Accounting for Uncertainty in the Tree Topology Has Little Effect on the Decision-Theoretic Approach to Model Selection in Phylogeny Estimation

Zaid Abdo,†‡ Vladimir N. Minin,§ Paul Joyce,*†‡ and Jack Sullivan*†¶*

*Initiative in Bioinformatics and Evolutionary Studies (IBEST), †Program of Bioinformatics and Computational Biology, and ‡Department of Mathematics, University of Idaho, Moscow; §Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles; and ¶Department of Biological Science, University of Idaho, Moscow

Currently available methods for model selection used in phylogenetic analysis are based on an initial fixed-tree topology. Once a model is picked based on this topology, a rigorous search of the tree space is run under that model to find the maximum-likelihood estimate of the tree (topology and branch lengths) and the maximum-likelihood estimates of the model parameters. In this paper, we propose two extensions to the decision-theoretic (DT) approach that relax the fixed-topology restriction. We also relax the fixed-topology restriction for the Bayesian information criterion (BIC) and the Akaike information criterion (AIC) methods. We compare the performance of the different methods (the relaxed, restricted, and the likelihood-ratio test [LRT]) using simulated data. This comparison is done by evaluating the relative complexity of the models resulting from each method and by comparing the performance of the chosen models in estimating the true tree. We also compare the methods relative to one another by measuring the closeness of the estimated trees corresponding to the different chosen models under these methods. We show that varying the topology does not have a major impact on model choice. We also show that the outcome of the two proposed extensions is identical and is comparable to that of the BIC, Extended-BIC, and DT. Hence, using the simpler methods in choosing a model for analyzing the data is more computationally feasible, with results comparable to the more computationally intensive methods. Another outcome of this study is that earlier conclusions about the DT approach are reinforced. That is, LRT, Extended-AIC, and AIC result in more complicated models that do not contribute to the performance of the phylogenetic inference, yet cause a significant increase in the time required for data analysis.

Introduction

Because molecular phylogenetics has come to be dominated by estimation methods that model explicitly the process of nucleotide substitution, greater attention is now being paid to the manner in which models are selected. Because of the computational intensity of model-based methods, one desirable property of a model-selection method is the ability to choose a model before performing an extensive phylogenetic analysis (Minin et al. 2003). Currently, the hierarchical likelihood-ratio test (LRT) approach and the Akaike information criterion (AIC), implemented in ModelTest (Posada and Carandall 1998), and the decision-theoretic (DT) approach and the Bayesian information criterion (BIC), implemented in DT-ModSel, follow this philosophy. Under these methods, the model is chosen based on an initial fixed-tree topology that is generated using a fast, approximate tree-building approach such as neighbor-joining or parsimony (e.g., Frati et al. 1997; Huelsenbeck and Crandall 1997; Sullivan, Markert, and Kilpatrick 1997). A rigorous search of the tree space is then run under this chosen model to find the maximum-likelihood estimate of the tree (topology and branch lengths) and the maximum-likelihood estimates of the model parameters. Posada and Crandall (2001) argue, based on their simulations, that, unless the initial tree is introduced at random, the initial topology will not have a major impact on model selection using LRT or AIC. This is true because there is negligible variation in parameter estimates across topologies that maintain well-supported nodes (Sullivan, Holsinger, and Simon 1996), and the order of models in terms of likelihood score rarely varies across trees that are close to optimal (e.g., Sullivan, Markert, and Kilpatrick 1997). However the extension of this conclusion to the decision theory method introduced by Minin et al. (2003), and implemented in DT-ModSel, is not clear, because that approach uses the Euclidian distance between branch-length vectors estimated under alternative models in erecting the risk function (see below).

In this paper, we address the issue of using an initial topology generated using neighbor-joining and holding it constant across models in model selection using DT-ModSel. We do this by developing two extensions to the DT approach that relax the fixed topology restriction. The aim of these extensions is to maintain a parsimonious selection in choosing the simplest models yielding the best performance, while accounting for variability across topologies into the selection criterion. As before (Minin et al. 2003), performance is measured by the loss of accuracy in branch lengths estimated under the alternative available models. These extensions violate the condition that a model be selected before extensive analyses by requiring a rigorous search of the tree space to find the optimal trees under each of the known models. We also relax the fixed initial-tree topology restriction for the BIC and the AIC methods.

