Evolution, 52(4), 1998, pp. 978–987

BEST-FIT MAXIMUM-LIKELIHOOD MODELS FOR PHYLOGENETIC INFERENCE: EMPIRICAL TESTS WITH KNOWN PHYLOGENIES

C. W. CUNNINGHAM,[1,2] H. ZHU,[1] AND D. M. HILLIS[3]
[1]*Zoology Department, Duke University, Durham, North Carolina 27708*
[2]*E-mail: cliff@acpub.duke.edu*
[3]*Department of Zoology and Institute of Cellular and Molecular Biology, University of Texas, Austin, Texas 78712*

Abstract.—Despite the proliferation of increasingly sophisticated models of DNA sequence evolution, choosing among models remains a major problem in phylogenetic reconstruction. The choice of appropriate models is thought to be especially important when there is large variation among branch lengths. We evaluated the ability of nested models to reconstruct experimentally generated, known phylogenies of bacteriophage T7 as we varied the terminal branch lengths. Then, for each phylogeny we determined the best-fit model by progressively adding parameters to simpler models. We found that in several cases the choice of best-fit model was affected by the parameter addition sequence. In terms of phylogenetic performance, there was little difference between models when the ratio of short:long terminal branches was 1:3 or less. However, under conditions of extreme terminal branch-length variation, there were not only dramatic differences among models, but best-fit models were always among the best at overcoming long-branch attraction. The performance of minimum-evolution-distance methods was generally lower than that of discrete maximum-likelihood methods, even if maximum-likelihood methods were used to generate distance matrices. Correcting for among-site rate variation was especially important for overcoming long-branch attraction. The generality of our conclusions is supported by earlier simulation studies and by a preliminary analysis of mitochondrial and nuclear sequences from a well-supported four-taxon amniote phylogeny.

Key words.—Best-fit models, long-branch attraction, maximum likelihood, minimum evolution, phylogeny reconstruction.

Received September 22, 1997. Accepted April 16, 1998.

Recent computational and theoretical advances have made it practical to apply increasingly realistic models of DNA sequence evolution to the problem of phylogenetic reconstruction. Although it is well known that serious systematic error can be introduced when data violate the assumptions of a reconstruction model (Felsenstein 1978; Penny et al. 1987; Yang 1995), it is by no means obvious that more complex models always improve phylogenetic accuracy (Yang et al. 1994; Gaut and Lewis 1995; Yang 1997). There are at least two circumstances where simpler models may be more accurate. First, if long branches are actually adjacent in the true phylogeny, then the well-known phenomenon of long-branch attraction may reinforce the correct tree, even though the length estimate of the joining internal branch will be exaggerated (Felsenstein 1978; Penny et al. 1987; Yang 1996a). Second, complex models require that more parameters be estimated from the same amount of data than for simple models. If superfluous parameters are added, the sampling variance increases without providing additional phylogenetic signal. Errors in parameter estimation may compromise phylogenetic accuracy, especially for small datasets.

In this study, we use DNA sequences from known phylogenies of bacteriophage T7 to evaluate the performance of a likelihood framework that seeks to strike a balance between realism and the error introduced by parameter estimation. In this framework, nested models are compared by adding parameters sequentially (Goldman 1993; Yang 1994). After each new set of parameters is added, a likelihood-ratio test is performed that determines whether the additional parameters significantly improve the fit between the model and the data. This process is continued until the addition of parameters no longer represents a significant improvement over the simpler model. The most-complex model that is accepted by this procedure will be referred to as the best-fit model.

Evaluating the performance of best-fit models using computer simulations is problematic. Simulations are generally based on fairly simple models of DNA sequence evolution. In most cases, the set of models being evaluated include the models under which the sequences evolved. In the ideal world of computer simulations, the model-fitting procedure is expected to choose the true model under which the sequences evolved (Goldman 1993). In most phylogenetic studies, however, investigators work with data that cannot be neatly described by simple models. Finally, even when rate heterogeneity is allowed, computer simulations generally assume an absence of selection and complete independence among sites (Gaut and Lewis 1995; Huelsenbeck 1995; but see Miyamoto and Fitch 1995).

In this study, we are interested in applying model-fitting methods to DNA sequences that have evolved under more realistic conditions than computer simulations. To this end, we used DNA sequences from a series of known phylogenies constructed using lineages of the bacteriophage T7 (Cunningham et al. 1997). Although our experimental phylogenies are generated under relatively artificial conditions, they present a series of challenges to phylogenetic reconstruction methods.

First, selection in these bacteriophage lineages has resulted in parallel evolution at the DNA sequence level (Cunningham et al. 1997), which poses a challenge to methods of estimating among-site rate variation (Yang 1996b). Second, the skewed mutational bias of the mutagen nitrosoguanidine (Bull et al. 1993) challenges methods of estimating nucleotide transformational probabilities and the number of invariable sites (Fitch and Markowitz 1970; Fitch 1986; Shoemaker and Fitch 1989; Gu et al. 1995; Lockhart et al. 1996). These are similar to the challenges produced by skewed mutational biases seen

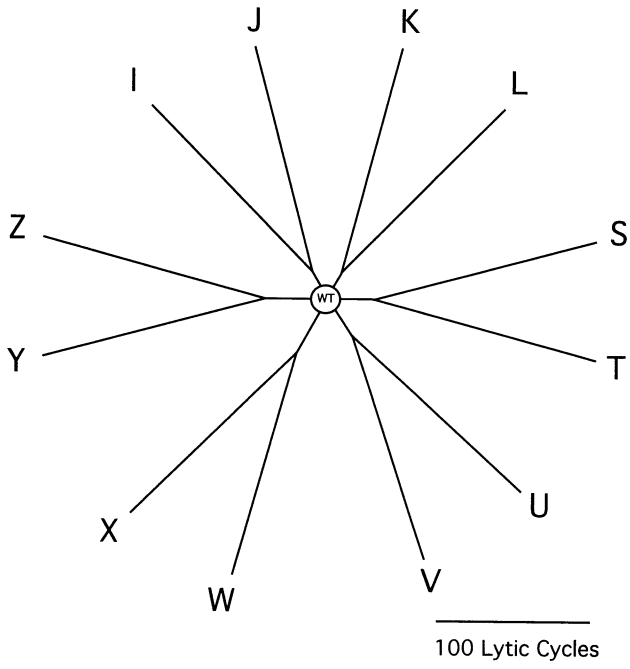

FIG. 1. Experimentally generated phylogeny of six independent bacteriophage T7 lineages. Each lineage was propagated from the same wild-type (WT) isolate and bifurcated once. The lengths of branches correspond to the number of lytic cycles, as described in Cunningham et al. (1997). The internal branches of the six lineages are not identical: IJ and KL = 10 cycles; ST and UV = 20 cycles; WX and YZ = 30 cycles. These lineages can be assembled into four-taxon phylogenies as shown in Figure 2.

in some natural systems (Gojobori et al. 1982; Moriyama et al. 1991). The phylogenies were designed so that the internal branches are very small and have long external branches. Furthermore, ancestors were collected at several points during the construction of each external branch. These features allow models to be compared when variation between branch lengths becomes progressively more extreme.

