Syst. Biol. 53(2):265–277, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490423520

The Importance of Proper Model Assumption in Bayesian Phylogenetics

ALAN R. LEMMON AND EMILY C. MORIARTY

Section of Integrative Biology, University of Texas, 1 University Station C0930, Austin, Texas 78712, USA;
E-mail: alemmon@evotutor.org (A.R.L.), chorusfrog@mail.utexas.edu (E.C.M.)

Abstract.—We studied the importance of proper model assumption in the context of Bayesian phylogenetics by examining >5,000 Bayesian analyses and six nested models of nucleotide substitution. Model misspecification can strongly bias bipartition posterior probability estimates. These biases were most pronounced when rate heterogeneity was ignored. The type of bias seen at a particular bipartition appeared to be strongly influenced by the lengths of the branches surrounding that bipartition. In the Felsenstein zone, posterior probability estimates of bipartitions were biased when the assumed model was underparameterized but were unbiased when the assumed model was overparameterized. For the inverse Felsenstein zone, however, both underparameterization and overparameterization led to biased bipartition posterior probabilities, although the bias caused by overparameterization was less pronounced and disappeared with increased sequence length. Model parameter estimates were also affected by model misspecification. Underparameterization caused a bias in some parameter estimates, such as branch lengths and the gamma shape parameter, whereas overparameterization caused a decrease in the precision of some parameter estimates. We caution researchers to assure that the most appropriate model is assumed by employing both a priori model choice methods and a posteriori model adequacy tests. [Bayesian phylogenetic inference; convergence; Markov chain Monte Carlo; maximum likelihood; model choice; posterior probability.]

Model choice is becoming a critical issue as the number of available models of nucleotide evolution increases rapidly. Recent studies have shown that adequate model choice is important by demonstrating that violations of model assumptions can produce biased results (Felsenstein, 1978; Huelsenbeck and Hillis, 1993; Yang et al., 1994; Swofford et al., 2001). When the model assumed is overparameterized (too complex relative to the true underlying model), unnecessary sampling variance is introduced from estimation of extra parameters. This added variance may compromise phylogenetic accuracy (Cunningham et al., 1998). Cases in which the model assumed is underparameterized (too simple relative to the true underlying model) are especially problematic for phylogeny estimation because of the phenomenon of long-branch attraction, where the confidence in estimation of an incorrect bipartition increases as more data are included (Swofford et al., 2001). Most studies of the importance of model choice have concentrated on the four-taxon case, often comparing maximum parsimony and/or distance-based methods with maximum-likelihood methods under simple model assumptions (Felsenstein, 1978; Huelsenbeck and Hillis, 1993; Gaut and Lewis, 1995; Swofford et al., 2001).

Traditional likelihood, parsimony, and distance-based methods of phylogeny reconstruction are giving way as Bayesian approaches to phylogeny inference rapidly gain in popularity (Huelsenbeck et al., 2001). Traditional methods yield a single (best) tree, and the uncertainty of each clade is assessed through repeatability tests, such as the bootstrap. The end product of a Bayesian analysis is fundamentally different, consisting of a distribution of "best" trees with associated model parameters sampled in proportion to their posterior probabilities. Uncertainty in the phylogeny and parameter estimates is expressed in the posterior probability distribution. (For a more detailed introduction to the use of Bayesian methods in phylogenetics, see Huelsenbeck et al., 2001.)

Because Bayesian methods have only recently emerged at the forefront of phylogenetics, research concerning the proper application of these methods and the interpretation of their results is still inadequate (Huelsenbeck et al., 2002). Progress has been made with regard to the relationship between bipartition posterior probabilities and nonparametric bootstrap values, although the relative accuracy of the two measures is still being debated (Suzuki et al., 2002; Wilcox et al., 2002; Alfaro et al., 2003; Cummings et al., 2003; Douady et al., 2003; Erixon et al., 2003). Further exploration of at least three other questions is especially critical: (1) how sensitive are these analyses to prior probability assumptions, (2) what is the most appropriate way to check for convergence and stationarity of Markov chains in the context of phylogenetics, and (3) how important is proper model assumption within the Bayesian framework?

We present here an analysis that addresses the third question. We investigated the effect of model misspecification on bipartition posterior probabilities, branch-length estimates, and other substitution-model parameter estimates by analyzing >5,000 Bayesian runs under a variety of nucleotide substitution models. To explore further how bipartition posterior probabilities are affected by model misspecification, we examined two special cases in which adequate model assumption is known to be important: the Felsenstein zone and the inverse Felsenstein zone (Swofford et al., 2001). We also discuss here the importance of proper model assumption and what can be done to assure that the most appropriate model available is assumed.

METHODS

Data Set Simulation

We selected six nested models of nucleotide substitution for our analyses: JC (Jukes and Cantor, 1969), K2P (Kimura, 1980), HKY (Hasegawa et al., 1985), GTR

(Lanave et al., 1984; Tavaré, 1986; Rodríguez et al., 1990), GTR+Γ (Steel et al., 1993; Yang, 1993), and GTR+Γ+I (Gu et al., 1995; Waddell and Penny, 1996). We simulated 100 replicate data sets of 1,000 bp sequence length (Seq-Gen 1.2.5; Rambaut and Grassly, 1997) assuming each of these six substitution models and the following parameter values (as appropriate for each model): transition/transversion ratio (κ) = 2.0, $\pi_A = 0.35$, $\pi_C = 0.22$, $\pi_G = 0.18$, $\pi_T = 0.25$, $r_{CT} = 30.7$, $r_{CG} = 0.225$, $r_{AG} = 7.35$, $r_{AT} = 6.125$, $r_{AC} = 2.675$, gamma shape parameter (α) = 0.67256, and proportion of invariable sites = 0.25. With the exception of the transition/transversion ratio and the proportion of invariable sites, all of these parameter values were obtained from a mitochondrial DNA analysis used to construct a phylogeny of North American chorus frogs (*Pseudacris*; Moriarty and Cannatella, 2004). We used data sets of 1,000 bp because this length was typical of empirical data sets at the onset of this study.

The 30-taxon tree used to simulate the data sets (tree 1) was generated using DNA-Sim (program written by A.R.L.) that assumes a birth–death process (speciation rate = 10^{-4}, extinction rate = 10^{-5}). The branch lengths were assigned in the following fashion. We numbered the internal branches from 0 to 26, choosing the order of the branches randomly, and then each branch was assigned a branch length using the following equation: $f(x) = 10^{2x/26-3}$, where x is the number assigned to that branch. This method of assigning branch lengths assured that a wide range of bipartition posterior probabilities would result from our Bayesian analyses. The procedure was repeated for the 30 external branches, using the equation: $f(x) = 10^{2x/29-3}$. Tree 1 is illustrated in Figure 1.

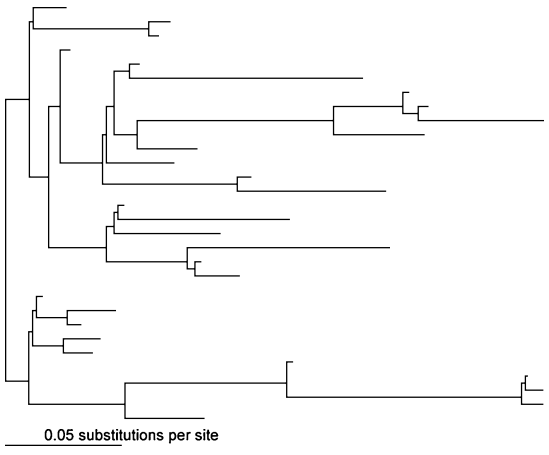

FIGURE 1. Tree 1, was used to simulate data sets for the first set of analyses. The topology for this tree was generated using a birth–death process. The branch lengths were assigned randomly.

