Syst. Biol. 52(5):674–683, 2003
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150390235494

Performance-Based Selection of Likelihood Models for Phylogeny Estimation

VLADIMIR MININ,[1,2,3] ZAID ABDO,[1,2] PAUL JOYCE,[1,2] AND JACK SULLIVAN[2,4]

[1]*Department of Mathematics, P.O. Box 441103, University of Idaho, Moscow, Idaho 83844-1103, USA*
[2]*Initiative in Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, Moscow, Idaho 83844, USA*
[3]*Department of Biomathematics, UCLA School of Medicine, AV-321 CHS, Los Angeles, California 90025-1766, USA*
[4]*Department of Biological Science, P.O. Box 443051, University of Idaho, Moscow, Idaho 83844-3051, USA; E-mail: jacks@uidaho.edu*

Abstract.— Phylogenetic estimation has largely come to rely on explicitly model-based methods. This approach requires that a model be chosen and that that choice be justified. To date, justification has largely been accomplished through use of likelihood-ratio tests (LRTs) to assess the relative fit of a nested series of reversible models. While this approach certainly represents an important advance over arbitrary model selection, the best fit of a series of models may not always provide the most reliable phylogenetic estimates for finite real data sets, where all available models are surely incorrect. Here, we develop a novel approach to model selection, which is based on the Bayesian information criterion, but incorporates relative branch-length error as a performance measure in a decision theory (DT) framework. This DT method includes a penalty for overfitting, is applicable prior to running extensive analyses, and simultaneously compares all models being considered and thus does not rely on a series of pairwise comparisons of models to traverse model space. We evaluate this method by examining four real data sets and by using those data sets to define simulation conditions. In the real data sets, the DT method selects the same or simpler models than conventional LRTs. In order to lend generality to the simulations, codon-based models (with parameters estimated from the real data sets) were used to generate simulated data sets, which are therefore more complex than any of the models we evaluate. On average, the DT method selects models that are simpler than those chosen by conventional LRTs. Nevertheless, these simpler models provide estimates of branch lengths that are more accurate both in terms of relative error and absolute error than those derived using the more complex (yet still wrong) models chosen by conventional LRTs. This method is available in a program called DT-ModSel. [Bayesian model selection; decision theory; incorrect models; likelihood ratio test; maximum likelihood; nucleotide-substitution model; phylogeny.]

The last decade has witnessed the emergence of statistical phylogenetics as the dominant view in systematics, and phylogeny estimation from DNA sequence data has reached a rather high level of sophistication. Increasingly complex likelihood models of sequence evolution are continually being developed, and the application methods such as the Markov chain Monte Carlo estimation have allowed these complex models to be incorporated into phylogeny estimation of large data sets (e.g., Leaché and Reeder, 2002). Furthermore, the ability to compare likelihood models objectively is a frequently cited advantage of explicitly model-based methods, such as maximum likelihood (ML) and Bayesian estimation, over competing methods (e.g., Swofford et al., 1996; Sullivan et al., 1997). This comparison is critical because it has been repeatedly demonstrated, both in simulations (e.g., Gaut and Lewis, 1995) and with real data sets (e.g., Sullivan and Swofford, 1997), that violation of model assumptions can lead to inconsistency of ML estimation. Such demonstrations have led many authors (e.g., Frati et al., 1997; Sullivan et al., 1997) to use statistical methods such as hierarchical likelihood-ratio tests (LRTs; Huelsenbeck and Crandall, 1997) to select a model for phylogeny estimation. Posada and Crandall (1998) automated model selection via LRTs (as well as other methods) with the production of Modeltest, which (in most cases) uses a decision tree and successive pairwise comparisons of nested models to traverse model space and select a model. This process has led to widespread use of hierarchical LRTs in model selection, which is an enormously important contribution to statistical phylogenetics.

Nevertheless, in spite of increasing model complexity, it is certainly the case that even our most complex and parameter-rich models are simplifications of the true evolutionary process that has generated any set of sequences. Thus, even the most well-justified model is surely wrong. Further, hierarchical LRTs can only provide information regarding the *relative* fit of the nested alternatives that have been examined; they can tell us nothing about the absolute goodness of fit of the chosen model. Although an absolute goodness-of-fit test does exist (e.g., Goldman, 1993; Sullivan et al., 2000; Demboski and Sullivan, 2003), there is no guarantee that the best-fit models will produce the best estimates of phylogeny. The reason being, at least for finite data, the relationship between fit (as measured by likelihood score) and performance is not as straightforward as one might wish. For example, Buckley et al. (2001) examined the performance of several models with regard to branch-length estimation from a data set containing 25 sequences of three mtDNA genes (COI, A6, and tRNAAsp) from *Maoricicada* and two outgroups. They found that both GTR+I+Γ and GTR+Γ models (applied to all the sites) provided uniformly larger (and probably better) estimates of branch lengths than did a 10-class site-specific rates (SSR) model (GTR+SSR_{10}), in spite of the fact that the SSR model is more parameter rich and has a better likelihood score than the former models. The improvement in fit Buckley et al. (2001) observed with the SSR model was attributable to more finely matching base frequencies that vary across partitions. The poor performance in branch length estimates under SSR was due to the presence of rate heterogeneity that was not accounted for under SSR. Thus, a single unit of improvement in fit with respect to rate heterogeneity has dramatically more effect on performance than a similar improvement in fit with respect to base frequencies. Models with the best likelihood score are not guaranteed consistently to

produce the best estimates of branch lengths from finite data and by extension should not necessarily be expected to perform best in phylogeny estimation.

Therefore, in spite of the improvement that LRTs represent over arbitrary model selection, there are at least two ways that further improvement in model selection may be achieved. First, the requirement that models be examined in series of pairwise comparisons presents difficulties in order of parameter subtraction (or addition in a bottom up approach). Thus, improvements in model selection may be achievable by employing simultaneous comparison of all models under consideration. Second, if the motivation for evaluation of alternative models is to choose one for estimation of phylogeny, improvement in model selection may be achievable by incorporating some measure of performance into a method for model selection. Here, we present a new method of model selection that incorporates both these modifications. With this method, all models under consideration are compared simultaneously using a Bayesian information criterion (BIC) approach that is modified by inclusion of a decision-theory framework (Bernardo and Smith, 1994) to evaluate performance of branch-length estimation.