Our purpose is to evaluate the impact of relaxing the reliance on the initial tree on model selection. To do so, we thoroughly evaluate the performance of the proposed extensions by comparing them to the other available methods, such as LRT, AIC, DT, and BIC. Model complexity and distance from the true tree are two main criteria used for this comparison. We, also, contrast the relative performance of the different methods to one another by measuring the distance between the resulting trees under each of them. These methods are then implemented on an

Key words: Decision-theoretic model selection, DT-ModSel, Bayesian information criterion, Akaike information criterion, hierarchical likelihood testing, ModelTest.

E-mail: abdo9538@uidaho.edu.

Mol. Biol. Evol. 22(3):691–703. 2004
doi:10.1093/molbev/msi050
Advance Access publication November 17, 2004

actual data set to compare the resulting selected models and the feasibility of using the extended methods in terms of the time expense.

The Extended Decision-Theoretic Approach

Minin et al. (2003) introduced a DT approach for selecting models for phylogentic estimation. The method weights the choice of a model (M_i) by its performance in estimating the branch lengths of a phylogenetic tree, in addition to its fit. Assuming that one of the available models is correct, using another model to estimate the branch lengths will result in error (loss) that can be measured by the difference between the branch lengths estimated under each of these models. The DT method chooses the model that has the minimum expected loss (posterior risk), which is calculated for each model M_i by multiplying the loss resulting from choosing model M_i, when another model, M_j, is "true," by the posterior probability of the "true" model, M_j, and then summing over all j models. We use "true" here in a loose sense to mean the model that provides the best approximation of reality (Burnham and Anderson 2002). Equation (1) represents the expected loss (posterior risk) of model M_i:

$$R_i = \sum_{j=1}^{m} \|\hat{\boldsymbol{B}}_{\boldsymbol{i}} - \hat{\boldsymbol{B}}_{\boldsymbol{j}}\| P(M_j \mid D) \tag{1}$$

R_i is the risk; $\|\hat{\boldsymbol{B}}_{\boldsymbol{i}} - \hat{\boldsymbol{B}}_{\boldsymbol{j}}\|$ is the loss function: the Euclidian distance between the estimated branch-length vector $\hat{\boldsymbol{B}}_{\boldsymbol{i}}$ using model M_i and the estimated branch-length vector $\hat{\boldsymbol{B}}_{\boldsymbol{j}}$ using model M_j. Each of these branch-length vectors has a length of $2N$–3 corresponding to the number of estimated branches in an unrooted tree (N is the number of taxa). $P(M_j \mid D)$ is the posterior probability of model M_j given the data. The posterior probabilities of the different models are used as a weighting measure to give the more likely models (i.e., those with better fit) higher weights than those that are less likely.

Applying Bayes' theorem and assuming flat priors on the models ($P(M_i) = 1/m$), the posterior probability of the model given the data becomes

$$P(M_i \mid D) = \frac{P(D \mid M_i)}{\sum_{j=1}^{m} P(D \mid M_j)} \tag{2}$$

Implementing the approximation by Raftery (1995):

$$\ln[P(D \mid M_i)] \approx \ln[P(D \mid M_i, \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{B}}_i)] - \frac{d_i + 2N - 3}{2} \ln[n] = -\frac{BIC_i}{2} \tag{3}$$

where $\hat{\boldsymbol{\theta}}_i$ is the vector of the maximum-likelihood estimates of the parameter vector, $\boldsymbol{\theta}_i = (\theta_1, \theta_2, \ldots, \theta_{d_i})$, of model M_i, and d_i is the number of parameters associated with that model. Equation (3) corrects a typographical error in the BIC equation presented in Minin et al. (2003); the branch lengths are parameters that are also estimated. These parameters contribute to the BIC, although their contribution results in a constant shift that does not affect the choice based on the BIC and that cancel out when calculating the posterior distribution of the models in the DT approach (as can be seen from equations (4) and (5) below). Substituting equation (3) in equation (2) we have

$$P(M_i \mid D) \approx \frac{e^{-BIC_i/2}}{\sum_{j=1}^{m} e^{-BIC_j/2}} = \frac{1}{\sum_{j=1}^{m} e^{(BIC_i - BIC_j)/2}} = \frac{1}{\sum_{j=1}^{m} e^{\left(2\left\{\ln[P(D \mid M_j, \hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{B}}_j)] - \ln[P(D \mid M_j, \hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{B}}_i)]\right\} + \ln[n](d_i - d_j)\right)/2}} \tag{4}$$