Bayesian Analyses

To investigate the effect of model misspecification on bipartition posterior probabilities, we performed 3,600 Bayesian analyses using the program MrBayes 3.0b3 (Huelsenbeck, 2001; Huelsenbeck and Ronquist, 2001). For each of the 600 simulated data sets, we conducted six MrBayes searches, each assuming a different one of the nested models. Thus, we examined 36 model combinations: 15 in which the assumed model was underparameterized, 15 in which the assumed model was overparameterized, and 6 in which the assumed model was appropriate relative to the model used to simulate the data sets. This design is depicted in Figure 2. We compared results from runs that used the same data set but were analyzed under different substitution models. Using nested substitution models allowed us to systematically test the effect of the presence or absence of each type of parameter on the estimation of bipartition posterior probabilities, branch lengths, and other model parameters. We limited our sampling to 100 replicates because of computational constraints. To assure that results obtained from 100 replicates were reliable, we analyzed an additional 400 replicates for one model combination (GTR+Γ+I–JC). We also assessed the sensitivity of our sample design using power analyses.

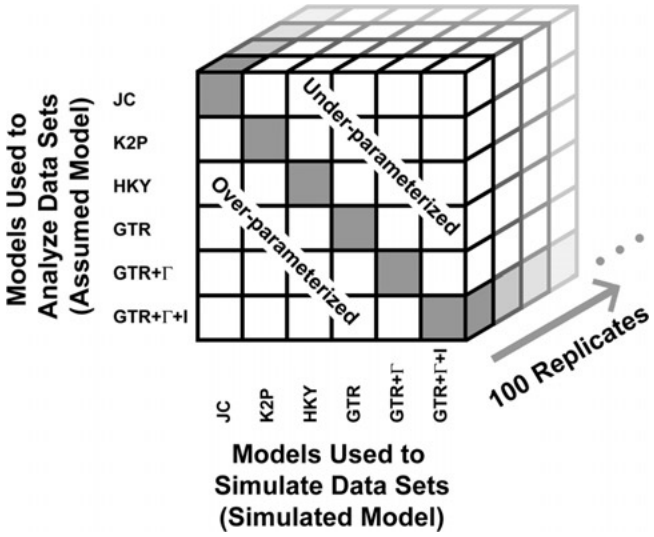

FIGURE 2. Study design of 36 model combinations. Shaded squares represent the model combinations in which the assumed model matches the simulated model. Model combinations above the diagonal contain an assumed model missing one or more parameters of the simulated model. Model combinations below the diagonal contain an assumed model including one or more parameters not present in the simulated model.

We conducted extensive preliminary analyses to determine the sample size, sample interval, and burn-in period appropriate for our data sets. The goal of these preliminary analyses was to determine the conditions that minimized the amount of variation among independent Bayesian analyses run under identical conditions. We assumed default priors for all parameters except for the GTR rate matrix. For the GTR rate matrix, a flat prior yielded incorrect substitution rate estimates and poor convergence to the true posterior distribution. This phenomenon has been studied more extensively by Zwickl and Holder (unpubl. data). An exponential prior (revmatpr = exponential(0.2)) allowed for reasonable convergence to the true posterior distribution when

the correct model was assumed (by reasonable convergence we mean that the maximum likelihood estimate of each parameter was at or very near the true value). Four Markov chains with a temperature of 0.15 assured proper mixing. Each MrBayes run spanned 500,000 generations. We sampled every 25 generations, yielding 20,000 total samples per run. Based on our preliminary tests, we chose an appropriate burn-in time of 25,000 generations (1,000 samples). Thus, each run was analyzed using 19,000 post-burn-in samples.

Convergence Testing

We employed several methods to assure that our runs had converged on the posterior distribution and that we had collected enough samples to obtain reliable results. First, we examined the stationarity of likelihood scores for all 3,600 runs performed. However, because of the large number of runs we could not examine each one independently. Instead, we visualized the likelihood curves (generation plotted on the x-axis, log likelihood plotted on the y-axis) for all 100 replicates of each model combination on the same graph, plotting only those samples for which the likelihood score was greater than that of any previous sample (data not shown). By plotting the likelihood scores in this manner, we quickly identified any runs that failed to reach stationarity within the chosen burn-in period.

Second, we examined the convergence of bipartition posterior probabilities, maximum likelihood scores, and model parameter estimates. We expect two converged runs performed on the same data set and under the same model assumptions to produce very similar bipartition posterior probability distributions, maximum likelihood scores, and model parameter estimates. Thus, a comparison of results from duplicate runs can be used to test for convergence. However, because of computational constraints we could not duplicate all 3,600 runs. Instead, we chose to concentrate on the eight model combinations in which the simulated and assumed models were either equivalent (e.g., JC–JC) or showed the greatest disparity (i.e., JC–GTR+Γ+I and GTR+Γ+I–JC). The duplicate runs were compared by observing the degree of correlation (across all 100 replicates) of bipartition posterior probabilities, maximum likelihood scores, and model parameter estimates. Checking for convergence in this way required an additional 800 Bayesian analyses.

Third, we checked the nature of the tree space to assure that 500,000 generations allowed convergence upon and proper sampling of the true posterior distribution. We repeated the analysis of five randomly chosen replicates from each of the four extreme model combinations (JC–JC, JC–GTR+Γ+I, GTR+Γ+I–JC, and GTR+Γ+I–GTR+Γ+I) but allowed these chains to run for 5,000,000 generations. We then compared the posterior distribution sampled in the shorter (500,000 generations) runs with that sampled in the longer (5,000,000 generations) runs to determine whether shorter chains were prone to entrapment in local optima. Each of the 20 pairs of posterior distributions was compared using the Mesquite (Maddison and Maddison, 2003) module Tree Set Viz (Amenta and Klinger, 2002). Tree Set Viz uses multidimensional scaling to represent the relationships among topologies (in this case, the topologies in the posterior distribution) as a scatter of points in two-dimensional space. The software arranges the points such that they group according to the distance between the trees (distance between trees was calculated using Robinson–Foulds differences; Robinson and Foulds, 1981). In addition to the visual comparisons, we compared the posterior distributions of topologies by examining the correlation of posterior probabilities of the topologies found in the shorter runs and the posterior probabilities of the topologies found in the longer runs.

Determining the Effects of Model Misspecification

We studied the effects of model misspecification on estimates of bipartition posterior probabilities, branch lengths, and other model parameters. Because we simulated the data sets, we know the true model parameter values and can compare those values with the estimates obtained through our Bayesian analysis. However, we do not have true values for the bipartition posterior probabilities (though we know which bipartitions are correct, we do not know their posterior probabilities a priori because these will depend on the data set assumed). Consequently, we compared the bipartition posterior probability estimates obtained when an incorrect model was assumed with the estimates that were obtained when the correct model was assumed. This procedure tells us the effect of model misspecification relative to the results that would have been obtained had the correct model been assumed. This comparison gives us a way to measure the bias in bipartition posterior probability estimates induced by model misspecification.

How might biased bipartition posterior probabilities affect conclusions regarding the relationships among taxa? To answer this question, we specified a rule for deciding whether a particular observed bipartition was true based on comparison of the posterior probability of that bipartition to a predetermined threshold (the threshold varied from 0.5 to 1.0). For example, a threshold of 0.5 implies that all bipartitions with a posterior probability $\geq$0.5 are accepted as true (i.e., all bipartitions in the majority-rule consensus tree are accepted). Because we know the true bipartitions, we can use the decision rule to estimate the probability of type I and type II errors for each posterior distribution observed. The probability of type I error was estimated as the proportion of true bipartitions observed that were rejected based on their posterior probability. Conversely, the probability of type II error was estimated as the proportion of false bipartitions observed that were accepted as true. We compared the probabilities of type I and type II error across the 36 model combinations to see how model misspecification might affect conclusions about the relationships among taxa.