We followed Buckley et al. (2001) in choosing branch-length estimation as the performance measure for a number of reasons. Since we're trying to optimize the model choice for phylogeny estimation, we looked for a method with at least the following attributes. First, it should be applicable prior to conducting extensive phylogenetic analyses. Second, given the first consideration, such a method should be based on some property that is intimately linked to the performance of phylogeny estimation. It has been repeatedly demonstrated (e.g., Gaut and Lewis, 1995; Sullivan and Swofford, 1997) that strong violation of model assumptions can mislead ML analyses in a manner similar to that in which parsimony is misled in the Felsenstein zone (i.e., where long branches are separated by a short internal branch). It is also well documented that this inconsistency is attributable to the systematic underestimation of branch length that affects long branches disproportionately more than short branches (e.g., Swofford et al., 2001). Thus, accuracy of branch-length estimation fulfils the second criterion of performance-based model selection. Furthermore, Sullivan and Swofford (2001) demonstrated that, even for Felsenstein-zone trees, phlyogeny estimation with some violated models performs as well as does estimation using the true model that was used to simulate the data. In that study, even some clearly wrong models were apparently able to estimate branch lengths sufficiently well to avoid long-branch attraction that afflicted estimation with very poor models. Thus, given the benefits of lower variance and shorter run times associated with simpler models, use of a simpler model should be favored over a more parameter-rich model when the simple model estimates branch lengths sufficiently well. This holds even when the simple model can be rejected using conventional LRTs of relative goodness of fit.

Here, we developed a performance-based method for selecting a likelihood model that can be used for subsequent ML or Bayesian estimation of phylogeny. We evaluated the alternative methods using a combination of simulations and analyses of real data sets. We took the approach that all the models being considered were wrong, and we attempted to find the model that would incur the least risk (i.e., perform the best) while still attempting to minimize the number of model parameters. In both real and simulated data sets, the decision theory (DT) method (as implemented in DT-ModSel) selects models that are simpler on average than those selected using current model selection procedures (e.g., LRTs as implemented in Modeltest; Posada and Crandall, 2001), yet these simpler models provide as good or better estimates of branch lengths as the more complex models selected by methods based on fit alone.

Decision Theory Background

The following conceptualization is very useful to illustrate decision theory. Let us suppose one is playing a game, the object of which is to choose an evolutionary scenario. At the end of the game, the true state of nature is revealed and penalties are assessed according to how far off one's guess is from the truth. However, the game is not played in complete ignorance. The data provide clues that, if used wisely, should lead to a reasonable choice that, if not completely correct, will at least receive a low penalty at the end of the game. Suppose, for example, that it is revealed a priori that one of M possibilities is the correct model of evolution under which the observed data were generated. One point is deducted for choosing the incorrect model and nothing is gained by choosing correctly. As we demonstrate below, under this penalty scheme, the optimal model-choice strategy is to calculate a statistic called the BIC (Bayesian Information Criterion) score and choose the model with the lowest score.

We begin by establishing some general notation. Let D be a particular data set that we are interested in analyzing. D is the information we use to make a model selection. The collection of models being considered can be denoted by $M_1, M_2, \ldots, M_m$. Associated with each model is a vector of parameters denoted by $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_m$, where $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{id_i})$, and d_i is the number of parameters for model M_i. We denote the likelihood of the data under model M_i with parameters $\boldsymbol{\theta}_i$ as

$$P(D \mid M_i, \boldsymbol{\theta}_i).$$

If we place a prior distribution of $\boldsymbol{\theta}_i$ under model M_i denoted by $g(\boldsymbol{\theta}_i \mid M_i)$ then the marginal probability of the data given only the model is defined as

$$P(D \mid M_i) = \int P(D \mid M_i, \boldsymbol{\theta}_i) g(\boldsymbol{\theta}_i \mid M_i)\, d\boldsymbol{\theta}_i.$$

The above represents a d_i-dimensional integral that would in principle be difficult to compute. However, Bayesian statistical theory (Raftery, 1995) provides the

following useful approximation:

$$\ln P(D \mid M_i) \approx \ln P(D \mid M_i, \hat{\boldsymbol{\theta}}_i) - (d_i/2) \ln n,$$

where n is the number of observations in the sample and $\hat{\boldsymbol{\theta}}_i$ is the vector of ML estimates. For phylogenetic analysis, where we assume that the tree topology is known and each site evolves independently along the branches of the tree, the sample size is the number of sites. The right side of the above equation is related to the BIC as follows:

$$\text{BIC} = -2 \ln p(D \mid M_i, \hat{\boldsymbol{\theta}}_i) + d_i \ln n \approx -2 \ln P(D \mid M_i).$$

Note that the approximation does not depend on the form of the prior distribution. Since we used the above approximation throughout, selection of a model was not influenced by prior probabilities of model parameters; only the data were used to make our final model selection.

For Bayesian model selection, the posterior probability of the model given the data is of primary concern. If we place a uniform prior probability on each model, that is $P(M_i) = 1/m$ for all i, then the usual Bayes formula relates BIC to posterior probabilities:

$$\begin{aligned} P(M_i \mid D) &= \frac{P(D \mid M_i)}{\sum_{j=1}^{m} P(D \mid M_j)} \\ &\approx \frac{e^{-BIC_i/2}}{\sum_{j=1}^{m} e^{-BIC_j/2}}. \end{aligned} \tag{1}$$

Since the denominator in Equation 1 is the same for all models, lower BIC scores correspond to higher posterior probabilities, and picking the model with the lowest BIC score is equivalent to picking the model with the highest posterior probability of being correct.

Now let us return to the game. Let l_{ij} be 1 if model i is chosen when model j is correct and $l_{ii} = 0$ if the true model is chosen. Since the true state of nature will not be revealed, we cannot directly compute this penalty. We can, however, compute an expected or average penalty conditional on the data. In the language of decision theory, this is the posterior risk associated with choosing model i. The posterior risk of choosing model i given the above 0 or 1 penalty function and assuming one of the models is correct, is defined by

$$R_i = \sum_{j=1}^{m} l_{ij} P(M_j \mid D) = 1 - P(M_i \mid D) \tag{2}$$

The optimal Bayesian model choice is the one that minimizes the posterior risk. We see from Equations 1 and 2 that minimizing the posterior risk is equivalent to maximizing the posterior probability of the model, which in turn leads to the BIC decision rule.