Substituting equation (4) in equation (1) results in Minin et al.'s (2003) risk function (adjusted for computational purposes):

$$R_i = \sum_{k=1}^{m} \frac{\|\hat{\boldsymbol{B}}_{\boldsymbol{i}} - \hat{\boldsymbol{B}}_{\boldsymbol{k}}\|}{\sum_{j=1}^{m} e^{(BIC_k - BIC_j)/2}} = \sum_{k=1}^{m} \frac{\|\hat{\boldsymbol{B}}_{\boldsymbol{i}} - \hat{\boldsymbol{B}}_{\boldsymbol{k}}\|}{\sum_{j=1}^{m} e^{\left(2\left\{\ln[P(D \mid M_j, \hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{B}}_j)] - \ln[P(D \mid M_k, \hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{B}}_k)]\right\} + \ln[n](d_k - d_j)\right)/2}} \tag{5}$$

In calculating this risk function, the method assumes a fixed topology. We relax this assumption by including the topology τ_i in the parameter vector $\boldsymbol{\theta}_{\boldsymbol{i}} = (\theta_1, \ldots, \theta_{d_i}, \tau_i)$. The added topology represents one more (complex) parameter to be estimated. The approximation presented in equation (3) changes to

$$\ln[P(D \mid M_i)] \approx \ln\left[P(D \mid M_i, \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{B}}_{\boldsymbol{i}})\right] - \frac{d_i + 2N - 2}{2} \ln[n] = -\frac{BIC_i}{2} \tag{6}$$

Substituting equation (6) in equation (2) results in equation (7):

$$P(M_i \mid D) \approx \frac{1}{\sum_{j=1}^{m} e^{\left(2\left\{\ln[P(D \mid M_j, \hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{B}}_j)] - \ln[P(D \mid M_i, \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{B}}_i)]\right\} + \ln[n](d_i - d_j)\right)/2}} \tag{7}$$

Note that the difference between equations (4) and (7) is the likelihood function; the added parameter will not affect the posterior probabilities.

The loss function is also changed to accommodate the addition of the topology to the parameter vector. We have evaluated two functions for this purpose. The first function fixes the tree tied to the "true" model and uses it to calculate the branch-length vectors for model M_i (that we want to find the risk for) and the "true" model, M_j. The Euclidean distance between these two vectors is the loss and is presented in equation (8):

$$(\|\hat{\boldsymbol{B}}_{\boldsymbol{i}} - \hat{\boldsymbol{B}}_{\boldsymbol{j}}\| \mid \hat{\tau}_i) = \sqrt{\sum_{i=1}^{2N-3} \left[(\hat{\boldsymbol{B}}_{il} - \hat{\boldsymbol{B}}_{jl})^2 \mid \hat{\tau}_j\right]} \tag{8}$$

The risk function associated with this loss function is

$$R_i = \sum_{k=1}^{m} \frac{(\|\hat{\boldsymbol{B}}_{\boldsymbol{i}} - \hat{\boldsymbol{B}}_{\boldsymbol{k}}\| \mid \hat{\tau}_k)}{\sum_{j=1}^{m} e^{(BIC_k - BIC_j)/2}} \tag{9}$$

The second function does not fix the tree when calculating the loss. Hence, each of the models uses its own optimal tree and branch-length vector to evaluate the loss. The loss function in this case is

$$\|\hat{\boldsymbol{B}}_i - \hat{\boldsymbol{B}}_j\| = \sqrt{\sum_{l \in C_{ij}} (\hat{\boldsymbol{B}}_{il} - \hat{\boldsymbol{B}}_{jl})^2 + \sum_{l \notin C_{ij}} (\hat{\boldsymbol{B}}_{il})^2 + \sum_{l \notin C_{ij}} (\hat{\boldsymbol{B}}_{jl})^2} \quad (10)$$

where C_{ij} is the set of branch lengths that belong to both of the optimum topologies resulting under models M_i and M_j. The risk function can be written as in equation (5).