The experimental phylogenies were analyzed as three replicated four-taxon phylogenies and together as a 12-taxon phylogeny. For each phylogeny, we compared nested models that differed in parameters associated with base composition, models of substitution, and among-site variation. Among-site variation was incorporated in two ways, first by estimating the proportion of invariable sites and then by using a discrete gamma distribution to estimate the rate variation among the variable sites. Finally, although our experimental phylogenies are much more realistic than computer simulations, they may not be subjected to the same constraints as DNA sequences that have evolved over millions of years. Because of this, we investigated the generality of our conclusions with mitochondrial and nuclear genes from a widely accepted four-taxon amniote phylogeny.

MATERIALS AND METHODS

Experimental Phylogeny

Six lineages of bacteriophage T7 were propagated in the presence of the mutagen nitrosoguanidine from a single wild-type (WT) isolate as illustrated in Figure 1 (Cunningham et al. 1997). Each lineage was bottlenecked to a single individual (by random plaque selection) after either 10 (IJ and KL), 20 (ST and UV), or 30 lytic cycles (WX and YZ). A lytic cycle represents the time to fully lyse a culture of bacteria, which requires several generations of viruses. Each bottlenecked isolate was divided into two descendant terminal lineages as shown in Figure 1. Each of these terminal lineages was bottlenecked to a single individual at intervals of 50 lytic cycles. The purpose of this design was to create six independent pairs of sister taxa that could be assembled into various four-taxon phylogenies or analyzed together as shown in Figure 1.

Because we periodically bottlenecked the terminal lineages we had certain knowledge of the ancestral states at several points during propagation. This allowed us to vary branch lengths by simply considering these ancestors as terminal taxa. For each phylogeny, the pair of opposing branches that experienced the most convergence to one another remained long and the length of the other pair was progressively shortened. In each phylogeny, two opposing branches remained long and two branches were progressively shortened. For the KLWX, STUV, and IJYZ phylogenies, the LX, SU, and JY branches, respectively, remained long.

The Bacteriophage Data

DNA sequences were collected from both the Early and Late Regions of the T7 genome to yield a total of 2733 aligned nucleotide positions as described in Cunningham et al. (1997). The first segment ranged from position 797–3100 in the T7 genome (Dunn and Studier 1983). Because each lineage experienced deletions in the Early Region (Cunningham et al. 1997), the segment from 1210–2979 was not present in every lineage and was therefore omitted from the analysis. The remaining portion of the Early Region included 128 bp of intergenic sequence as well as portions of the 0.3 and 0.7 genes. The Late Region sequence extended from positions 34624–36822 and included the entire 17.0, 17.5, and 18.0 genes as well as some intergenic sequences. The Early Region data were presented in Cunningham et al. (1997), and the Late Region data are described here for the first time. The aligned sequences for all 2733 aligned positions are available from the EBI FTP server under accession code DS33256 either by anonymous ftp from FTP.EBI.AC.UK in directory /pub/databases/embl/align, from the world wide web at ftp://ftp.ebi.ac.uk/pub/databases/embl/align/, or by sending an e-mail message to netserv@ebi.ac.uk including the line GET ALIGN:DS33256.DAT.

The Amniote Data

One criticism of an experimental viral system is that our conclusions may not extend to naturally evolving DNA sequences. For this reason, it is especially important to analyze sets of extant taxa whose relationships are relatively noncontroversial. We applied our methods to DNA sequences from two nuclear (18S and 28S) and three mitochondrial genes (12S, 16S, valine tRNA) taken from representative taxa from a relatively noncontroversial and well-studied four-taxon amniote phylogeny (*Sceloporus undulatus, Alligator mississippiensis, Gallus gallus*, and *Mus musculus*: alignments

TABLE 1. Model fitting for the KLWX phylogeny with highly unequal branch lengths. JC, Jukes and Cantor 1969; F81, Felsenstein 1981; HKY, Hasegawa et al. 1985; GTR, general time reversible model, Lanave et al. 1984; INV, invariable-sites method, Hasegawa et al. 1985; GAM, discrete gamma distribution, four categories, Yang 1994. Data for the best-fit model appears in bold.

	Estimated parameters	Additional df	Likelihood	*P*-value
Sequence 1: Adding rate variation *after* increasing number of substitutional classes from two to six				
JC			4253.8	
F81	Base composition	3	4241.2	< 0.001
Substitution classes				
HKY	2 substitution classes	1	4162.2	< 0.001
GTR	6 substitution classes	4	4157.2	< 0.04
Among-site rate variation				
GTR/INV	**% of invariable sites**	**1**	**4148.6**	**< 0.001**
GTR/INV/GAM	discrete gamma distribution	1	4148.6	> 0.90
Sequence 2: Adding rate variation *before* increasing number of substitutional classes from two to six				
JC			4253.8	
F81	Base composition	3	4241.2	< 0.001
Substitution classes				
HKY	2 substitution classes	1	4162.2	< 0.001
HKY/INV	**% of invariable sites**	**1**	**4152.5**	**< 0.001**
HKY/INV/GAM	discrete gamma distribution	1	4152.5	> 0.90
GTR/INV	6 substitution classes	4	4148.6	> 0.10

and GENBANK accession numbers as described in Hedges et al. 1990; Hedges 1994; Huelsenbeck and Bull 1996).

Model Fitting

Because the topology has a strong effect on likelihood values, most approaches to model fitting compare likelihoods on the same topology (Yang, 1996b). First, a maximum-likelihood tree was estimated using a test version of PAUP* 4.0d61 (written by David L. Swofford, Smithsonian Institution) using the Jukes-Cantor (JC) model (Jukes and Cantor 1969). We then used the topology of this tree as a basis for calculating maximum-likelihood scores for progressively more complex models (see Table 1). The parameters added include base composition, numbers of substitutional classes (e.g., transitions vs. transversions represent two classes), and two approaches to incorporating among-site rate variation: the invariable sites method (Hasegawa et al. 1985; Palumbi 1989; Gu et al. 1995) and the four-category discrete gamma distribution (Yang 1994, 1996b).