Model misspecification may negatively affect parameter estimation by either decreasing accuracy or

decreasing precision (Cunningham et al., 1998). To assess how the accuracy of parameter estimates may be affected by model misspecification, we compared (for each parameter) the maximum likelihood estimate of the parameter obtained when the correct model was assumed with that obtained when the model was misspecified. To quantify the degree of bias for each parameter, we employed the two-tailed, paired-sample *t*-test (Zar, 1999). In this case, we are testing whether the distribution of differences (value assuming correct model minus value assuming incorrect model) is significantly different from zero. We calculated the *P* value associated with the amount of bias observed for all applicable model combinations in which the model was misspecified. To assess how the precision of parameter estimates may be affected by model misspecification, we repeated these tests using the width of the 95% credible set from the posterior distribution of the parameter as our measure of precision. For each *t*-test performed, we estimated the minimum difference in accuracy and precision that we are able to detect 99% of the time ($\beta = 0.01$), given a level of significance of 0.01 and the variance estimated from the distribution of differences (Zar, 1999).

Robustness Tests

To test for robustness of our results, we performed a second set of analyses assuming a different topology and set of model parameters. Because of time constraints, we focused on the four extreme model combinations: JC–JC, JC–GTR+Γ+I, GTR+Γ+I–JC, and GTR+Γ+I–GTR+Γ+I. The parameter values chosen were $\pi_A = 0.313735$, $\pi_C = 0.285552$, $\pi_G = 0.18302$, $\pi_T = 0.217693$, $r_{CT} = 33.79102$, $r_{CG} = 0.55726$, $r_{AG} = 11.10442$, $r_{AT} = 3.44797$, $r_{AC} = 4.16568$, gamma shape parameter (α) = 0.583564, and proportion of invariable sites = 0.454113. These values were obtained from a phylogenetic analysis of Siluriformes (D. Hillis, unpubl. data). We created a 16-taxon tree (tree 2, Fig. 3) containing structures that are difficult to recover when the assumed substitution model is inappropriate, the structures found in the Felsenstein and inverse Felsenstein zones (Swofford et al., 2001).

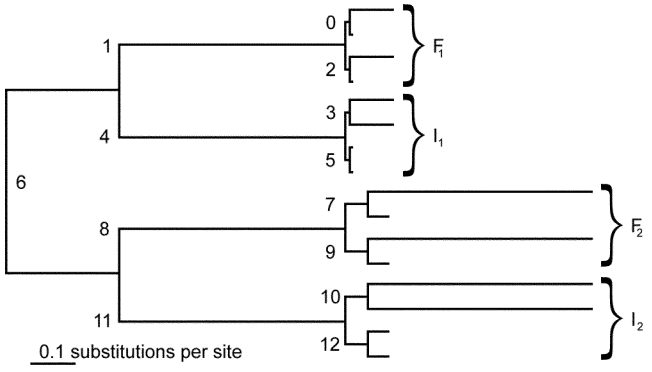

FIGURE 3. Tree 2 was used to simulate data sets for the second set of analyses. This tree contains two Felsenstein structures (F_1 and F_2) and two inverse Felsenstein structures (I_1 and I_2). Internal nodes are labeled for reference.

Tree 2 contained two Felsenstein structures (containing two long branches separated by a third, much shorter branch) and two inverse Felsenstein structures (containing a pair of long branches that are adjacent to a pair of much shorter branches). We included two structures of each type to provide some variation in the difficulty of the problem, although we could not perform an extensive analysis. The branches separating the four structures were fairly long (0.5 substitutions per site), allowing us to avoid confounding effects of interactions among two or more structures. We assumed the same prior distributions, sample size, sample interval, and burn-in times. The 800 Bayesian analyses (400 unique runs, each duplicated) conducted under these conditions were analyzed in a fashion similar to that for the previous analysis.

We constructed tree 2 using two Felsenstein structures and two inverse Felsenstein structures with the hopes of determining how the properties of a particular bipartition affect the type of bias produced by model misspecification. We can determine the effect of misspecification for each bipartition by comparing the posterior probability obtained for that bipartition when the model is misspecified with the posterior probability obtained when the model is correctly specified. The proper test for this type of comparison, given that the distributions of bipartition posterior probabilities across the 100 replicates are neither normal nor homoschedastic, is the nonparametric sign test (Zar, 1999). We employed this test for each of the bipartitions in tree 2 that are part of either a Felsenstein structure or an inverse Felsenstein structure.

RESULTS

Bipartition Posterior Probabilities

Model misspecification can strongly bias bipartition posterior probability estimates (Fig. 4). Although overparameterization had no noticeable effect on bipartition posterior probability estimates, underparameterization produced a strong bias. The bias observed for a particular bipartition depends on how well supported that bipartition is when the correct model is assumed: well-supported bipartitions tend to be overestimated, whereas poorly supported bipartitions tend to be underestimated. Although this is the general trend, the effect of model misspecification on a particular bipartition is likely to be affected by the length of the branch at that bipartition, the length of the branches surrounding that bipartition, and the data set used to infer the phylogeny. Results from an additional 400 replicates for the model combination GTR+Γ+I–JC (see Fig. 5) suggest that increasing sampling efforts would not have affected our qualitative results regarding the effect of model misspecification on bipartition posterior probability estimates.

The largest bias in bipartition posterior probabilities was seen when the assumed model failed to incorporate rate variation across sites. This bias was especially pronounced when gamma distributed rate heterogeneity was neglected. We also observed that failing to account for unequal rates of base substitution (i.e., the transition