The above scenario demonstrates the usefulness of the conceptualization of model choice as a game with specific rules and penalties. However, the above game includes departures from an application to phylogenetics. The m statistical models that a phylogeneticist might propose are all only rough approximations to the truth; to assume a priori that one is correct is not reasonable. Also, suppose model i fits the data slightly better than does model j, but both models produce nearly the same inferences. Now suppose model k produces completely ridiculous results. It is not reasonable to assess the same penalty to model j as model k. Thus, performance may be incorporated into model selection through a risk function that allows for a nonbinary penalty function.

Performance-Based Model Selection

There are two aspects of a phylogeny that are of fundamental importance: the tree topology and the branch lengths (the rate of evolution times the time between each node or speciation event in the tree). Under model-based frameworks, if we assume momentarily that topology is known, we can focus attention on accurate branch-length estimates; rather than worry about the somewhat artificial criterion of whether or not a model is correct, we will focus on the accuracy of the branch lengths estimated under various models. If a simple model is returning estimates of branch lengths that are nearly identical to those from a more complex model, there will be little difference in phylogenetic estimation under the two models.

We assume that the phylogeny is described by an unrooted binary tree with k terminal nodes. As in current model selection procedures, we use a neighbor-joining tree based on LogDet distances. Therefore, there will be $2k - 3$ branches. We denote the vector of branch lengths by $\mathbf{B} = (B_1, B_2, \ldots, B_{2k-3})$. Let $\hat{\mathbf{B}}_i$ be the estimated branch lengths under the assumptions of model M_i. That is, $\hat{\mathbf{B}}_i$ is a function of the data D, the model M_i, and the ML estimates of the parameters $\hat{\boldsymbol{\theta}}_i$ under model M_i. Instead of a 0 or 1 penalty function, we develop a decision theoretic approach that penalizes models according to their performance with regard to branch length estimation. Consider the estimated vector of branch lengths under models M_i and M_j. The squared Euclidean distance between the branch-length estimates is given by

$$\|\hat{\mathbf{B}}_i - \hat{\mathbf{B}}_j\|^2 = \sum_{l=1}^{2k-3} (\hat{B}_{il} - \hat{B}_{jl})^2 \tag{3}$$

and the risk of choosing model M_i is given by

$$\begin{aligned} R_i &= \sum_{j=1}^{m} \|\hat{\mathbf{B}}_i - \hat{\mathbf{B}}_j\| P(M_j \mid D) \\ &\approx \sum_{j=1}^{m} \|\hat{\mathbf{B}}_i - \hat{\mathbf{B}}_j\| \frac{e^{-BIC_j/2}}{\sum_{j=1}^{m} e^{-BIC_j/2}}. \end{aligned} \tag{4}$$

R_i can therefore be calculated for each model, and that model with the minimum posterior risk can be chosen.

The following theoretical comparisons point out the advantages to our approach. First, in our approach each model is compared to all competing models at once; existing automated model-selection methods (e.g., LRTs in Modeltest) require a sequence of pairwise comparisons. Second, each model is weighted according to the posterior probability of the model conditional on the data. Since any model of evolution is only a crude approximation to reality, rather than focus attention on trying to find the "correct model" (e.g., Posada and Crandall, 2001), we have a measure of how plausible a model is given the data. Third, the decision theoretic framework allows for much flexibility. One can decide based on biologically relevant criteria what makes a model useful and use this criterion to assess a higher penalty to models that do not meet the criterion than to those that do. Our method combines branch-length estimates, model fit, and a penalty for overfitting in a statistically rigorous way.

Data Sets and Simulations

We chose several data sets with different patterns of evolution to assess the performance of the model selection methods under different conditions. First, we examined 12 primate mitochondrial DNA (mtDNA) sequences (Hayasaka et al., 1988) with a high average rate of evolution and strong heterogeneity among sites. This data set is included as an example in the distribution of PAUP* (Swofford, 1998) and has been used in many studies in which new phylogenetic methods have been developed (e.g., Yang, 1994). Second, we decided to focus on nuclear elongation factor 1α (EF-1α) sequences in *Ips* beetles (Cognato et al., 2001). In analyses of this data set, we excluded closely related taxa to do large scale simulations (AF397613, AF397617, AF397619, AF397621 AF397623, AF397624, AF397625, AF397626, AF397631, AF397632, AF397633, AF397634, AF397635, AF397636, AF397638, AF397642, AF397648, AF397649). The choice of this data set was motivated by the fact that nuclear DNA sequences have slower average rate of evolution and less rate heterogeneity among sites relative to mtDNA. Third, we examined a typical phylogeographic data set. We took 43 mtDNA cytochrome *b* (cyt *b*) sequences of Sumichrast's harvest mice, *Reithrodontomys sumichrasti* (Sullivan et al., 2000) and excluded redundant and closely related individuals to reduce the number of sequences to 14. Fourth, we examined another rodent cyt *b* data set, which includes 22 species of sigmodontine rodents (T. Rinehart et al., unpubl.), that we downloaded from GenBank (accessions AY041185–AY041206). We emphasized this last data set because, as is commonly the case, use of LRTs to compare models resulted in acceptance of the most general and parameter rich of the commonly used alternatives (i.e., the GTR+I+Γ model; Table 1).

For each of the data sets, we first conducted an ML search using PAUP* (Swofford, 1998) under a single model determined by conventional LRTs. We then parsed each data set by codon position and used the previous ML tree to optimize branch lengths and parameters of the GTR+I+Γ model for each codon position (again using PAUP*). We then used the codon-specific branch lengths and model parameters in conjunction with Seq-Gen, v. 1.2.5 (Rambaut and Grassly, 1997) to generate first, second, and third codon-position data sets for each replicate. Finally, we merged codon positions to form a single replicate data set, which was thus generated with a separate GTR+I+Γ model for each codon position (each of which was estimated from the original data). This strategy was designed to generate simulated data using a model that is much more complex than any of the candidate models typically used for phylogeny estimation. Each replicate was then subjected to model selection using both conventional LRTs (as implemented in Modeltest) and our new performance-based DT method (as implemented in DT-ModSel). We conducted 1,000 replicates for each of these four real data sets described above, and evaluated the accuracy of branch lengths estimated under the model chosen by each method. This was done both in terms of absolute and relative branch-length error. For the former, the vector of true branch lengths was calculated by averaging position-specific true branch lengths across codon positions (which is valid because branch lengths are expressed in expected substitutions per site and there is the same number of sites in each codon position). In mathematical language, if $\mathbf{B}$ is the vector of "true" branch lengths and $\hat{\mathbf{B}}_i$ is the estimated branch-length vector under model i chosen by the particular model selection criteria, then $\|\hat{\mathbf{B}}_i - \mathbf{B}\|^2$ is the absolute error. If model j produces branch-length estimates closest to the true branch lengths but model i is chosen, then $\|\hat{\mathbf{B}}_i - \hat{\mathbf{B}}_j\|^2$ is the relative error.