As each new set of parameters was added, a likelihood-ratio test was performed to determine whether the simpler model could be rejected (Goldman 1993; Yang 1996b). These tests were carried out assuming that the likelihood-ratio statistic ($\delta = 2[\ln L_1 - \ln L_0]$) is distributed according to a chi-square distribution, which is valid because the likelihoods are calculated on the same topology (Yang 1996b). If the simpler model could not be rejected, any remaining parameters were added to the simpler model. For example, if increasing the substitutional classes from two to six types did not represent a significant improvement, rate variation was added to the model with two substitutional types. This procedure is illustrated in Table 1.

Some of these parameters are hierarchically nested. For example, the JC model, which assumes equal base frequencies, must be nested within the F81 model, which allows for unequal base frequencies. Similarly, models with different numbers of nucleotide substitution classes must be nested within one another in a hierarchical manner. For example, the F81 model (one class) is nested within the HKY model (two classes), which is nested within the GTR model (six classes). Finally, the invariable-sites method of accommodating rate variation can be nested within a combined model that not only estimates the proportion of invariable sites, but assumes a gamma distribution for the remaining sites.

It is important to note that these categories are not hierarchically nested with respect to one another. For example, rate variation can be added either before or after increasing the number of nucleotide substitutions from two to six (as in Table 1). Varying the parameter addition sequence can affect the choice of best-fit models. If these two sequences differed in their choice of best-fit model, the simplest of the best-fit models was preferred. Although the two addition sequences we used do not exhaust all of the possibilities, they are representative of the parameter-addition sequences found in the literature (Goldman 1993; Yang 1996b; Huelsenbeck 1997).

Phylogenetic Analysis, Tree Support, and Data Transformation

The performance of each model was evaluated by determining the support for correct internal branches across bootstrap pseudoreplicates (Hillis et al. 1994; Cunningham 1997). The starting seed for all bootstraps was always the same. This means that exactly the same pseudoreplicate datasets were evaluated by each reconstruction method being compared. For the four-taxon phylogenies, 10,000 bootstrap pseudoreplications were performed with PAUP* 4.0d61 (Swofford 1997) using exact search strategies. Distance searches were carried out using the minimum-evolution criterion (Kidd and Sgaramella-Zonta 1971; Rzhetsky and Nei 1992). During each bootstrap pseudoreplicate, if more than one resolution of a polytomy was equivalent with respect to the minimization criterion being applied, each alternative resolution was included in the final bootstrap consensus (the "no-collapse" option in PAUP* 4.d61).

For the 12-taxon phylogeny, bootstrapping was performed

for 100 pseudoreplicates with heuristic searches using a starting tree obtained by stepwise addition and tree-bisection-reconnection branch swapping with "no steepest descent" option in effect. Due to the time required for swapping during maximum-likelihood searches, all heuristic searches were performed with a maximum limit of 20 trees saved. Estimating tree support for the 12-taxon phylogeny was less straightforward. Because the six lineages are all descended from WT, the internal branches form a polytomy (Fig. 1). For this reason, only the bootstrap support values for each of the six pairs of sister taxa were considered.

Because the bootstrap support for the correct tree across the models and phylogenies varied from 0.04 to 0.99, bootstrap proportions were transformed to allow comparison across replicates and treatments. This was necessary because proportions are bounded, so that the difference between 0.49 and 0.50 is not comparable to the difference between 0.99 and 1.00. This bounded effect can be easily overcome by performing an arcsine transformation on the bootstrap proportion for every method being compared. This allowed the performance of each model to be expressed in terms of the deviation from the mean of all models being compared for a particular branch length/phylogeny combination.

Parameter Estimation

Before bootstrapping with maximum likelihood, all parameters were estimated in an iterative fashion. First, each parameter was estimated on the JC tree. Then the parameters were set to their estimated values and the search was repeated to yield a new tree. This process was repeated twice before setting the parameters to their final values for bootstrapping. Only base composition was estimated for every bootstrap pseudoreplicate. In most cases, it was prohibitively time consuming to estimate other parameters while bootstrapping. On limited runs, we found that estimating substitutional and rate variation anew for every bootstrap pseudoreplicate made little or no difference to our estimates of tree support (results not shown).

For all minimum-evolution searches, maximum-likelihood distances were calculated. These distances are calculated as the length of the single "branch" in a tree composed of the two taxa being compared, as estimated under the given model (Felsenstein 1995; Swofford et al. 1996). Maximum likelihood was used to estimate the appropriate parameters in the following manner. First, the shortest tree was determined under the minimum-evolution criterion for maximum-likelihood distances under the JC model. Then, the criterion was shifted to maximum likelihood to estimate the appropriate parameters. The criterion was shifted back to minimum evolution, and searches were performed with maximum-likelihood distances with the estimated parameter settings. This iteration was repeated again before setting the parameters for bootstrapping.

RESULTS

Four-Taxon Analyses: Tree with the Smallest Internal Branch

Because each of the six pairs of sister taxa are descended from the same WT individual, they can be assembled to form any of 15 possible four-taxon phylogenies. Of these possible phylogenies, the one with the smallest internal branch should be the most difficult to reconstruct when terminal branch lengths are allowed to vary (Felsenstein, 1978; Huelsenbeck and Hillis 1993). The phylogeny with the smallest internal branch (in terms of actual substitutions, not time) is formed by joining the lineages KL and UV. This phylogeny experienced only two substitutions during the 30 lytic cycles that compose its internal branch, compared with seven parallel evolutionary events between its two longest branches. Many of our major conclusions can be illustrated using this phylogeny.

Model Fitting

The phylogeny was subjected to three branch-length treatments: equal, with all terminal branches 150 lytic cycles in length; intermediate, with two branches 150 cycles long and two 50 cycles long; and highly unequal, with two branches 150 cycles long and two branches of length zero. Best-fit models were determined separately for each branch-length treatment. The GTR with invariable sites model was the best-fit model for the equal treatment (Fig. 2A), with HKY with invariable sites being preferred for the intermediate and highly unequal treatments (Fig. 2B,C). For these phylogenies, estimating a discrete gamma distribution for rate heterogeneity did not significantly improve the likelihood over simply estimating the proportion of invariable sites (e.g., Table 1).

Relative and Absolute Performance

For both the intermediate and equal branch-length treatments, the differences between the models were small (Fig. 2A,B). For the intermediate treatment, adding parameters improved tree support until the number of substitutional classes was increased from two to six, whereupon it dropped. For the equal treatment, adding parameters actually decreased performance relative to parsimony.