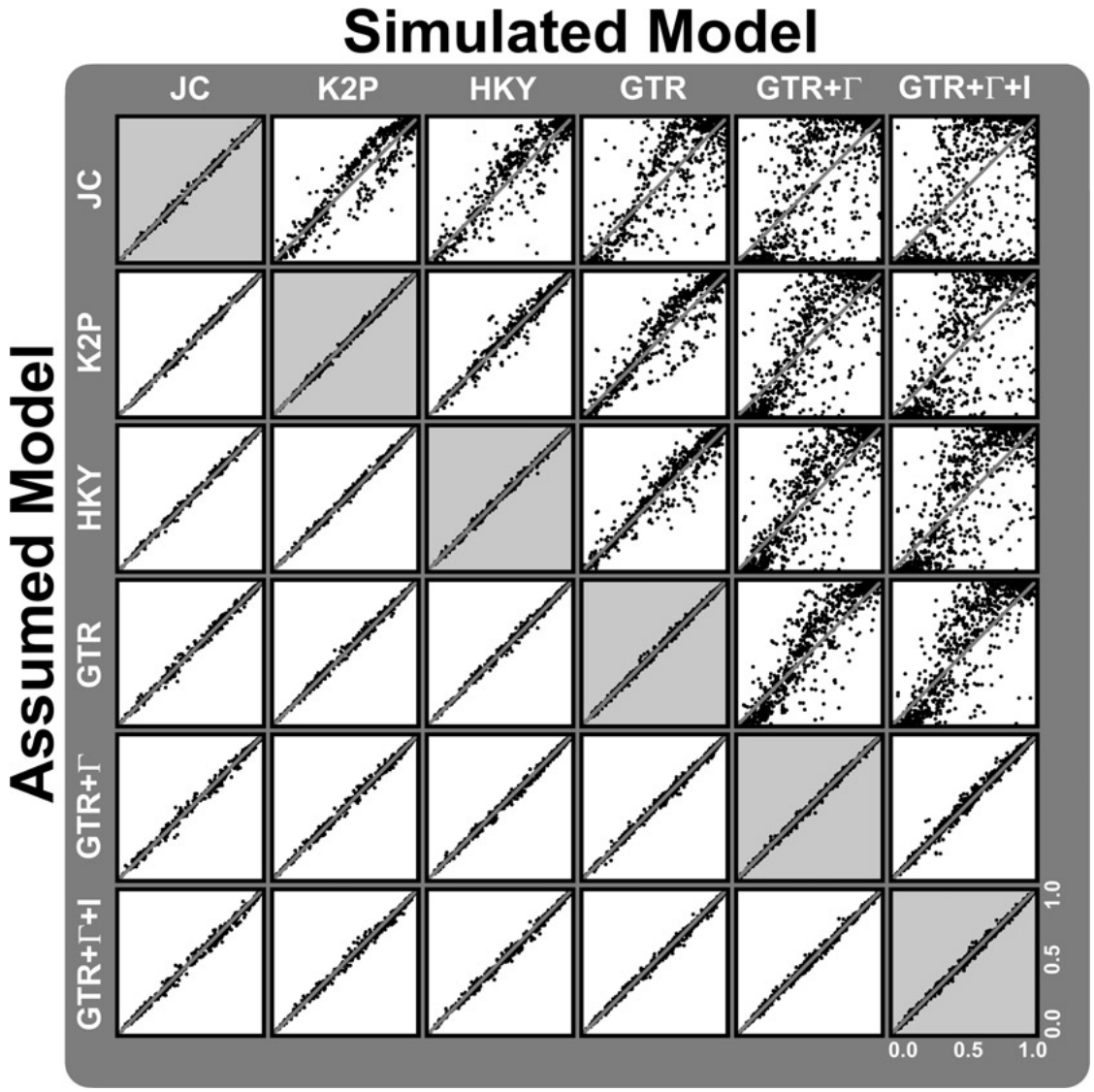

FIGURE 4. The effect of model misspecification on bipartition posterior probability estimates. The six graphs on the shaded diagonal demonstrate the convergence of the bipartition posterior probabilities for 100 pairs of independent runs when the correct model is assumed. Each of the 30 unshaded graphs compare the bipartition posterior probabilities obtained when the correct model was assumed (plotted on the *x*-axis) with those obtained when an incorrect model was assumed (plotted on the *y*-axis). The posterior probabilities of all 27 true bipartitions are plotted on the same graph for all 100 replicates, yielding 2,700 points per graph. To determine the effect of model misspecification involving a single type of parameter, compare a graph on the shaded diagonal with a graph either directly above (underparameterized) or directly below (overparameterized). Only the bipartitions found in the true tree (tree 1) are represented.

bias or the GTR rate matrix) led to slightly biased bipartition posterior probability estimates. However, inappropriately assuming equal base frequencies had very little effect on bipartition posterior probability estimates under the conditions tested.

Bias caused by underparameterization resulted in an increased incidence of type II error (Fig. 6), i.e., assuming an underparameterized model led to the acceptance of a greater number of false bipartitions. This pattern was consistent across all threshold values tested. The opposite trend occurred for type I error; assuming an underparameterized model resulted in the rejection of fewer true bipartitions. As one might expect, increasing the threshold resulted in an increase in type I error and a decrease in type II error. Overparameterization had very little effect on the probability of either type of error.

Branch Lengths and Other Model Parameters

Branch lengths were also affected by model misspecification. Model underparameterization led to underestimated branch lengths, especially for long branches (Fig. 7). Failing to account for rate heterogeneity had the largest effect on branch-length estimates, although neglecting to include other parameters also produced slightly underestimated branch lengths.

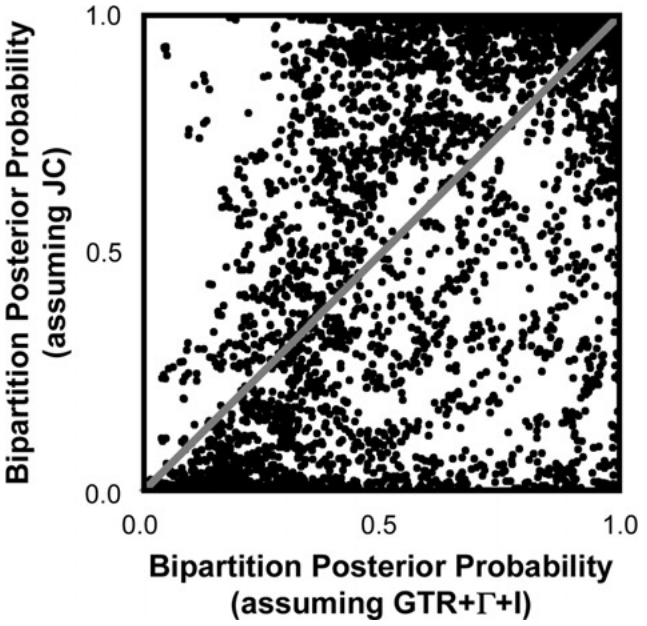

FIGURE 5. Graph of 400 additional replicates for the model combination GTR+Γ+I–JC. Compare this graph with the graph in the upper right corner of Figure 4.

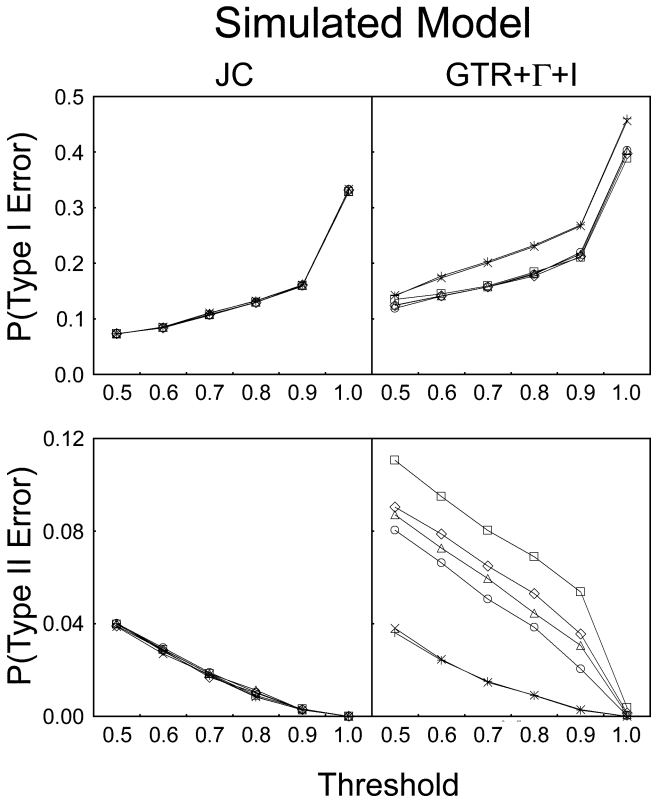

FIGURE 6. The effect of model misspecification on type I and type II error rates for hypothesis tests using bipartition posterior probabilities. The threshold is the posterior probability below which a particular bipartition is rejected. Type I error was calculated as the proportion of true bipartitions observed that were rejected based on their posterior probability. Conversely, type II error was calculated as the proportion of false bipartitions observed that were accepted as true. The assumed models are JC (□), K2P (◇), HKY (△), GTR (○), GTR+Γ (×), and GTR+Γ+I (+). Each point represents the average across 100 replicates for the model combination depicted.

Overparameterization produced little if any bias in branch-length estimates.

In many cases, parameter estimates were biased when the model assumed was underparameterized (Fig. 8). However, estimation of rate matrix parameters seems to be fairly robust to model misspecification, at least under the conditions tested. Nucleotide base frequencies were only biased when the GTR rate matrix was inappropriately neglected. Estimates of the gamma shape parameter appeared to be biased when the proportion of invariable sites was inappropriately included or ignored. The gamma shape parameter was also biased when the simulated model assumed homogeneity of rates across sites (Fig. 9). This result is expected because the true value of α in these cases is infinity.

We observed decreased precision of some parameter estimates when the assumed model was overparameterized (Fig. 8). Tree length (TL), base frequencies, and the gamma shape parameter showed a strong decrease in precision under some conditions. Estimates of rate matrix parameters and transition bias appeared to be more robust to changes in precision with model overparameterization. Not surprisingly, underparameterization tended to result in an increase in the precision of most parameter estimates.