TABLE 1. Four sample data sets and the results of model selection using traditional LRTs and the DT methods presented here applied to the real data.

Data set	No. taxa	Sequence length (bp)	Model selected, no. parameters	
			LRT	DT
Primate mtDNA	9	693	TVM+Γ, 8	TrN+I+Γ, 7
Beetle EF-1α	20	553	TrN+Γ, 6	TrN+Γ, 6
Harvest mouse Cyt *b*	14	1,130	TrN+I+Γ, 7	TrN+I+Γ, 7
Sigmodontine Cyt *b*	22	772	GTR+I+Γ, 10	HKY+I+Γ, 6

Simulation Results

In general, application of DT-ModSel to the replicate data sets chose simpler models on average than did application of LRTs (Fig. 1). The one exception to this is the harvest mouse phylogeography data set in which the LRTs chose a slightly simpler model on average (Fig. 1c). This data set represents the most recent divergences we examined. Interestingly, for the fourth data set, which is from the same gene (mitochondrial cyt *b*) in a group

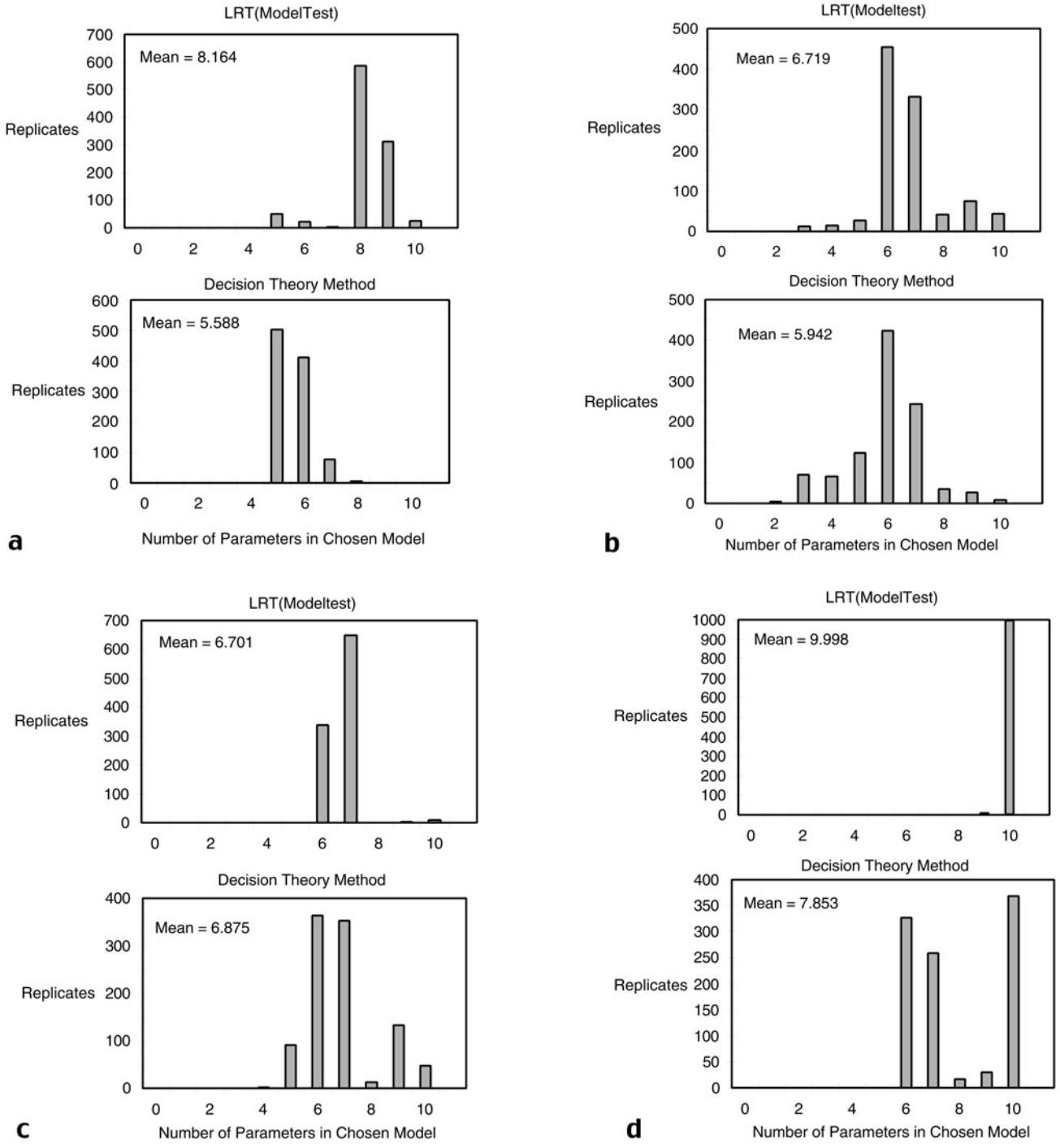

FIGURE 1. The distributions of the number of parameters in the model chosen by conventional LRTs (as implemented in Modeltest; top) and the new DT method presented here (as implemented in DT-ModSel; bottom) for the four conditions simulated. Data were simulated using conditions estimated from (a) the primate mtDNA, (b) the beetle EF1-α data, (c) the harvest mouse cyt *b* data, and (d) the sigmodontine cyt *b* data. Each distribution represents 1,000 simulations conducted with the codon-based model described in the text, and the only models considered were applied across codon positions.