In contrast, adding model parameters to the highly unequal branch treatment had dramatic effects. When parsimony was applied, the incorrect phylogeny was supported in 96% of the bootstrap pseudoreplicates (Fig. 2C). The best-fit model, HKY with invariable sites, supported the correct tree with 86% bootstrap support. Of the parameters added, estimating the proportion of invariable sites most improved performance (Fig. 2C). For all three treatments, the performance of the most parameter-rich model—GTR with invariable sites—was lower than simpler models (Figs. 2B,C). The decreased performance of the most parameter-rich model is consistent with the hypothesis that too many parameters can have a negative effect.

Four-Taxon Analyses: Three Replicate Phylogenies

Model Fitting

Of the 15 possible rearrangements of the six lineages, three were chosen to form identical, replicate four-taxon phylogenies with internal branches of 40 lytic cycles in length (KLWX, STUV, IJYZ). Each of the three replicate phylogenies was subjected to the branch-length treatments described above. As before, in no case did estimating a discrete gamma

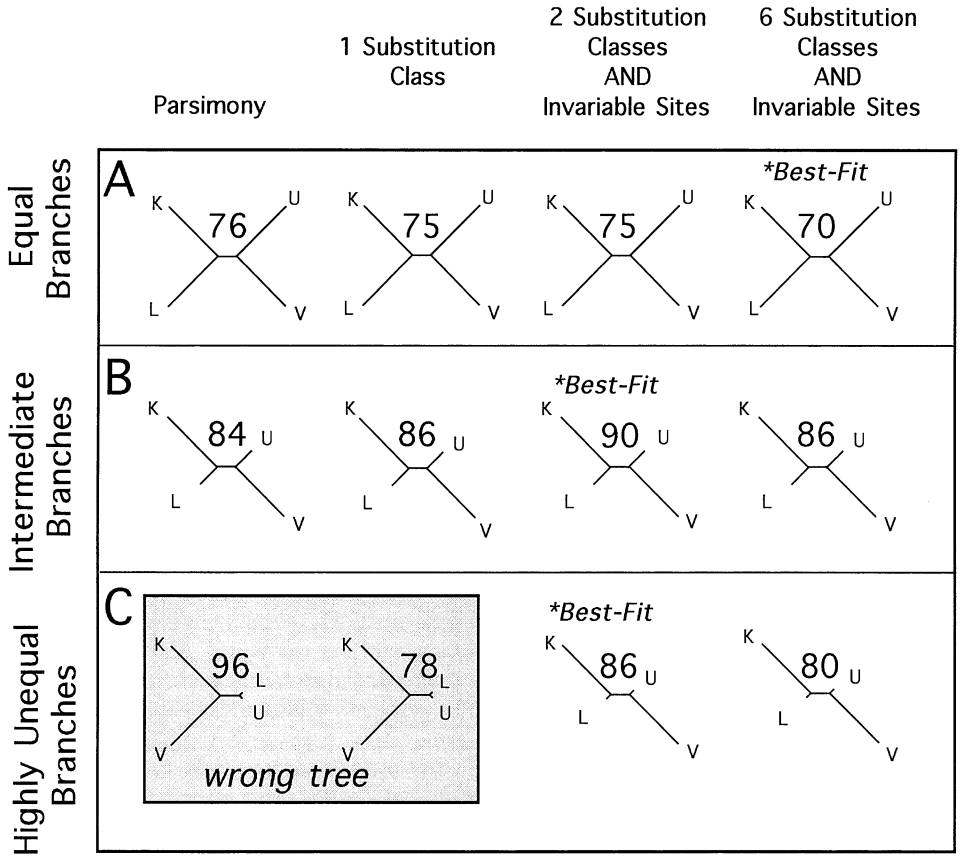

FIG. 2. Adding parameters to nested models can overcome long-branch attraction, especially when branch lengths are highly unequal (see trees in shaded box). Of the 15 possible four-taxon phylogenies that can be assembled from the six lineages, the phylogeny shown was the one with the smallest number of substitutions along the internal branch. The numbers represent bootstrap support (10,000 pseudoreplicates).

distribution for rate heterogeneity significantly improve the likelihood over simply estimating the proportion of invariable sites. The GTR model with invariable sites was preferred in six phylogeny/treatment combinations, and the HKY with invariable sites was preferred in three.

Relative Performance

For each branch-length treatment and model, the relative levels of support for the correct tree are shown in Figure 3. The choice of models made little difference in performance when the branch lengths were equal (Fig. 3A), but made increasingly more difference as the branches became more unequal (Fig. 3B,C). In all the phylogenies, adding the discrete gamma distribution for rate heterogeneity gave virtually the same performance as estimating the proportion of invariable sites alone. These results are shown from a different perspective for a subset of these models in Figure 4. The standard errors of the performance of the various models overlapped considerably for the equal treatment and were progressively more distinct in the intermediate, highly unequal branch treatments. For the highly unequal treatment, the best-fit model had either the highest level of performance or nearly so for every phylogeny (Figs. 3C, 4).

12-Taxon Analysis

Model Fitting

To investigate the effect of increasing number of taxa, the six lineages were analyzed simultaneously so that their internal branches formed a polytomy (Fig. 1). For this phylogeny, the addition of the discrete gamma distribution for rate heterogeneity was only rejected for the highly unequal treatment (Fig. 5).

Relative Performance

The best-fit model never resulted in the most support for the correct tree under any of the three treatments (Fig. 5A–C). When branches were equal or intermediate, the best-fit models showed among the lowest levels of performance (Fig. 5A,B). In contrast, in the highly unequal branch treatment, the best-fit model was among the best performing models (Fig. 5C).

As with the four-taxon analysis, the differences among models were much greater in the highly unequal branch treatment. Also, as before, adding the discrete gamma distribution for rate heterogeneity made little difference relative to simply estimating the proportion of invariable sites.