Our model design allowed us to detect small changes in accuracy and precision for estimates of TL, κ, and base frequencies and moderate changes for estimates of rate matrix parameters and α. The ranges in the minimum detectable difference estimates for our accuracy tests are as follows: TL, 0.015–0.030; κ, 0.096–0.104; π_A, 0.005–0.006; π_C, 0.004–0.007; π_G, 0.004–0.005; π_T, 0.004–0.007; r_{CT}, 2.969–3.466; r_{CG}, 0.072–0.099; r_{AG}, 0.799–0.886; r_{AT}, 0.605–0.760; r_{AC}, 0.304–0.351; α, 0.126–0.169. The ranges in the minimum detectable difference estimates for our precision tests are as follows: TL, 0.002–0.011; κ, 0.007–0.009; π_A, 4×10^{-4}–0.001; π_C, 4×10^{-4}–0.001; π_G, 4×10^{-4}–0.001; π_T, 4×10^{-4}–0.001; r_{CT}, 2.229–2.666; r_{CG}, 0.024–0.052; r_{AG}, 0.499–0.559; r_{AT}, 0.406–0.480; r_{AC}, 0.177–0.229; α, 0.152–0.174. These results suggest, for example, that we would be able to detect a difference of $\leq$0.006 between the maximum likelihood estimate of π_A obtained when the model was correctly assumed and the estimate obtained when the model was misspecified. Given that the estimate for this parameter was always >0.268 (for the applicable model combinations), we would have a 99% chance of detecting a change on the order of 2.2% in the maximum likelihood estimate of π_A. Likewise, we would be able to detect a difference of 0.001 between the width of the 95% credible set of π_A obtained when the correct model was assumed and the width obtained when the model was misspecified. Given that the estimate for this parameter was always >0.036 (for the applicable model combinations), we would have a 99% chance

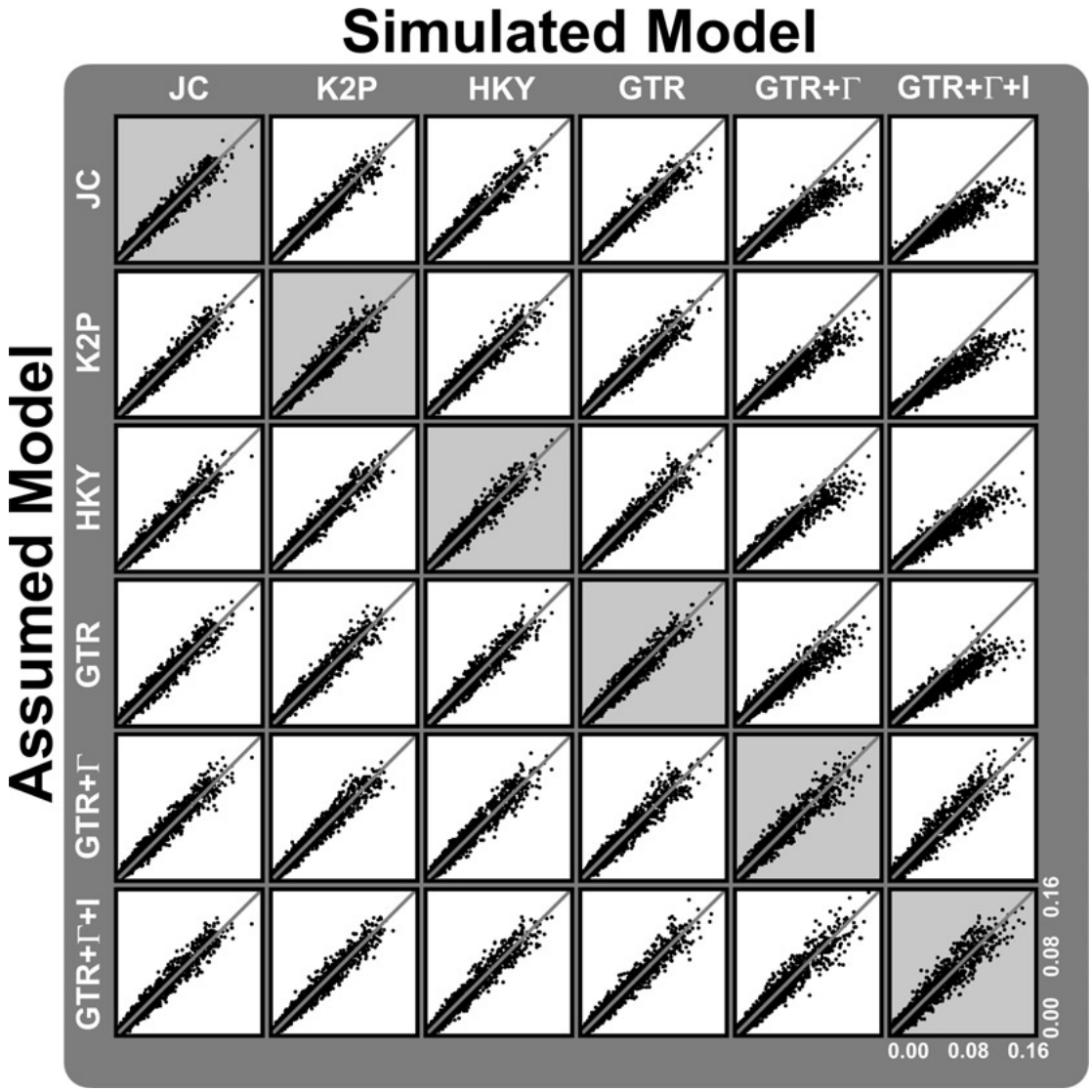

FIGURE 7. The effect of model misspecification on branch-length estimates. The format for this figure is the same as that for Figure 4, except that the values plotted are the maximum likelihood estimates of the branch lengths (when the maximum likelihood tree did not contain a particular true internal branch, no value was plotted). Only internal branches found in the true tree are represented here. Plots of external branches demonstrate very similar results.

of detecting a change on the order of 2.8% in the width of the 95% credible set.

Convergence

We observed adequate convergence of bipartition posterior probabilities (Fig. 4) and model parameter estimates (data not shown). When observing likelihood burn-in plots, we found that all runs reached stationarity before 25,000 generations, the chosen burn-in time. Good convergence of bipartition posterior probabilities can be observed in correlation plots of the duplicate runs (see the shaded diagonal of Fig. 4). These plots demonstrate the congruence of bipartition posterior probability distributions between pairs of independent Bayesian analyses (model combinations not shown, JC–GTR+Γ+I and GTR+Γ+I–JC, demonstrated a pattern very similar to that seen in the diagonal of Fig. 4). Posterior distributions of the substitution model parameters were congruent with the true parameter values (denoted by arrows in Fig. 9) when the correct model was assumed. Parameter estimates were also very similar between independent runs under the same conditions (data not shown). Moreover, running the Markov chains for 5,000,000 instead of 500,000 generations did not substantially change the resulting sample taken from the posterior distribution of topologies; the posterior distributions of the shorter and longer runs were congruent in all 20 visual comparisons made using Tree Set Viz (data not shown). We also observed a strong correlation of topology posterior