of rodents that spans much deeper divergences, the DT algorithm selected substantially simpler models, both in the simulations (Fig. 1d) and in the real data set (Table 1). This is probably a very representative data set in that application of the LRT approach to the real data indicated that the most general and parameter-rich alternative (GTR+I+Γ) was required; this is frequently the case for real data. In terms of relative branch-length error, the models chosen by the DT algorithm always performed slightly better than did the models chosen by the LRT (Fig. 2). This result was expected because DT-ModSel was designed to choose the simplest model that minimizes relative branch-length error. The performance of models chosen by DT-ModSel was improved in terms

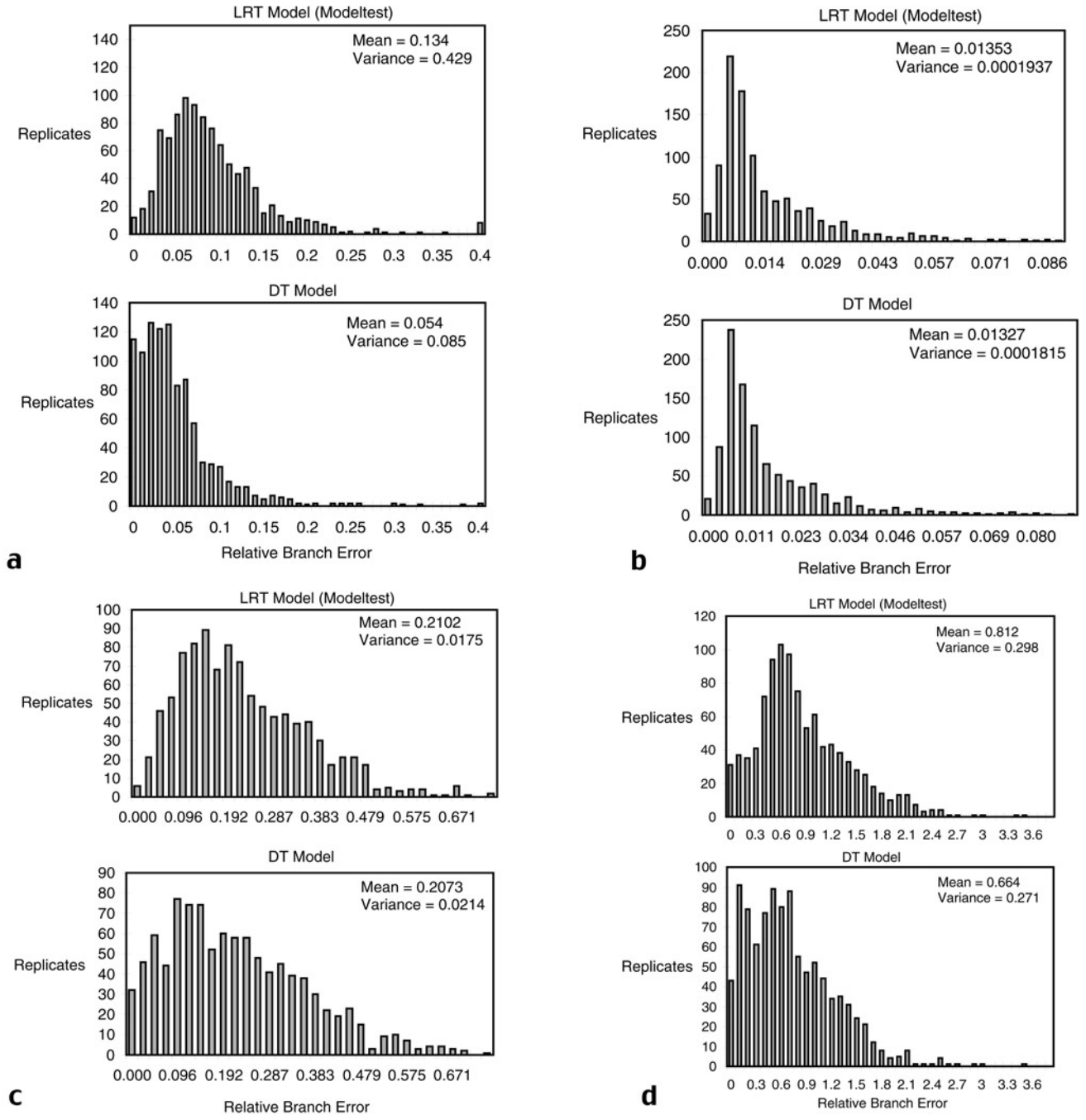

FIGURE 2. The distributions of the relative error in branch lengths estimated by the model chosen by conventional LRTs (as implemented in Modeltest) and the new DT method presented here (as implemented in DT-ModSel) for the four conditions simulated. Data were simulated using conditions estimated from (a) the primate mtDNA, (b) the beetle EF1-α data, (c) the harvest mouse cyt b data, and (d) the sigmodontine cyt b data. Relative branch error was calculated as described in the text.

of relative branch-length error obtained across data sets, but improvement was most pronounced in simulations based on the primate mtDNA and sigmodontine rodent cyt *b*. These two data sets are more complex than the others (Table 1). In the beetle EF-1α data set, there is relatively little heterogeneity in the rate matrix ($R = 1.00$, 3.89, 1.00, 1.00, 8.47, 1.00), whereas in the harvest mouse cyt *b* data the divergences are relatively shallow. Some very interesting patterns emerged in the assessment of absolute branch-length error (Fig. 3). First, in all the data sets, the model chosen via the DT algorithm estimated the true branch lengths slightly better on average than did the model chosen using LRTs. This observation is particularly interesting when coupled with the observation that DT-ModSel generally selected simpler models than did Modeltest (Fig. 1). One might expect that more complex incorrect models would always outperform incorrect simpler models, but this clearly is not the