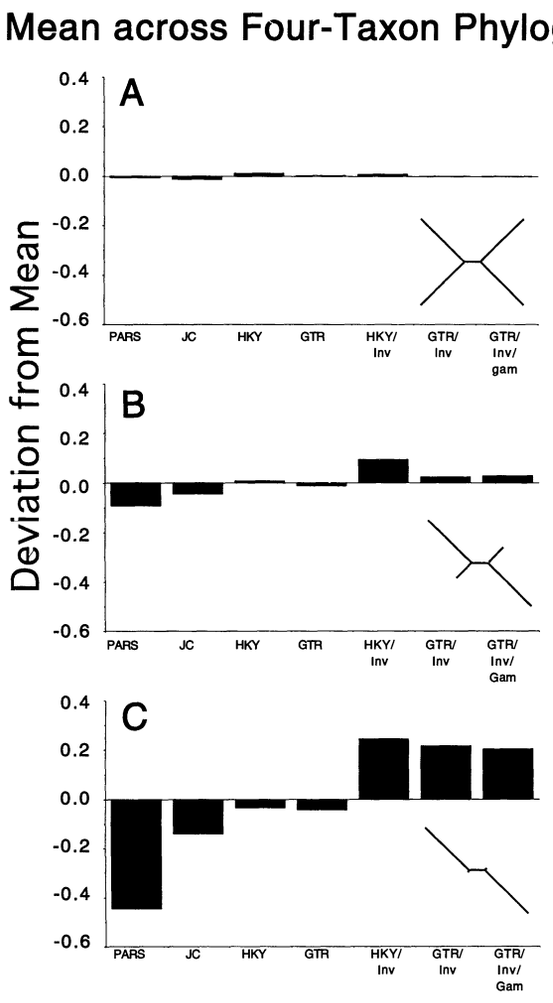

FIG. 3. Performance of nested models averaged across three replicate four-taxon phylogenies, each with an internal branch of 40 lytic cycles. Performance for each model/phylogeny combination was measured relative to the mean performance of all seven reconstruction methods shown. Units are in arcsine-transformed bootstrap proportions. As in Figure 2, the effect of adding parameters was the strongest when branch lengths are highly unequal. The abbreviations for the models are from Table 1.

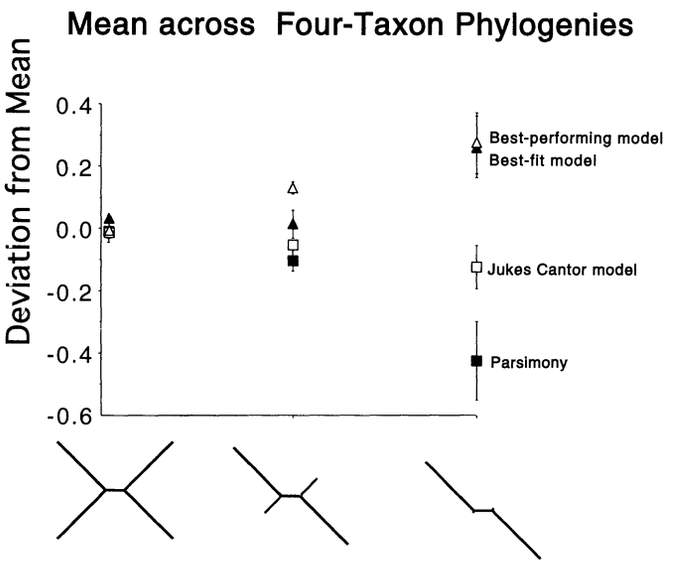

FIG. 4. The relative performance of a subset of the models shown in Figure 3: the best-fit model for each four-taxon phylogeny; the most succesful model for each phylogeny; and two simple models for comparison (PARS and JC). The means and standard errors are calculated from the three replicate phylogenies used in Figure 3. Units are in arcsine-transformed bootstrap proportions. As before, adding parameters makes the biggest difference when branches are highly unequal, although a smaller advantage was already apparent in the intermediate treatment.

Absolute Performance of Parsimony and Maximum Likelihood in the Bacteriophage Phylogenies

Although the main focus of this study is to evaluate the relative performance of phylogenetic methods, their absolute performance is also of interest (summarized in Table 2). In the four-taxon phylogenies, the range of support for the correct tree is very low for the KLWX phylogeny (with four substitutions on the internal branch, and six parallel substitutions between the L150 and X150 lineages). The degree of parallel evolution is not sufficient to overwhelm the phylogenetic signal from the much larger internal branches of the STUV and IJYZ phylogenies (13 and 15 substitutions, respectively).

The absolute performance of the various methods is much lower in the 12-taxon phylogenies, where the percentage of bootstrap support for two of the correct nodes is sometimes in the single digits and rarely rises above 30%, no matter what method is appliezd (nodes IJ, KL, Table 2). This finding is consistent with the expected difficulty of estimating a "star-burst" phylogeny with any method, especially when parallel substitutions are common among long branches connected by very small internal branches. Improvement in phylogenetic accuracy under these conditions is expected only by subdividing the long branches by adding appropriate taxa (Hillis 1996).

Likelihood versus Distance

When identical models were applied to each of the four-taxon phylogenies, the performance of likelihood and minimum-evolution-distance methods only differed when branch lengths were highly unequal (shown for three models in Fig. 6). In the highly unequal treatment, however, likelihood methods showed considerably higher levels of support for the correct tree than did the minimum-evolution criterion for all of the models shown in Figure 6 as well as for every other model being compared (results not shown). Virtually the same pattern was also observed in the 12-taxon phylogeny (results not shown).

In the 12-taxon analyses above, performance was measured by the bootstrap support for each of the six pairs of sister taxa. Another consideration is the extent to which methods falsely resolved the basal polytomy. In the highly unequal treatment, the strongest bootstrap support for an incorrect resolution of the basal polytomy ranged from only 20% to 34% under the likelihood criterion, but ranged from 52% to 64% under the minimum-evolution criterion. This suggests that the distance methods may be more likely to incorrectly

