(22,36)),((9,27),19)),((11,(12,20)),21))),(5,6)),((14,15),26)),(7,33)),(3,40)),10),39));
5 (1,(((((((((2,13),4),(((((8,16),((17,38),18)),36),((9,27),19)),((11,(12,20)),21))),(5,6)),((14,15),26)),(7,33)),(3,40)),10),39));
6 (1,(((((((((2,13),4),((((8,16),((17,38),18)),((9,27),19)),((11,(12,20)),21))),(5,6)),((14,15),26)),(7,33)),(3,40)),10),39));
7 (1,(((((((((2,13),4),(((9,27),19),((11,(12,20)),21))),(5,6)),((14,15),26)),(7,33)),(3,40)),10),39));
8 (1,(((((((((2,13),4),((11,(12,20)),21)),(5,6)),((14,15),26)),(7,33)),(3,40)),10),39));
9 (1,(((((((((2,13),4),(5,6)),((14,15),26)),(7,33)),(3,40)),10),39));
10 (1,(((3,40),(((5,6),((14,15),26)),(7,33))),10),39));
11 (1,((((3,40),((7,33),((14,15),26))),10),39));
12 (1,((((3,40),(7,33)),10),39));
13 (1,(((3,40),10),39));
14 (2,((4,(((5,6),(((14,15),26),33)),(((((8,16),((17,38),18)),((22,(((23,((28,(34,37)),30)),((25,32),35)),((24,29),31))),36)),((9,27),19)),((11,(12,20)),21)))),13));
15 (2,((4,(((5,6),((7,33),((14,15),26))),(((((8,16),((17,38),18)),((22,((23,(28,30)),((24,29),31))),36)),((9,27),19)),((11,(12,20)),21)))),13));
16 (2,((4,(((5,6),(7,33)),(((((8,16),((17,38),18)),((22,(((23,((28,(34,37)),30)),((25,32),35)),((24,29),31))),36)),((9,27),19)),((11,(12,20)),21)))),13));
17 (2,((4,(((5,6),((7,33),((14,15),26))),((((8,16),22),((9,27),19)),((11,(12,20)),21)))),13));
18 (2,((4,(((((8,((17,38),18)),((22,((24,29),31)),36)),((9,27),19)),((l1,(12,20)),21)),14)),13));
19 (23,((25,33),((28,(34,37)),30)));
20 (2,((4,((5,6),(((9,27),19),((11,(12,20)),21)))),13));
21 (8,(((((9,27),19),((22,(((23,((28,(34,37)),30)),(32,35)),((24,29),31))),36)),((17,38),18)),16));
22 (23,(((24,29),31),((28,(34,37)),30)));
23 (8,(((9,27),19),((17,38),18)),16));
24 (2,((4,(6,((11,(12,20)),21))),13));
25 (3,(((7,((13,16),(14,15))),10),40));
26 (3,(((((5,6),(((11,(12,20)),(16,24)),13)),(14,15)),7),40));
27 (8,((((((9,27),19),(20,21)),((22,((24,29),31)),36)),((17,38),18)),16));
28 (22,(((23,((28,(34,37)),30)),35),((24,29),31)));
29 (8,(((((9,27),19),(20,21)),((17,38),18)),16));
30 (1,((((3,40),((7,33),((14,15),26))),10),39));
31 (5,(6,((14,15),26)));
32 (9,((11,(12,20)),21));
33 (8,(16,((17,38),18)));
34 (22,((24,29),31),36));
35 (23,((28,(34,37)),30));
36 (1,(((((((((2,13),4),(((((8,16),((17,38),18)),((22,(((24,29),31),((25,32),35))),36)),((9,27),19)),((11,(12,20)),21))),(5,6)),((14,15),26)),(7,33)),(3,40)),10),39));
37 (1,((((3,40),((7,33),((((((8,16),((17,38),18)),((22,(((23,((28,(34,37)),30)),((25,32),35)),((24,29),31))),36)),((9,27),19)),((11,(12,20)),21)),((14,15),26)))),10),39));
38 (1,(((((((((2,13),4),((((9,27),19),((22,(((23,((28,(34,37)),30)),((25,32),35)),((24,29),31))),36)),((11,(12,20)),21))),(5,6)),((14,15),26)),(7,33)),(3,40)),10),39));
39 (2,(((((((3,40),(1O,39)),(7,33)),((14,15),26)),(5,6)),(((((8,16),((17,38),18)),((22,(((23,((28,(34,37)),30)),((25,32),35)),((24,29),31))),36)),((9,27),19)),((11,(12,20)),21))),4),13));
40 (18,(((22,(((23,((28,(34,37)),30)),((25,32),35)),((24,29),31))),36),38));

LITERATURE CITED

BRENT, R. P. 1973. Algorithms for minimization without derivatives. Prentice-Hall, Englewood Cliffs, N.J.