The extensions of the BIC and the AIC approaches are straightforward and are presented in equations (11) and (12):

$$BIC_i = -2\ln[P(D \mid M_i, \hat{\theta}_i)] + (d_i + 2N - 2)\ln[n] \quad (11)$$

$$AIC_i = -2\ln[P(D \mid M_i, \hat{\theta}_i)] + 2(d_i + 2N - 2) \quad (12)$$

with $\hat{\theta}_i = (\hat{\theta}_1, \ldots, \hat{\theta}_{d_i}, \hat{\tau}_i)$; that is, including the topology. Again, the only significant change in the BIC and AIC is caused by the change in the approximate likelihood; the change in the number of parameters will result in a constant shift, affecting all models and, hence, canceling out.

Methods

Choosing the Best Model Using the New Extensions

In our original description of the DT approach (Minin et al. 2003), we explored conditions under which different model-selection methods chose different models. In general, we found that different approaches select different models when the evolutionary process is complex (i.e., simple models are very poor) and when divergences are relatively deep. Therefore, here, we revisit this condition where different approaches to model selection result in different choices (i.e., where model-selection approach makes a difference). Thus, we utilized simulated data associated with the fourth data set used by Minin et al. (2003). This is a rodent cyt *b* data set that includes 22 species of sigmodontine rodents (T. Rinehardt et al., personal communication) downloaded from GenBank (accession numbers AY041185 to AY041206), and it represents a very difficult phylogenetic problem. The substitution process appears to be quite complex in this data set, and it appears that the phylogeny has very short internal branches, with long, but varying terminal branches. As described in Minin et al. (2003), GTR+I+Γ was used to obtain the optimum ML tree. The data were then parsed into three different data sets, each associated with one of the codon positions. The parameters of the GTR+I+Γ model and the branch lengths of the tree were estimated again utilizing these three data sets and using the tree topology introduced in the previous step. Hence, the resulting three trees had the same topology but different branch lengths. Simulated data were generated using Seq-Gen version 1.2.5 (Rambaut and Grassly 1997) based on these three combinations of model and tree. Each run of the simulation gave rise to three data sets, each corresponding to one codon position. These data sets were combined to form one data set. One thousand combined data sets resulted from this simulation. Accordingly, replicate data sets were simulated with a more complex model (30 parameters in the substitution model) than any of those commonly used in phylogentic inference (where the maximum number of parameters is 10).

Implementing the above extensions to the DT approach (loss equations (8) and (10)) involved finding an optimal tree (topology and branch lengths) under each of the 56 substitution models examined by current automated model-selection approaches (e.g., Posada and Crandall 2001) for each replicate data set. To accommodate this, we conducted simultaneous optimizations to search tree space under each of these 56 models as follows. First, for each model, we introduced an initial tree using neighbor-joining with LogDet distances (Lake 1994; Felsenstein 2004). Based on these initial trees and the relevant models, we conducted a heuristic search using a maximum-likelihood criterion and a tree bisection-reconnection (TBR) search strategy. This search resulted in 56 optimal trees and 56 likelihood scores, each associated with one of the models evaluated. To accommodate the loss function of equation (8) we fixed the best topologies ensuing under each of the models and evaluated the score and the branch lengths of these optimal topologies under each of the other models. This caused us to consider 56 trees per model. So, for each data set, we had 56×56 (3,136) trees and scores. These 3,136,000 searches were conducted using PAUP* (Swofford 2001), on a 64-node Beowulf cluster supported by University of Idaho's Initiative in Bioinformatics and Evolutionary Studies (IBEST).

The posterior probabilities of the models, $P(M_j \mid D)$'s (equation (2)), were calculated using the scores of the optimal trees under each model for each data set as shown in equation (7). Calculating the risk for the first extension involved traversing a matrix of distances. Each row of this matrix represents the loss associated with a certain model, whereas each column is associated with a fixed tree. Accordingly, summing the rows of such a matrix, after multiplying each cell by the posterior probability of a model corresponding to a tree, provides the desired posterior risk.