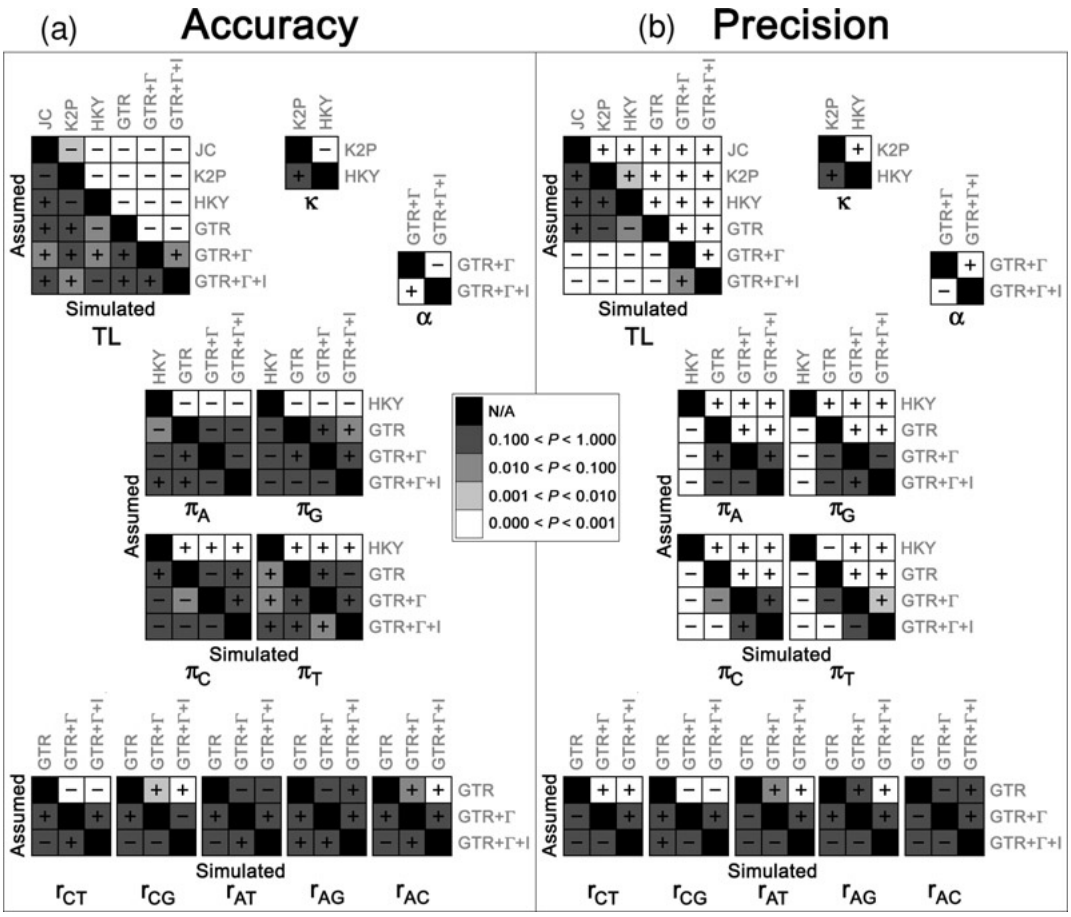

FIGURE 8. The effect of model misspecification on the accuracy and precision of parameter estimates. TL = tree length. (a) The maximum likelihood estimate of each parameter obtained assuming the correct model was compared with the estimate obtained assuming an incorrect model. Model misspecification produced either a positive bias (+) or a negative bias (−). (b) Width of the 95% credible set of each parameter obtained assuming the correct model was compared with the width obtained assuming an incorrect model. Model misspecification increased precision of the parameter estimate (+) or decreased precision (−). The boxes are shaded according to the P value obtained from a two-tailed, paired-sample t-test across 100 replicates.

probabilities between long and short runs for all four model combinations tested (data not shown).

Robustness Tests

The results of the second set of analyses, which investigated Felsenstein and inverse Felsenstein structures, agreed with those of the first set, although the bias was even more pronounced (Fig. 10). The one exception is that in the second set of analyses we were able to detect a bias in bipartition posterior probabilities caused by model overparameterization under some conditions. However, this bias was much less pronounced than the bias caused by underparameterization. The branch lengths were also more strongly biased by underparameterization under the second set of conditions, although this effect could be due to the fact that branches in tree 2 (Fig. 3) were longer than those in tree 1 (Fig. 1).

We were surprised to observe a slight bias in bipartition posterior probabilities when the assumed model was overparameterized (Fig. 10a). Other authors have found similar patterns when using simulated data sets of 1,000 nucleotides but found that the bias disappeared with increased sequence length (Sullivan and Swofford, 2001; Swofford et al., 2001). To determine whether the bias we observed was also attributable to sequence length, we constructed data sets of length 5,000, 10,000, and 50,000 nucleotides by successively concatenating randomly chosen data sets (without replacement) from the pool of 100 replicate data sets used in the robustness tests. When these data sets were analyzed in the same fashion as were those containing 1,000 nucleotides, we found that increasing sequence length corrected the slight bias seen in the case of overparameterization but amplified the already large bias seen in the case of underparameterization (data not shown).

After examining more carefully how model misspecification affected the posterior probabilities of each of the bipartitions, we found that posterior probabilities of bipartitions found in Felsenstein structures were biased by underparameterization but not overparameterization. For three of the four bipartitions found in Felsenstein

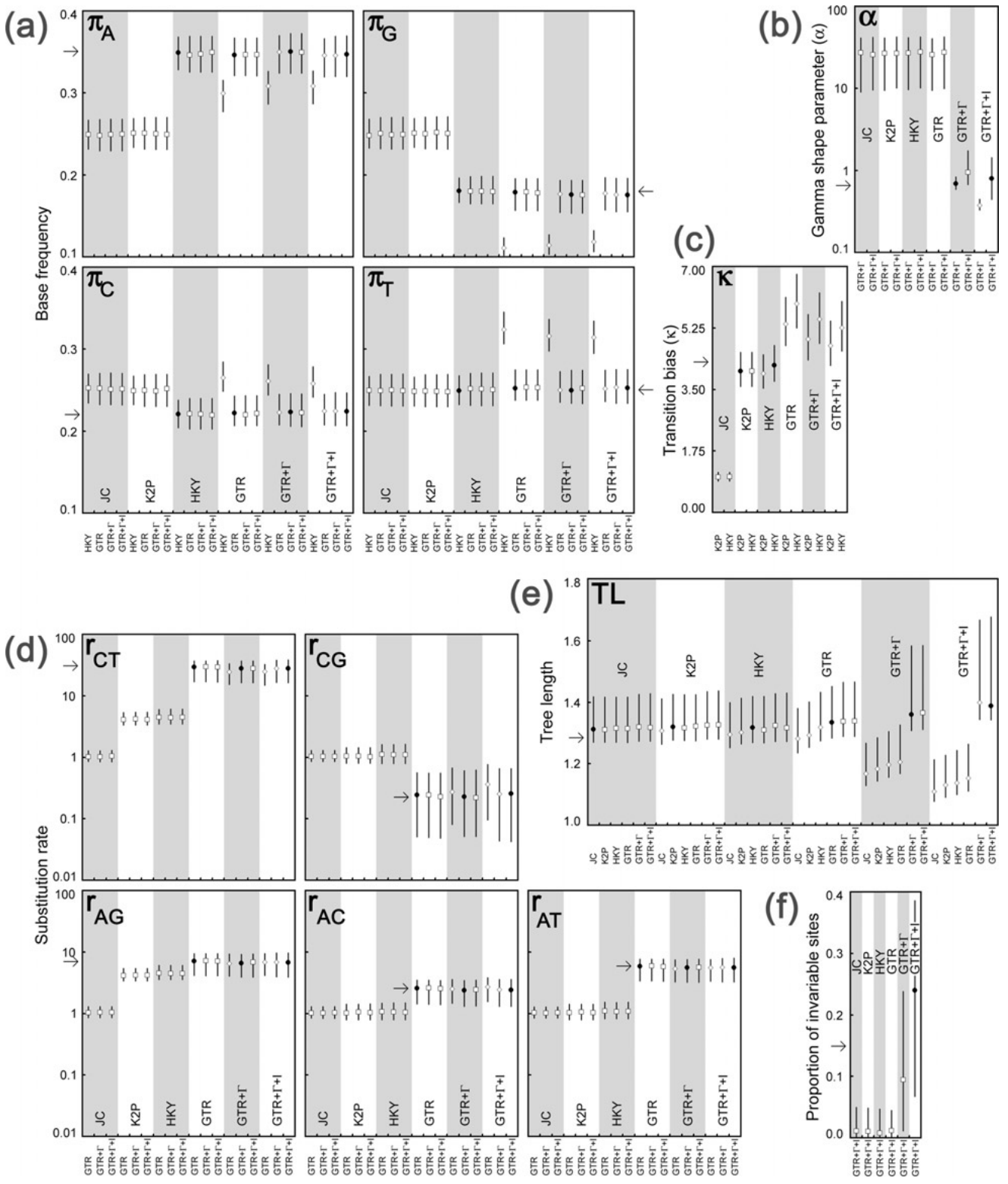

FIGURE 9. The effect of model misspecification on estimates of base frequencies (a), the gamma shape parameter α (b), the transition bias κ (c), substitution rates (d), tree length (TL) (e), and the proportion of invariable sites (f). The assumed model is labeled on the x-axis, and the simulated model is labeled in the alternating panels; each panel groups the model combinations that share the same simulated model. The maximum likelihood estimate (averaged across 100 replicates) is plotted for model combinations in which the assumed model is correct (●), overparameterized (□), or underparameterized (◇). Vertical bars represent the range of the 95% credible set, also averaged across the 100 replicates. The effect of model misspecification can be observed by comparing points represented by solid circles with other points within the same panel. Arrows indicate true parameter values (but are only pertinent to those model combinations where the simulated model includes the parameter of interest).