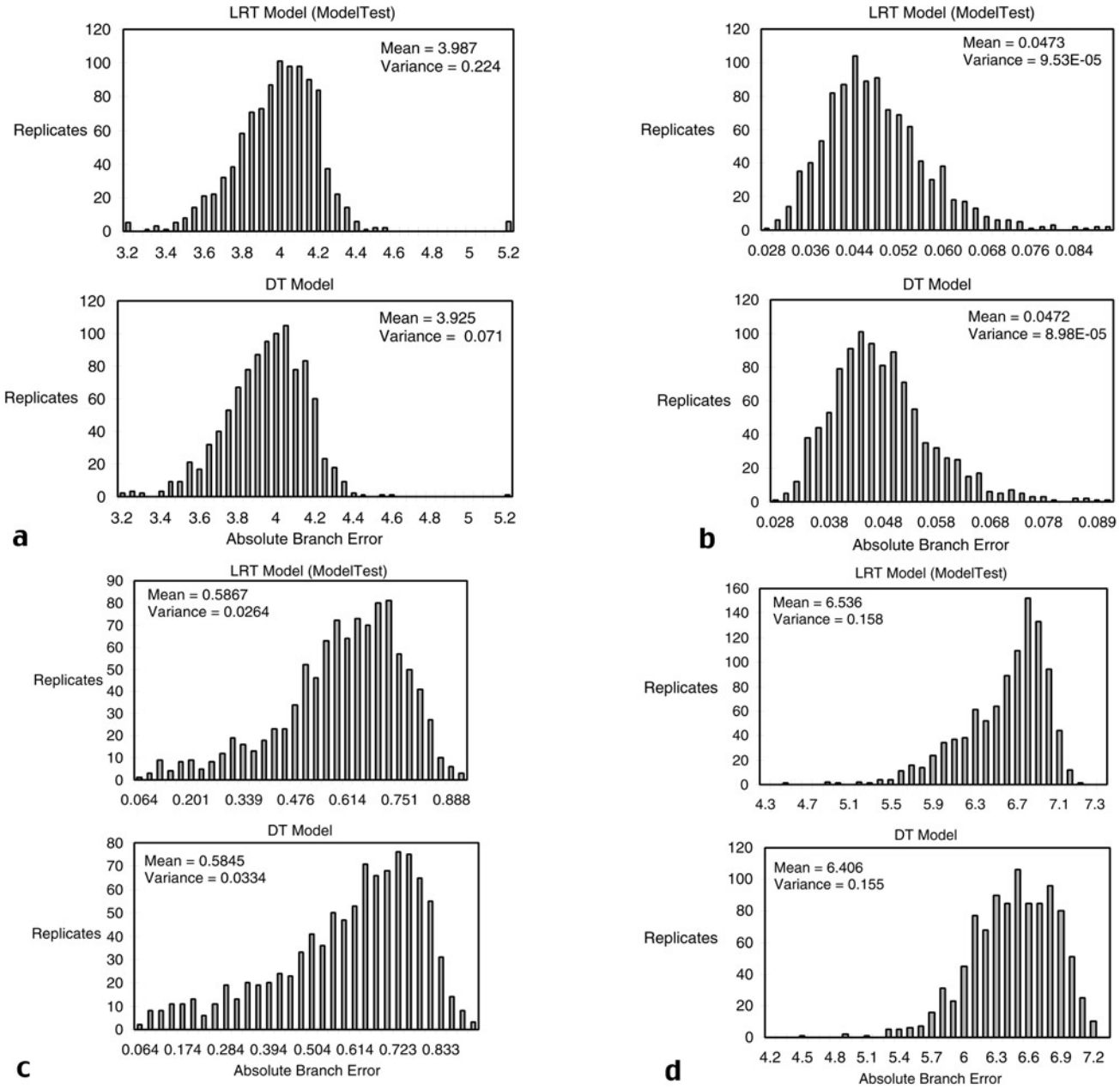

FIGURE 3. The distributions of the absolute error in the branch lengths estimated by the model chosen by conventional LRTs (as implemented in Modeltest; top) and the new DT method presented here (as implemented in DT-ModSel; bottom) for the four conditions simulated. Data were simulated using conditions estimated from (a) the primate mtDNA, (b) the beetle EF1-α data, (c) the harvest mouse cyt *b* data, and (d) the sigmodontine cyt *b* data. The absolute branch-length error is calculated as the Euclidean distance between the vector of ML branch lengths estimated under the model and the vector of true branch lengths used to simulate the data.

case. When the data were generated by a separate GTR+I+Γ model for each codon position and analyzed with a single model applied to all the sites, the simpler model chosen by the DT algorithm performed better on average than did the more complex model chosen by LRTs. Second, for the primate mtDNA and sigmodontine cyt *b* data, there is a large amount of absolute error in branch-length estimation. Even though there is not a great deal of relative branch-length error for either of these data sets, even the best relative estimates are quite poor in absolute terms. This suggests that an important extension of DT-ModSel will be to incorporate tests of likelihood models based on codon position.

Discussion

One obvious limitation of our method (both in application to real data and in our simulations) is that we used a single tree topology at the very beginning and never changed it, whereas for real data, the true tree is unknown. However, Posada and Crandall (2001) showed that use of different initial trees (e.g., a neighbor-joining or maximum-parsimony tree) does not affect the accuracy of model selection methods unless a random tree is chosen as the initial tree. This finding is expected given the nature of the relationship between topology and model parameters (Sullivan et al., 1996), use of an initial tree should not confound model selection using LRTs. However, this conclusion may not extend to our DT method. Not only do we rely on an initial topology, but we also use the vector of branch lengths from that topology under the model being considered, using point estimates of its parameters. This is certainly a limitation of our DT method. Suchard et al. (2002) developed a method for selecting models that incorporates uncertainty across topologies, and Bollback (2002) developed a test of the adequacy of a model that incorporates uncertainty in both topology and model parameters. Nevertheless, our use of initial point estimates of topology and model parameters permits the selection of a model prior to running extensive analyses that will perform as well as or better than any of the alternatives considered.