For the second loss function, the process was somewhat less complicated. A matrix was also formed that contained distances between the branch-length vectors; this matrix is symmetric. Summing the rows or the columns after multiplication by the posterior probability will result in the same risk, and this second extension requires less computational effort to accommodate.

The last step was to apply the introduced extensions to the actual data under study. The purpose was to measure the feasibility of implementing such extensions evaluated by the time required for implementation and the resulting chosen model. To do so, we used PAUP* to conduct a heuristic search under the likelihood criterion utilizing a newer Beowulf cluster also supported by IBEST composed of 200 nodes with 2.4 GHz CPUs. The initial trees were built by stepwise addition, with 10 random addition sequences, to make the search less susceptible to local optima. Branch swapping was then done using TBR as above.

Comparing the Different Methods

We compared our new extensions to the DT approach of Minin et al. (2003), as well as to the results of the LRT and the AIC implemented in ModelTest (each based on a single topology). We also compared our results to the results of the BIC, Extended-AIC, and Extended-BIC

implemented in our own software. Further, taking advantage of having estimated optimal trees under each of the models for each of the simulated data, we compared the results of each of these methods to one another.

Two criteria were adopted for these comparisons. First, the chosen models under each of the methods were contrasted with one another. In particular, we calculated the percentage of times we obtained the same model for each of the replicates under each of the methods. The standard error was calculated based on each of these percent matches to give a measure of the simulation error and to highlight the significance in the differences in these methods. We used the distribution of the difference in the number of parameters between the mismatched models to measure the difference in the complexity between them. To test the significance of these differences, we define a "*P* value" to be a measure of the probability of having a parameter difference of zero or more in the mismatched models. In other words, we use the distribution of the differences of the parameters, conditional on having a mismatch for any replicate to find the probability of choosing models with similar or higher complexity when comparing any two of the methods understudy. A very small or very large *P* value indicates that zero is in the tail of the distribution of the differences; hence, there would be a small probability that the complexity of the chosen models is similar. We estimate this probability using the distribution of the differences of the mismatched model parameters.

The second criterion was aimed at comparing the optimal trees under the chosen models using each method. We used symmetric difference distances (Felsenstein 2004; Robinson and Foulds 1981) to compare differences in topology and square root of the squared distance (Felsenstein 2004; Kuhner and Felsenstein 1994) to compare topology and branch lengths. Absolute distances were measured from the true tree to evaluate the ability of the different methods to choose a model that can retrieve the true tree or trees close to the truth. Relative distances were measured between the pairs of optimal trees chosen under any two methods to evaluate the closeness of the performance of the different methods. The standard error was calculated based on the percent matches associated with these measures, as well, to quantify the simulation error. We also calculated a measure of the bias of branch-length estimation. This is a signed difference between the branch-length vector of an optimal tree under a model and the branch-length vector of the true tree. The aim of this measure was to assess whether the resulting trees tend to underestimate or overestimate the true branch lengths.

Results

Comparing the Two Extensions

The results from the two extended DT approaches introduced in equations (8) and (10) were identical. This is a direct consequence of the similarity between the optimal tree topologies associated with the most-probable models. Poor models have a very small probability and, hence, are in essence dropped from consideration (very small probability compared with the distance weighting it and very small probability compared with the other more probable models). Assuming 10^{-6} is a small probability with no significant contribution, then, on average, only about 5.3 models contributed to the calculation of the risk and, therefore, to selecting the best model per data set. The main competing models are listed in table 1, along with the number of times they were considered (out of 1,000 simulations). It is clear that the most probable models are the complex ones.

Table 1
Models with Posterior Probabilities Greater Than 10^{-6} That Had Main Impact on Model Choice Using the Extended-DT and the Number of Times They Were Considered in 1,000 Simulations

Model	Number of Times Considered
HKY+I+Γ	853
TrN+I+Γ	948
K81uf+I+Γ	804
TIM+I+Γ	915
TVM+I+Γ	738
GTR+Γ	24
GTR+I+Γ	989

According to the above, we refer to these two extensions as the Extended-DT method without distinction. Because the time needed to perform a comparison using the first method is drastically more than that using the second method, we focus on the second extension.