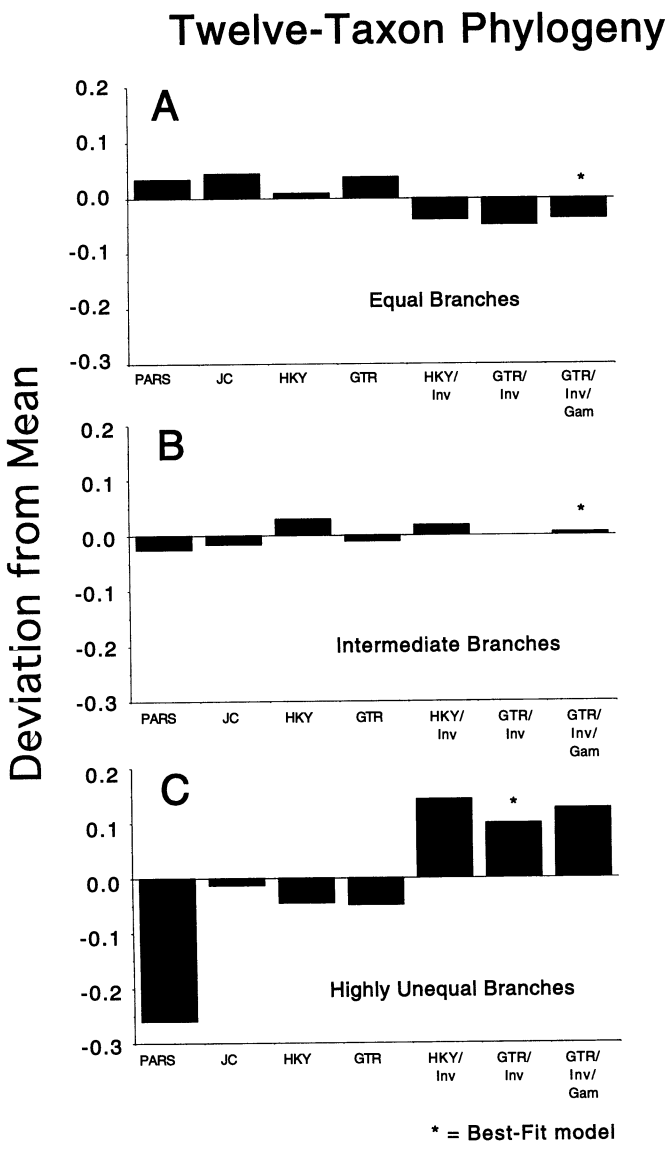

FIG. 5. Performance of each model when applied to the 12-taxon phylogeny shown in Figure 1. Because this phylogeny has a basal polytomy, performance was measured for the nodes supporting each of the six pairs of sister taxa. Deviations from mean are calculated as described in Figure 3. All performance is calculated from 100 pseudoreplicates.

resolve a true polytomy caused by the simultaneous divergence of lineages.

Amniote Phylogeny

When the amniote genes are mapped onto the expected phylogeny, the branch lengths for the mitochondrial genes are nearly equal and the branch lengths for nuclear genes are highly unequal (Fig. 7). For both nuclear and mitochondrial genes, adding the discrete gamma distribution for rate heterogeneity did not significantly improve the likelihood when compared to estimating the proportion of invariable sites alone.

These data confirm our major conclusions. First, differences among models are the greatest when branches are highly unequal (Fig. 7). Second, the best-fit model performs the best when the differences among the models are the greatest (Fig. 7). Third, adding a discrete gamma distribution for rate heterogeneity does not improve performance over estimating the proportion of invariable sites by itself. Finally, as with the bacteriophage phylogenies (Fig. 6), the differences between maximum-likelihood and minimum-evolution-distance methods were small when branch lengths were equal, whereas when branch lengths were highly unequal, the best-fit model was considerably more accurate when applied in a discrete maximum-likelihood framework than when it was used to convert DNA sequences to pairwise maximum-likelihood distances (results not shown).

DISCUSSION

We have used experimentally generated bacteriophage phylogenies to evaluate the performance of a likelihood framework for choosing among models of DNA sequence evolution. These phylogenies were designed to examine the performance of various models when branch lengths are allowed to vary. Because these bacteriophage sequences represent actual genes evolving under conditions of positive selection, they represent a far more realistic and complex model system than is possible in computer simulations.

In every phylogeny we studied, the best-fit models always included among-site rate heterogeneity. This result is not surprising, because our bacteriophage lineages are known to contain a large number of invariable sites due to a pronounced mutagen bias (Cunningham et al. 1997), and is consistent with empirical studies that have found significant rate heterogeneity in most DNA sequences surveyed, ranging from viruses to mammals (Sullivan et al. 1995; Yang et al. 1995; Huelsenbeck 1997).

When the branch lengths in our phylogenies were highly unequal, accounting for among-site variation had a stronger effect on phylogenetic performance than any other parameter (Figs. 3, 5). Consider our four-taxon bacteriophage phylogeny with the smallest internal branch, where there were only two substitutions along the internal branch and seven parallel evolutionary events between the long branches. Parsimony found 96% bootstrap support for the wrong tree, while the best-fit model found 86% bootstrap support for the *correct* tree (Fig. 2C). Similarly, for nuclear genes in the amniote phylogeny, with highly unequal branch lengths, accounting for rate heterogeneity made the most difference (Fig. 7), although no model was able to recover the expected tree (shown in Fig. 7).

These results are consistent with theoretical expectations and with simulation studies. Rate heterogeneity is known to increase the level of parallel evolution between lineages (von Haeseler and Churchill 1993), which will have the strongest negative effect on phylogeny reconstruction when some lineages are very long and unbranched. Simulation studies have also shown that models that do not account for rate heterogeneity can perform quite badly when branch lengths are highly unequal, even when these models are considered in a likelihood framework (Gaut and Lewis 1995).

Although complex best-fit models performed well with

TABLE 2. Range of absolute performance of parsimony and maximum-likelihood methods for each phylogeny.

Four-taxon phylogenies (proportion of 10,000 bootstraps)	KLWX phylogeny	STUV phylogeny	IJYZ phylogeny
Equal branches range	40.9–51.9	97.8–98.8	94.3–96.8
Intermediate branches range	36.5–60.8	91.0–95.2	98.9–99.1
Highly unequal branches range	26.0–97.6	77.0–99.9	93.7–99.9

12-taxon phylogeny (proportion of 100 bootstraps)	IJ node	KL node	ST node	UV node	WX node	YZ node
Equal branches range	13.0–24.8	10.0–20.8	73.7–89.0	27.0–50.0	15.0–20.0	67.0–89.1
Intermediate branches range	23.2–36.5	2.3–9.5	69.6–89.0	36.6–45.0	38.0–52.3	92.0–97.3
Highly unequal branches range	4.0–31.8	7.9–19.6	83.0–99.0	1.0–43.5	36.3–90.5	95.0–100.0

highly unequal branches for both the bacteriophage and amniote phylogenies, the difference in performance between simple and complex models was much smaller when branch lengths were intermediate and equal (Figs. 2–7), which agrees with the conclusions of a recent computer-simulation study (Yang 1997). In general, then, it seems that best-fit models appear to perform the best when the choice of the appropriate model is most important.

Explicit models of DNA sequence evolution can be applied to discrete characters in a maximum-likelihood framework or can be used to generate a distance matrix that can then be analyzed by any number of methods. When we compared the same models, maximum likelihood generally outperformed the minimum-evolution criterion of reconstructing phylogenies from distance matrices (Fig. 6). The advantage of maximum likelihood over distances was especially clear when branch-length variation was extreme. Our results are consistent with theoretical arguments that important information is lost when discrete characters are converted to distances. Although it has been argued that the information lost should increase with the number of taxa (Penny 1982), our results show that this loss of information is already apparent in the four-taxon case.

The advantage of discrete-character maximum-likelihood methods over distance methods incorporating the same models has been previously observed in simulation studies (Huelsenbeck 1995). This conclusion was also supported by a recent empirical study. Although the true phylogeny was not known, Huelsenbeck (1997) found that both simple and complex models were able to separate long branches in a likelihood framework, but only the most complex models were able to separate the long branches in a distance framework.