Performance Under Alternative Models

We used the sigmodontine cyt *b* data to assess the effect of selecting models using our DT method versus conventional LRTs. We subjected this data set to a series of analyses using the Hasegawa–Kishino–Yano HKY+I+Γ model selected by the DT method and the GTR+I+Γ model selected by LRTs. First, we conducted ML searches (stepwise addition, 10 random addition sequences with tree bisection–reconnection branch swapping) under each model using a 1-GHz Macintosh G4. The analysis using the HKY+I+Γ model ran in 1 hour 8 minutes, whereas the analysis using GTR+I+Γ ran in 1 hour 32 minutes. Slightly different trees were found using each model. The analysis under HKY+I+Γ found a single ML tree (Fig. 4), whereas the analysis under GTR+I+Γ found three different trees, one of which is not fully resolved (e.g., has one zero-length internal branch). The nonbifurcating tree of these four ML trees has a single internal polytomy (formed by collapsing node E in Fig. 4), and the other three trees represent the alternative resolutions compatible with that polytomy. Thus, searches under the two models essentially resulted in the same topology, but the search run under the model selected by DT-ModSel required one-third less time. We also conducted Bayesian analyses (Mr Bayes; Huelsenbeck and Renquist, 2001) under the two different models (Table 2). In general, there would be no differences in the interpretation of nodal supported as estimated under the model selected by DT-ModSel (HKY+I+Γ) and conventional LRTs (GTR+I+Γ). Nodes are strongly supported (posterior probability > 0.95) in both analyses, moderately supported (0.85 < posterior probability < 0.95) in

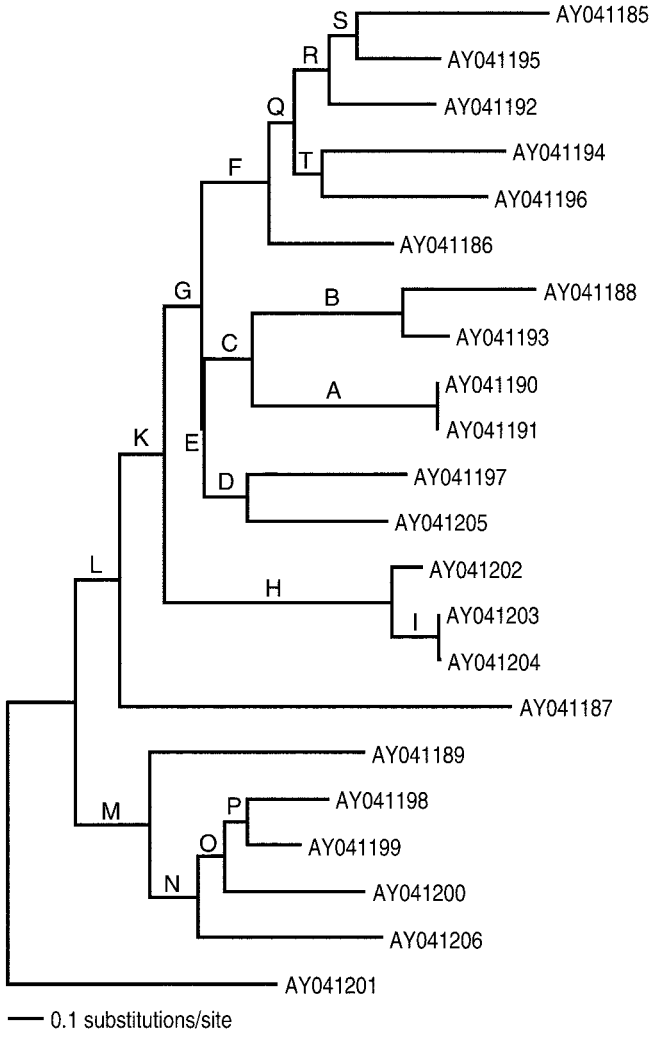

FIGURE 4. The ML tree for the sigmodontine cyt *b* data estimated using the HKY+I+Γ model selected by the DT method. Estimation under the GTR+I+Γ model (selected by conventional LRTs) resulted in three trees. In one of the GTR+I+Γ ML trees, node E was collapsed into a polytomy and the other two GTR+I+Γ ML trees are alternative resolutions of that polytomy.

TABLE 2. Posterior probabilities for sigmodontine cyt *b* data across nodes estimated under the GTR+I+Γ model, which was selected by conventional LRTs, and under the HKY+I+Γ model, which was selected by the DT method.

	Posterior probability[a]	
Node	HKY+I+G	GTR+I+G
A	100	100
B	100	100
C	83	80
D	81	82
E	31	NA
F	100	100
G	91	93
H	100	100
I	96	96
J	NA	33
K	100	100
L	100	100
M	98	97
N	87	94
O	70	72
P	61	66
Q	80	69
R	96	96
S	80	77
T	99	99

[a]Posterior probabilities under each model were estimated using MrBayes (Huelsenbeck and Ronquist, 2001), with two chains of 10^6 generations, a burn-in of 5×10^4 generations, and uniform priors. NA = not applicable.

both analyses, or poorly supported (posterior probability < 0.85) in both analyses (Fig. 5). This finding is consistent with those of Sullivan and Swoford (2001) that, even in the Felsenstein zone, phylongeny estimation with an incorrect HKY+I+Γ performs as well as estimation with the correct GTR+I+Γ model.

Next Logical Step

As currently implemented in DT-ModSel, our method is restricted to evaluating the same 56 models that are examined by Modeltest (Posada and Crandall, 2001); all these models are applied uniformly across a data set, without regard to codon position (or any other partitions that one may wish to consider). While the DT method we have developed here will allow for selection of the simplest model that outperforms the alternatives in terms of branch-length estimation, for two of the four conditions we simulated the chosen model still estimates branch lengths rather poorly (Fig. 3) in terms of absolute branch-length error. This result clearly indicates that application of a separate model to codon positons in protein-coding genes will frequently result in improved phylogenetic performance. The logical extension of the work presented here is to incorporate evaluation of partitions into model selection. For example, one might wish to treat all three codon positions separately or split just third positions from first and second positions. Similarly, one might wish to consider some formulation of a covarion model (Fitch and Markowitz, 1970; Miyamoto and Fitch, 1995). Basing such decisions on whether (and how) to incorporate codon structure and nonstationarity of the process across a tree on a metric that incorporates performance as well as fit (such as the DT method just described) should be an improvement over basing such decisions on fit alone (e.g., Huelsenbeck, 2002).