EDWARDS, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge, England.

EFRON, B., and R. J. TIBSHIRANI. 1993. An introduction to the bootstrap. Chapman and Hall, New York.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

FITCH, W. M., and E. MARGOLIASH. 1967. A method for estimating the number of invariant amino acid positions in a gene using cytochrome c as a model case. Biochem. Genet. **1**:65–71.

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.

GAUT, B. S., and P. O. LEWIS. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. **12**:152–162.

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

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating the human-ape split by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

HERSHKOVITZ, M. A., and L. A. LEWIS. 1996. Deep-level diagnostic value of the rDNA-ITS region. Mol. Biol. Evol. **13**:1276–1295.

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

———. 1995*b*. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol. Biol. Evol. **12**: 843–849.

HUELSENBECK, J. P., D. M. HILLIS, and R. JONES. 1996. Parametric bootstrapping in molecular phylogenetics: applications and performance. Pp. 19–45 *in* J. D. FERRARIS and S. R. PALUMBI, eds. Molecular zoology: advances, strategies, and protocols. Wiley-Liss, New York.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

LAKE, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA **91**:1455–1459.

LOCKHART, P. J., M. A. STEEL, M. D. HENDY, and D. PENNY. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11**:650–612.

POWELL, M. J. D. 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives. Comp. J. **7**:155–162.

ROGERS, J. S., and D. L. SWOFFORD. 1999. A fast method for approximating maximum likelihoods of phylogenetic trees from nucleotide sequences. Syst. Biol. (in press).

———. 1999. Multiple local maxima for likelihoods of phylogenetic trees: a simulation study. Mol. Biol. Evol. **16**: 1079–1085.

STEELE, M. 1994. The maximum likelihood point for a phylogenetic tree is not unique. Syst. Biol. **43**:560–564.

SULLIVAN, J., K. E. HOLSINGER, and C. SIMON. 1995. Among-site rate variation and phylogenetic analysis of 12S rRNA data in sigmodontine rodents. Mol. Biol. Evol. **12**:988–1001.

———. 1996. The effect of topology on estimates of among-site rate variation. J. Mol. Evol. **42**:308–312.

SULLIVAN, J., J. A. MARKERT, and C. W. KILPATRICK. 1997. Molecular systematics and phylogeography of the *Peromyscus aztecus* species group (Rodentia: Muridae) inferred using parsimony and likelihood. Syst. Biol. **46**:426–440.

SULLIVAN, J., and D. L. SWOFFORD. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mamm. Evol. **4**:77–86.

TOURASSE, N. J., and M. GOUY. 1997. Evolutionary distances between nucleotide sequences based on the distribution of substitution rates among sites as estimated by parsimony. Mol. Biol. Evol. **14**:287–298.

UZZELL, T., and K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. Science **172**: 1089–1096.

WADDELL, P., and D. PENNY. 1996. Evolutionary trees of apes and humans from DNA sequences. Pp. 53–73 *in* A. J. LOCK and C. R. PETERS, eds. Handbook of symbolic evolution. Clarendon Press, Oxford.

WAKELEY, J. 1993. Substitution rate variation among sites in hypervariable region I of human mitochondrial DNA. Mol. Biol. Evol. **36**:613–623.

YANG, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**:1396–1401.

———. 1994*a*. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39**:105–111.

———. 1994*b*. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

———. 1996*a*. Among-site rate variation and its impact on phylogenetic analyses. TREE **11**:367–372.

———. 1996*b*. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. **42**:294–307.

———. 1997. How often do wrong models produce better phylogenies? Mol. Biol. Evol. **14**:105–108.

YANG, Z., and S. KUMAR. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rate among sites. Mol. Biol. Evol. **13**: 650–659.

YANG, Z., N. GOLDMAN, and A. FRIDAY. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. **44**:384–399.

MASAMI HASEGAWA, reviewing editor

Accepted June 24, 1999