Conclusions and Recommendations

The best-fit model identified under the framework of a likelihood-ratio test seems to be a conservative choice of models. Although best-fit models did not always provide the highest level of support for the correct tree, their relative performance was the greatest when the choice of models was most important. More work is needed to determine whether it is possible to identify cases where simpler models should be preferred.

When adding parameters to models, it is important to remember that parameters can be added in different addition sequences. For example, rate variation can be added before or after increasing the number of substitutional categories from two to six (Table 1). We have shown that the order in which the parameters are added can affect the choice of parameters included in the best-fit model. We recommend varying the parameter addition sequence as we have and preferring the addition sequence that yields the simplest model.

Correcting for among-site variation can be extremely important, especially when branch lengths are highly unequal. In our phylogenies, the invariable-sites method was usually sufficient to account for rate heterogeneity, and the addition of the gamma distribution for the variable sites provided little or no additional resolution. This is significant because the invariable-sites method is considerably less computationally intensive than the discrete gamma method. For the datasets we examined, the discrete gamma method takes between two and five times longer than the invariable-sites method to evaluate the same number of trees. When we have applied these models to other datasets, the results of the two methods are very similar and the gamma method is always much slower. Nonetheless, we have obviously not examined all possible conditions, and the addition of a gamma distribution to model rate heterogeneity among variable sites may be critical for some datasets.

Finally, we have shown that discrete-character maximum likelihood methods show a considerable performance advantage over the distance-based minimum-evolution method, even when maximum-likelihood distances are used for minimum evolution. The advantage of discrete-character maximum likelihood appears to be greater for at least some empirical datasets than the small advantage sometimes seen for simulated data (e.g., Huelsenbeck 1995). There is a trade-off

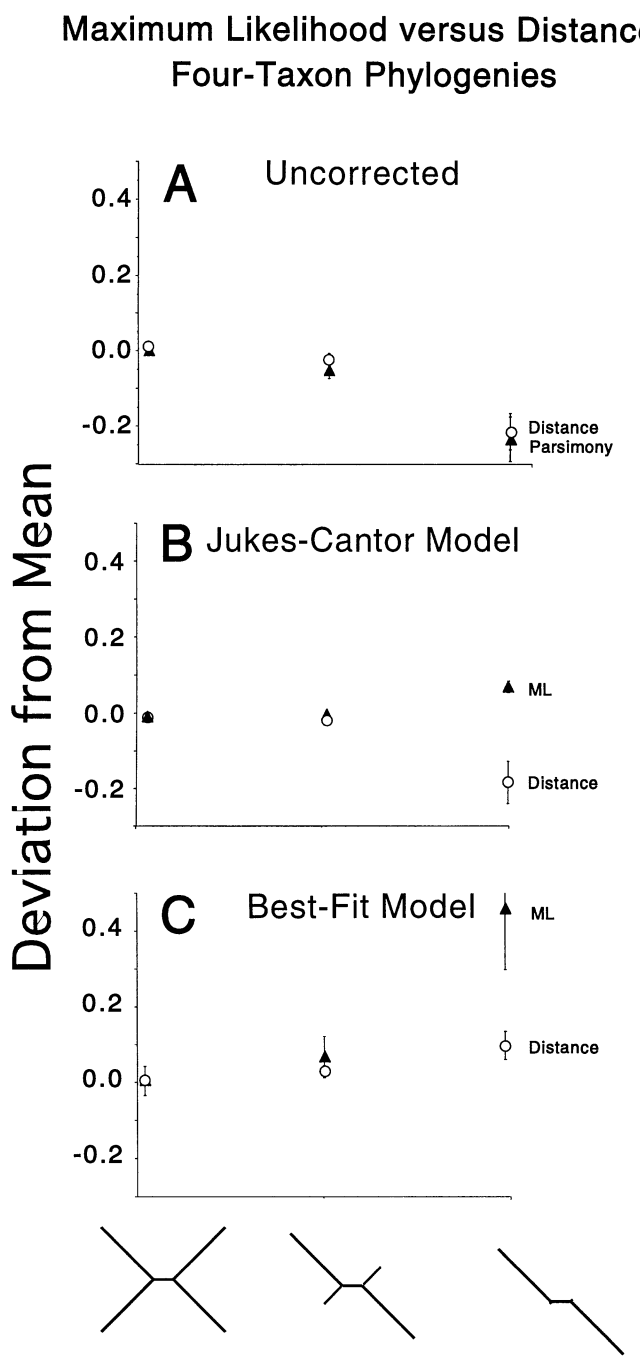

FIG. 6. Identical models are more successful at overcoming long-branch attraction in a likelihood framework than under the minimum-evolution criterion. These graphs represent the means and standard errors across the three replicate four-taxon phylogenies. Graphs are drawn as described in Figure 5. To allow the most appropriate comparison, all distances except for the uncorrected *p*-distance model were calculated using maximum-likelihood distances. The best-fit model and estimated parameters were determined under a maximum-likelihood framework as described in the text.