Program Availability

A Perl script that implements DT-ModSel is available at http://www.uidaho.edu/~jacks/DTModSel. We are

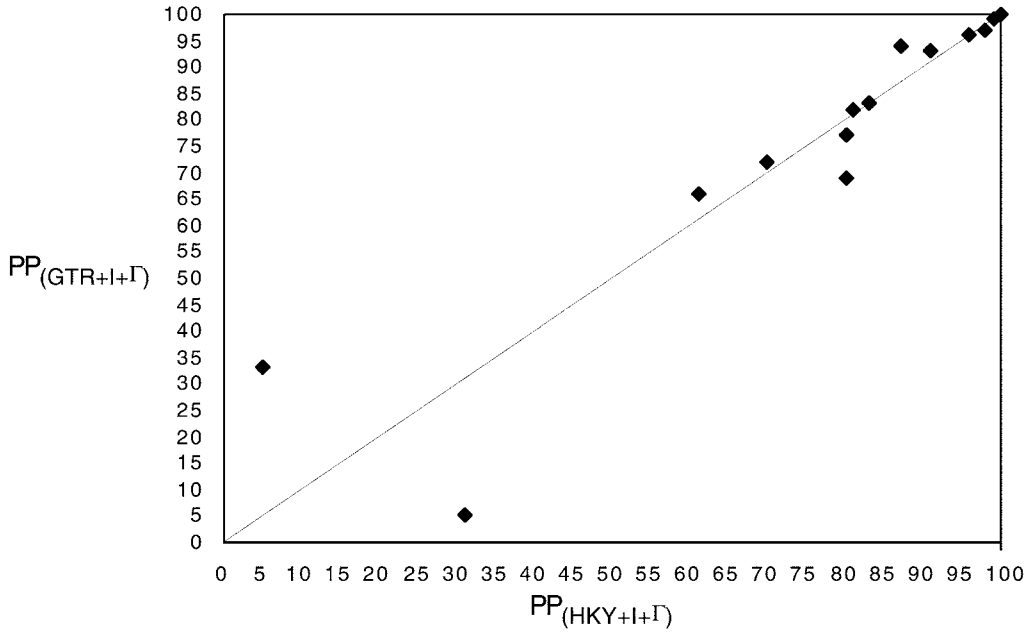

FIGURE 5. Posterior nodal probabilities (pp) for the sigmodontine cyt *b* data estimated under the GTR+I+Γ model (selected by LRTs) plotted against posterior probabilities estimated under the HKY+I+Γ model (selected by the DT method).

in the process of developing a Java-based program that will integrate with PAUP*; this program will replace the Perl script as soon as it is available.

ACKNOWLEDGMENTS

This research is part of the University of Idaho Initiative in Bioinformatics and Evolutionary Studies (IBEST). Funding was provided by NSF EPSCoR grant EPS-0080935 (to IBEST), NSF Systematic Biology Panel grant DEB-9974124 (to J.S.), NSF Probability and Statistics Panel grant DMS 0072198 (to P.J.), NSF EPSCoR grant EPS-0132626 (to P.J. and Z.A.), NSF Population Biology Panel grant DEB-0089756 (to P.J.), and NIH NCRR grant NIH NCRR 1P20RR016448-01 (to IBEST). We thank T. Rinehart, R. Graham, and H. Wichman for permission to use unpublished sigmodontine cyt *b* sequences, which were generated with funds from NIH grant GM38737 (to H. Wichman). We thank D. Swofford for many years of influential discussions regarding his dissatisfaction with simple strategies for model selection. We also thank members of IBEST for providing a stimulating intellectual environment and excellent comments and suggestions. Chris Simon, Mike Steel, David Posada, and an anonymous reviewer provided helpful comments that improved this paper.

REFERENCES

BERNARDO, J. M., AND A. F. M. SMITH. 1994. Bayesian theory. Wiley and Sons, New York.

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

BUCKLEY, T. R., C. SIMON, AND G. K. CHAMBERS. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: Effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst. Biol. 50:67–86.

COGNATO, A. I., AND A. P. VOGLER. 2001. Exploring data interaction and nucleotide alignment in a multiple gene analysis of *Ips* (Coleoptera: Scolytinae). Syst. Biol. 50:758–780.

DEMBOSKI, J. R., AND J. SULLIVAN. 2003. Extensive mtDNA variation within the yellow-pine chipmunk, *Tamias amoenus* (Rodentia: Sciuridae), and phylogeographic inferences for northwestern North America. Mol. Phylogenet. Evol. 26:389–408.

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

FRATI, F., C. SIMON, J. SULLIVAN, AND D. L. SWOFFORD. 1997. Evolution of the mitochondrial COII gene in Collembola. J. Mol. Evol. 44:145–158.

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.

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

HAYASAKA, K., T. GOJOBORI, AND S. HORAI. 1988. Molecular phylogeny and evolution of primate mitochondrial DNA. Mol. Biol. Evol. 5:626–644.

HUELSENBECK, J. P. 2002. Testing a covariotide model of DNA substitution. Mol. Bio. Evol. 19:698–707.

HUELSENBECK, J. P., AND K. A. CRANDALL. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. 28:437–466.

HUELSENBECK, J. P., AND F. RONQUIST. 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics 17:754–755.

LEACHÉ, A. D., AND T. W. REEDER. 2002. Molecular systematics of the eastern fence lizard (*Sceloporus undulatus*): A comparison of parsimony, likelihood, and bayesian approaches. Syst. Biol. 51:44–68.

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

POSADA, D., AND K. A. CRANDALL. 1998. Modeltest: Testing the model of DNA substitution. Biofinormatics. 14:817–818.

POSADA, D., AND K. A. CRANDALL. 2001. Selecting the Best-Fit Model of Nucleotide Substitution. Syst. Biol. 50:580–601.

RAFTERY, A. E. 1995. Bayesian Model Selection in Social Research (with discussion by A. Gelman, D. B. Rubin, and R. M. Hauser). Pages 111–196 *in* Sociological Methodology (V. Marsden ed.). Blackwells, Oxford, U.K.

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.

SUCHARD, M. A., R. E. WEISS AND J. S. SINSHEIMER. 2002. Bayesian selection of continuous-time Markov Chain evolutionary models. Mol. Biol. Evol. 18:1001–1013

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

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.

SULLIVAN, J., E. ARELLANO, AND D. S. ROGERS. 2000. Comparative phylogeography of mesoamerican highland rodents: Concerted versus independent response to past climatic fluctuations. Am. Nat. 155:754–768.

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

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

SWOFFORD, D. L. 1998. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0b10a. Sinauer Associates, Sunderland, Massachusetts.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference, pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

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.

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

First submitted 23 December 2002; reviews returned 30 March 2003; final acceptance 11 May 2003
Associate Editor: Mike Steel