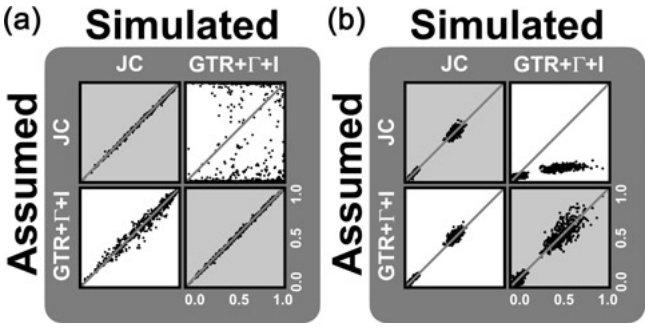

FIGURE 10. The effect of model misspecification on bipartition posterior probability (a) and branch-length estimates (b) for the second set analyses, which focus on Felsenstein and inverse Felsenstein structures. The formats are the same as those for Figures 4 and 7, respectively. Note the extreme effect of model underparameterization on bipartition posterior probabilities and branch-length estimates.

structures, a significant number of replicates showed a decrease in the bipartition posterior probability when the assumed model was underparameterized (one-tailed paired sign test, $\alpha = 0.05$; node 0: $P = 1.84 \times 10^{-1}$, node 2: $P = 3.32 \times 10^{-3}$, node 7: $P = 9.05 \times 10^{-8}$, node 9: $P = 1.53 \times 10^{-17}$). We attribute the one exception to type II error. However, when the assumed model was overparameterized no significant bias was observed for any of the four bipartitions (node 0: $P = 3.09 \times 10^{-1}$, node 2: $P = 7.95 \times 10^{-2}$, node 7: $P = 5.00 \times 10^{-1}$, node 9: $P = 4.19 \times 10^{-1}$).

For bipartitions found in inverse Felsenstein structures, we found that both overparameterization and underparameterization produced a bias in bipartition posterior probability estimates. The direction of the bias in each case depended upon whether the particular bipartition was adjacent to two long tips or two short tips. When the assumed model was underparameterized, a significant number of replicates showed a decrease in the posterior probabilities of bipartitions closest to two long tips (node 3: $P = 1.32 \times 10^{-25}$, node 10: $P = 9.05 \times 10^{-8}$), and a significant number of replicates showed an increase in the posterior probabilities of bipartitions closest to two short tips (node 5: $P = 7.97 \times 10^{-29}$, node 12: $P = 7.89 \times 10^{-31}$). Overparameterization produced the opposite bias for all nodes in the inverse Felsenstein structures: a significant number of replicates showed an increase in the posterior probabilities of bipartitions closest to two long tips (node 3: $P = 4.43 \times 10^{-2}$, node 10: $P = 9.16 \times 10^{-5}$), and a significant number of replicates showed a decrease in the posterior probabilities of bipartitions closest to two short tips (node 5: $P = 1.20 \times 10^{-3}$, node 12: $P = 7.81 \times 10^{-27}$).

The sign test does not tell us the magnitude of the bias, only the direction of the bias. However, Figure 10a clearly illustrates that the bias in bipartition posterior probabilities due to underparameterization is much more pronounced than that due to overparameterization. The bipartition posterior probabilities for nodes 1, 4, 6, 8, and 11 were always at or very near 1.0 for all 100 replicates and all model combinations; thus, our results were not influenced by interactions among two or more structures.

DISCUSSION

Model underparameterization can strongly bias estimates of bipartition posterior probabilities, branch lengths, and other model parameters. The bias is especially severe when rate heterogeneity is neglected. This result is not surprising; previous researchers have demonstrated that ignoring rate heterogeneity among sites can bias topology estimation (Kuhner and Felsenstein, 1994; Yang et al., 1994; Sullivan et al., 1995; Lockhart et al., 1996) and can lead to underestimation of branch lengths (Golding, 1983; Yang et al., 1994). Our results also agree with those of previous studies of model misspecification in that the bias seen in branch-length estimates increases disproportionately as branch length increases (Golding, 1983).

Results from our analyses of type I and type II errors suggest that the best approach to assuring accurate and informative phylogenies is to employ a sufficiently complex model and to accept bipartitions as true when their posterior probability are moderately high ($0.7 \leq$ decision threshold ≤ 0.9). Based on the results of this study, there appears to be little advantage to requiring posterior probabilities to be very near 1 when an adequate model is available. When an adequate model is not available, however, the conservative approach would be to accept bipartitions as true only when they have very high posterior probabilities (e.g., >0.9). These conclusions are based on results obtained under the particular set of conditions we examined here. Clearly, more research investigating the factors affecting error in phylogeny estimates is needed.

Model overparameterization also carries a cost: inclusion of unnecessary parameters can lead to decreased precision in estimates of branch lengths and other model parameters, as suggested by Cunningham et al. (1998). We have also seen in the case of the inverse Felsenstein zone that model overparameterization can sometimes lead to slightly biased estimates of bipartition posterior probabilities, although this bias is expected to decrease with increased sequence length. There are two additional negative consequences of model overparameterization that we have not investigated here: (1) computation time is likely to increase rapidly with the complexity of the model assumed, especially when data sets are large (Lemmon and Milinkovitch, 2002), and (2) overparameterization may affect the convergence of Markov chains (see Huelsenbeck et al., 2002, for a discussion of convergence).

If appropriate model assumption is so important, how are we to identify an appropriate model? Two steps should be taken to assure that a proper model is assumed. First, before a Bayesian analysis is performed, one should identify the available model that best fits the data set. Several methods facilitate this choice, including the likelihood-ratio test (Goldman, 1993), the

Akaike information criterion (Akaike, 1974), and the Bayesian information criterion (Schwarz, 1974). Posada and Crandall (2001) compared the performance of these methods in detail and found that the likelihood-ratio test was more accurate than the Akaike and Bayesian information methods under some conditions. None of these methods tell us whether or not a particular model adequately describes a particular data set; they are useful only for choosing the best model among a set that may or may not contain an adequate model. There is some evidence that when all of the available models are inadequate, the hierarchical likelihood-ratio test performs poorly relative to other model selection methods (Minin et al., 2003). Another disadvantage of the commonly used likelihood-ratio test is that it is appropriate only for choosing among nested models. When unnested models are being compared, an alternative method should be used, such as parametric bootstrapping (Goldman, 1993) or the recently developed decision theory methods (Minin et al., in 2003). Both of these methods also can be used to test for absolute goodness of fit of the model to the data set.

Second, following a Bayesian analysis, one should perform model-adequacy tests to assure that the model assumed in the analysis adequately explains the observed data. Model adequacy can be tested, for example, using posterior predictive distributions, as described by Bollback (2002). With this method, the posterior probability distribution from a Bayesian analysis is used to simulate numerous data sets (the posterior predictive distribution). The simulated data sets are then compared with the observed data (the one used in the Bayesian analysis). When the assumed model is adequate, the properties of the simulated data sets (e.g., the distribution of site patterns) are congruent with the properties of the observed data. However, when the assumed model is inadequate the properties of the simulated data sets will not be congruent with those of the observed data. In this case, the researcher should be very cautious when interpreting the results of a Bayesian analysis and perhaps should continue to search for a more appropriate model with which to reanalyze the data set.