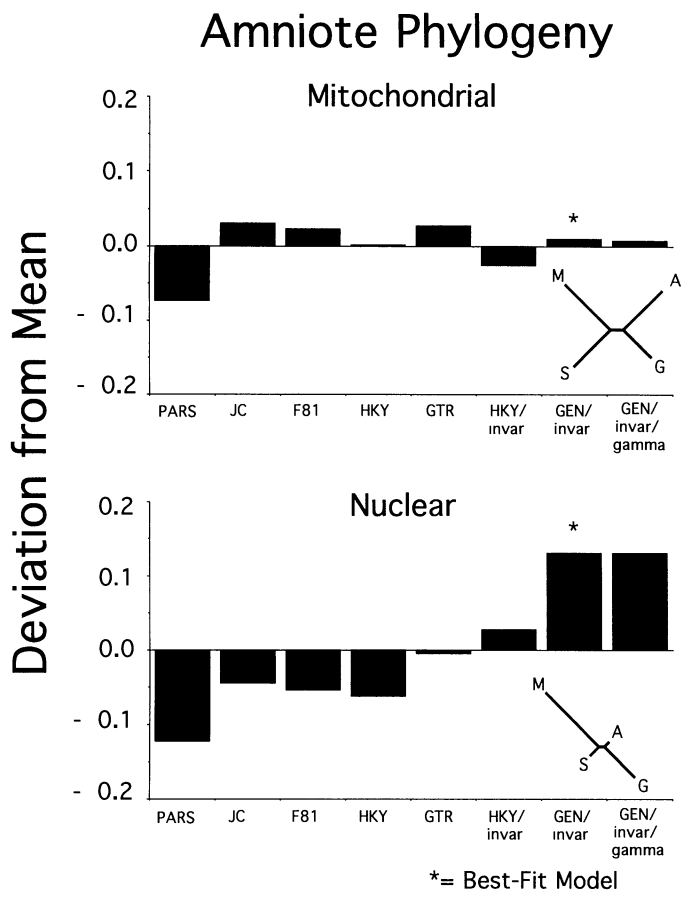

FIG. 7. Performance of nested models for nuclear and mitochondrial genes from a widely accepted amniote phylogeny (Hedges et al. 1990; Hedges 1994; Huelsenbeck and Bull 1996). As shown, the mitochondrial genes have roughly equal branches and the nuclear genes have highly unequal branches. As before, the best-fit models make the most difference when branches are not equal in length. The proportional length of the branches shown are estimated under the best-fit model. The actual lengths of the branches of the mitochondrial and nuclear genes are not comparable and are drawn to the same scale for heuristic purposes. M, *Mus musculus*; S, *Sceloporus undulatus*; A, *Alligator mississippiensis*; G, *Gallus gallus*.

between the advantage of distance methods in terms of computational speed and the advantage of discrete maximum likelihood in terms of performance. If phylogenetic accuracy is critical and the number of taxa is relatively small, the higher performance of maximum likelihood will usually be worth the additional computational effort.

ACKNOWLEDGMENTS

This paper benefited from the advice and encouragement of J. J. Bull as well as from the comments of two anonymous reviewers. We would like to thank D. Swofford for the use of test versions of PAUP* as well as K. Jeng, J. Husti, and M. Badgett for their sequencing efforts. This research was funded with the support of a National Science Foundation grant (DEB 9508987) and with start-up funds made available to CWC by Duke University.

LITERATURE CITED

BULL, J. J., C. W. CUNNINGHAM, I. J. MOLINEUX, M. R. BADGETT, AND D. M. HILLIS. 1993. Experimental molecular evolution of Bacteriophage T7. Evolution. 47:993–1007.

CUNNINGHAM, C. W. 1997. Is congruence between data partitions

a reliable predictor of phylogenetic accuracy? Empirically testing an iterative procedure for choosing among phylogenetic methods. Syst. Biol. 46:464–478.

CUNNINGHAM, C. W., K. JENG, J. HUSTI, M. BADGETT, I. J. MOLINEUX, D. M. HILLIS, AND J. J. BULL. 1997. Parallel molecular evolution of deletions and nonsense mutations in Bacteriophage T7. Mol. Biol. Evol. 14:113–116.

DUNN, J. J., AND F. W. STUDIER. 1983. Complete nucleotide sequence of Bacteriophage T7 DNA and the locations of T7 genetic elements. J. Mol. Biol. 166:477–535.

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

———. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

———. 1995. PHYLIP: phylogeny inference package. Vers. 3.5c. Distributed by the author, University of Washington, Seattle, WA.

FITCH, W. M. 1986. An estimation of the number of invariable sites is necessary for the accurate estimation of the number of nucleotide substitutions since a common ancestor. Pp. 146–160 *in* H. Gershowitz, D. L. Rucknagel and R. E. Tashian, eds. Evolutionary perspectives and the new genetics. Alan R. Liss, New York.

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. Gen. 4: 579–593.

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

GOJOBORI, T., W.-H. LI, AND D. GRAUR. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18:360–369.

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

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

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

HEDGES, S. B. 1994. Molecular evidence for the origin of birds. Proc. Nat. Acad. Sci. USA 91:2621–2624.

HEDGES, S., K. MOBERG, AND L. MAXSON. 1990. Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequences and a review of the evidence for amniote relationships. Mol. Biol. Evol. 7:607–633.

HILLIS, D. M. 1996. Inferring complex phylogenies. Nature. 383: 130–131.

HILLIS, D. M., J. HUELSENBECK, AND C. W. CUNNINGHAM. 1994. Application and accuracy of molecular phylogenies. Science 264:671–676.

HUELSENBECK, J. P. 1995. 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.

———. 1997. Is the Felsenstein zone a fly trap? Syst. Biol. 46: 69–74.

HUELSENBECK, J. P., AND J. J. BULL. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. Syst. Biol. 45:92–98.

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

JUKES, T. H., AND C. H. CANTOR. 1969. Evolution of protein molecules. Pp. 21–123 *in* H. M. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

KIDD, K. K., AND L. A. SGARAMELLA-ZONTA. 1971. Phylogenetic analysis: concepts and methods. Am. J. Hum. Genet. 23:235–52.

LANAVE, C., G. PREPARATA, C. SACCONE, AND G. SERIO. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20:86–93.

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. Nat. Acad. Sci. USA 93:1930–1934.

MIYAMOTO, M. M., AND W. M. FITCH. 1995. Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol. 12:503–513.

MORIYAMA, E. N., Y. INA, K. IKEO, N. SHIMIZU, AND T. GOJOBORI. 1991. Mutation pattern of human immunodeficiency virus genes. J. Mol. Evol. 32:360–363.

PALUMBI, S. R. 1989. Rates of molecular evolution and the fraction of nucleotide positions free to vary. J. Mol. Evol. 29:180–187.

PENNY, D. 1982. Towards a basis for classification: the incompleteness of distance measures, incompatibility analysis and phenetic classification. J. Theor. Biol. 96: 129–142.

PENNY, D., M. D. HENDY, AND I. M. HENDERSON. 1987. Reliability of evolutionary trees. Cold Spring Harbor Symp. Quant. Biol. 52:857–862.

RZHETSKY, A., AND M. NEI. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. 9:945–967.

SHOEMAKER, J. S., AND W. M. FITCH. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. 6:270–289.

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

SWOFFORD, D. L. 1997. PAUP*: phylogenetic analysis using parsimony (*and other methods). Vers. 4.0. Sinauer, Sunderland, MA.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. Hillis, C. Moritz, and B. K. Mable, eds. Molecular systematics, 2d ed. Sinauer, Sunderland, MA.

VON HAESELER, A., AND G. A. CHURCHILL. 1993. Network models for sequence evolution. J. Mol. Evol. 37:77–85.

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

YANG, Z. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105–111.

———. 1995. Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. J. Mol. Evol. 40:689–697.

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

———. 1996b. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–372.

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

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.

YANG, Z., I. J. LAUDER, AND H. J. LIN. 1995. Molecular evolution of the hepatitis B virus genome. J. Mol. Evol. 41:587–596.

Corresponding Editor: E. Martins