Numerous models have been developed in an attempt to account for one or another of the complexities of sequence evolution, including temporal variation in base frequencies (Lockhart et al., 1994; Galtier and Gouy, 1998), temporal variation in rates of evolution (Sanderson, 1997; Thorne et al., 1998; Huelsenbeck et al., 2000; Kishino et al., 2001), base-pairing interactions of RNA (Muse, 1995; Tillier and Collins, 1998), correlation between rates of adjacent sites (Felsenstein and Churchill, 1996), different rates of synonymous and nonsynonymous substitutions (Goldman and Yang, 1994; Muse and Gaut, 1994), effects of selection on protein-coding regions (Halpern and Bruno, 1998; Nielsen and Yang, 1998), and insertions or deletions (McGuire et al., 2001). However, consideration of these models in phylogenetic studies has yet to become common practice. Four reasons likely contribute to the lack of use of these more complex models: (1) most of these models are not implemented in commonly used phylogenetic analysis packages (although recent versions of MrBayes have improved in this respect), (2) these models typically require much greater computational effort, (3) commonly used model selection methods (such as hierarchical likelihood-ratio tests) are restricted to nested models, and (4) justification for the use of more complex models is still insufficient.

Given the well-demonstrated cost to model misspecification and the general reluctance of systematists to consider the use of more complex models, how should we proceed? First, the adequacy of available models needs to be assessed on a broad scale using real data sets. Although the intuition of many systematists suggests that our models are inadequate, no large-scale test of model adequacy has been performed. If we discover that the current set of available models is inadequate with respect to most real data sets, then more research should be conducted to identify the properties of sequence evolution that are inappropriately being ignored, and models should be developed to account for those complexities. Second, we should develop computationally feasible and broadly applicable (i.e., not restricted to nested models) methods of model choice and subsequently test the absolute and relative performance of these methods. The absolute and relative performances of model choice methods are especially important because different methods are likely to disagree on which model is most appropriate. For example, the hierarchical likelihood-ratio test and the Akaike information criterion for model choice chose the same model approximately 25% of the time when several hundred empirical data sets were analyzed (A. R. Lemmon, unpubl. data). Third, enough models should be incorporated into phylogenetic analysis packages to assure that adequate models are available for most real data sets.

The goals of this study were to determine the effects of model misspecification on the estimation of phylogeny and substitution model parameters in the context of Bayesian phylogenetics. The results of this study are congruent with those from previous studies of model choice outside of the Bayesian context (Golding, 1983; Kuhner and Felsenstein, 1994; Yang et al., 1994; Sullivan et al., 1995; Lockhart et al., 1996) and therefore underscore the importance of proper model assumption. Given the bias that may result from underparameterization and the imprecision that may result from overparameterization, we strongly caution researchers to refrain from choosing models haphazardly (e.g., by assuming the most complex model that is computationally feasible or by assuming the model that happens to be the default in their favorite phylogenetic inference package). Careful consideration of the caveats resulting from studies of the importance of proper model choice will enable systematists to have greater confidence in their choice of models and estimates of phylogeny.

ACKNOWLEDGMENTS

This research was supported by NSF Graduate Research Fellowships to both authors. We thank the participants of the IGERT program at

the University of Texas–Austin for many useful discussions and for use of the phylocluster. We are very grateful to David Hillis, Derrick Zwickl, Mark Kirkpatrick, Michel Milinkovitch, Jack Sullivan, and two anonymous reviewers for their comments on an earlier version of this manuscript.

REFERENCES

Akaike, H. 1974. A new look at statistical model identification. IEEE Trans. Automatic. Control. 19:716–723.

Alfaro, M. E., S. Zoller, and F. Lutzoni. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20:255–266.

Amenta, N., and J. Klinger. 2002. Case Study: Visualizing sets of evolutionary trees. Pages 71–74 *in* IEEE symposium on Information Visualization. Available at http://csdl.computer.org/comp/proceedings/infovis/2002/1751/00/175/toc.htm.

Bollback, J. P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171–1180.

Cummings, M. P., S. A. Handley, D. S. Myers, D. L. Reed, A. Rokas, and K. Winka. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52:477–487.

Cunningham, C. W., H. Zhu, and D. M. Hillis. 1998. Best-fit maximum-likelihood models for phylogenetic inference: Empirical tests with known phylogenies. Evolution 52:978–987.

Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. P. Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20:248–254.

Erixon, P., B. Svennblad, T. Britton, and B. Oxelman. 2003. The reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst. Biol. 52:665–673.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

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.

Galtier, N., and M. Gouy. 1998. Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15:871–879.

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.

Golding, G. B. 1983. Estimates of DNA and protein sequence divergence: An examination of some assumptions. Mol. Biol. Evol. 1:125–142.

Goldman, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

Goldman, N., and Z. H. Yang. 1994. Codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11:725–736.

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.

Halpern, A. L., and W. J. Bruno. 1998. Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. Mol. Biol. Evol. 15:910–917.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Huelsenbeck, J. P. 2001. MrBayes 3.0b3: Bayesian inference of phylogeny, Distributed by the author. Department of Biology, Univ. Rochester, Rochester, New York.

Huelsenbeck, J. P., and D. M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51:673–688.

Huelsenbeck, J. P., B. Larget, and D. Swofford. 2000. A compound Poisson process for relaxing the molecular clock. Genetics 154:1879–1892.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

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

Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol. Biol. Evol. 18:352–361.

Kuhner, M. K., and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.

Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20:86–93.

Lemmon, A. R., and M. C. Milinkovitch. 2002. The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. Proc. Natl. Acad. Sci. USA 99:10516–10521.

Lockhart, P. J., A. W. D. Larkum, M. A. Steel, P. J. Waddell, and D. Penny. 1996. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA 93:1930–1934.

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. 11:605–612.

Maddison, W. P., and D. R. Maddison. 2003. Mesquite: A modular system for evolutionary analysis, version 0.994. Available at http://mesquiteproject.org.

McGuire, G., M. C. Denham, and D. J. Balding. 2001. Models of sequence evolution for DNA sequences containing gaps. Mol. Biol. Evol. 18:481–490.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:674–683.

Moriarty, E. C., and D. C. Cannatella. 2004. Phylogenetic relationships of North American chorus frogs (*Pseudacris*). Mol. Phylogenet Evol. 30:409–442.

Muse, S. V. 1995. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. Genetics 139:1429–1439.

Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11:715–724.

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.

Posada, D., and K. A. Crandall. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50:580-601.

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.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Rodríguez, F., J. L. Oliver, A. Marín, and J. R. Medina. 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. 142:485–501.

Sanderson, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14:1218-1231.

Schwarz, G. 1974. Estimating the dimension of a model. Ann. Stat. 6:461–464.

Steel, M., P. J. L. Székely, P. L. Erdös, and P. J. Waddell. 1993. A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. N. Z. J. Bot. 31:289–296.

Sullivan, J., K. E. Holsinger, and C. Simon. 1995. Among-site rate variation and phylogenetic analysis of 12S rRNA in sigmodontine rodents. Mol. Biol. Evol. 12:988–1001.

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.

Suzuki, Y., G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA 99:16138–16143.

Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525–539.

Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.

Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.

Tillier, E. R. M., and R. A. Collins. 1998. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. Genetics 148:1993–2002.

Waddell, P., and D. Penny. 1996. Evolutionary trees of apes and humans from DNA sequences. Pages 53–73 *in* Handbook of symbolic evolution (A. J. Lock and C. R. Peters, eds.). Clarendon Press, Oxford, U.K.

Wilcox, T. P., D. J. Zwickl, T. A. Heath, and D. M. Hillis. 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Mol. Phylogenet. Evol. 25:361–371.

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

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.

Zar, J. H. 1999. Biostatistical analysis. Prentice-Hall, Englewood Cliffs, New Jersey.

First submitted 8 March 2003; reviews returned 24 September 2003; final acceptance 16 November 2003

Associate Editor: Jack Sullivan