Model Comparison

In general, it is expected that the Extended-DT approach would choose more complex models, on average, than would DT of Minin et al. (2003). This is because of the improvement in fitness resulting from factoring the topology in the optimization process. It is also expected that the LRT approach (LRT-ModelTest) and the AIC (simple and extended) would pick more complex models than would DT, BIC, Extended-DT, or Extended-BIC. This is because of the complexity of the process we used in generating the data, with 30 parameters in the generating model; AIC penalizes overparameterization less than BIC and DT by a factor of $\ln(n)/2$ ($\sim$3.3 in the case of this data set), where n is the sample size (number of nucleotides).

Extended-DT Versus Other Methods

Extended-DT and DT chose the same models almost 88% of the time (table 2). Comparing the mismatched models, figure 1*a* indicates that DT chose slightly simpler models on average (measured by the difference in the number of parameters in a model), in accordance with our expectations. The distribution of the differences of these models is somewhat symmetric around the mean; hence, there is no reason to believe that these differences are not random. Table 2 illustrates that there is also no reason to believe that the expected difference is not equal to zero, with the estimate of the probability of occurrence of zero or more in the difference of the mismatched models

Table 2
Match Proportions Between Models Chosen Using Extended-DT and Those Chosen Using the Other Methods

		Match Proportion[a]	*P* value
Extended-DT	DT	87.9% (1.03%)	0.5785
	Extended-BIC	95.4% (0.66%)	0.6087
	Extended-AIC	45.8% (3.2%)	0.0037
	LRT-ModelTest	37.6% (1.58%)	0.0000
	BIC	87.4% (1.05%)	0.5238
	AIC	46.1% (1.58%)	0.0111

NOTE.—Included is the proportion of the mismatched models, chosen by Extended-DT and by the other models, that had differences in the number of parameters greater than zero (*P* value)

[a] Number in parenthesis is equal to the standard error of the simulation.

(*P* value) being 0.5785. All newly selected models agreed on the importance of the rate heterogeneity. The disparity between the chosen models for any data set was related, for the most part, to the transition/transversion rates. These rates were either consolidated or expanded, which explains the clustering of the distribution of the differences on the odd numbers within the range [–4, 4] in the figure. Table 3 lists a detailed comparison of the mismatched models. The main contributors to the mismatches were the changes from HKY+I+Γ to TrN+I+Γ and visa versa. HKY+I+Γ assumes one transition rate, whereas TrN +I+Γ assumes a different rate for purines than for pyrimidines. Also contributing were the switches between TrN+I+Γ and GTR+I+Γ. These changes were caused by the collapse or expansion of the transversion rates (GTR+I+Γ assumes four trasversion rates, whereas TrN+I+Γ assumes one).

Table 2 illustrates the similarity of the results between Extended-DT and Extended-BIC, with almost 96% match and 0.6087 *P* value. This emphasizes the importance of the posterior probabilities in model choice using decision theory. The posterior probabilities of many of the models were extremely small, with not enough weighting by the loss function to be considered in the analysis. That and the fact that the models being considered had small branch-length differences reduced the impact of the loss function on model choice. Figure 1*b* shows that, on average, Extended-BIC tended to choose slightly more complicated models than did Extended-DT when a mismatch occurred.

BIC's performance was comparable to that of the DT in its similarity to the Extended-DT method with 87.4% match. Figure 1*c* indicates that, unlike DT, BIC tended to choose slightly more complicated models than the Extended-DT, although there is no reason to believe that the two methods are different (*P* value 0.5238 [table 2]).

LRT-ModelTest, Extended-AIC, and AIC all had similar results when compared with the Extended-DT. LRT-ModelTest choice matched that of Extended-DT only 37.6% of the time and chose significantly more complicated models than Extended-DT (with *P* value = 0 [fig. 1*d* and table 2]). The percentages of matches of models chosen using Extended-AIC and AIC to that chosen using Extended-DT were 45.8% and 46.1%, respectively (table

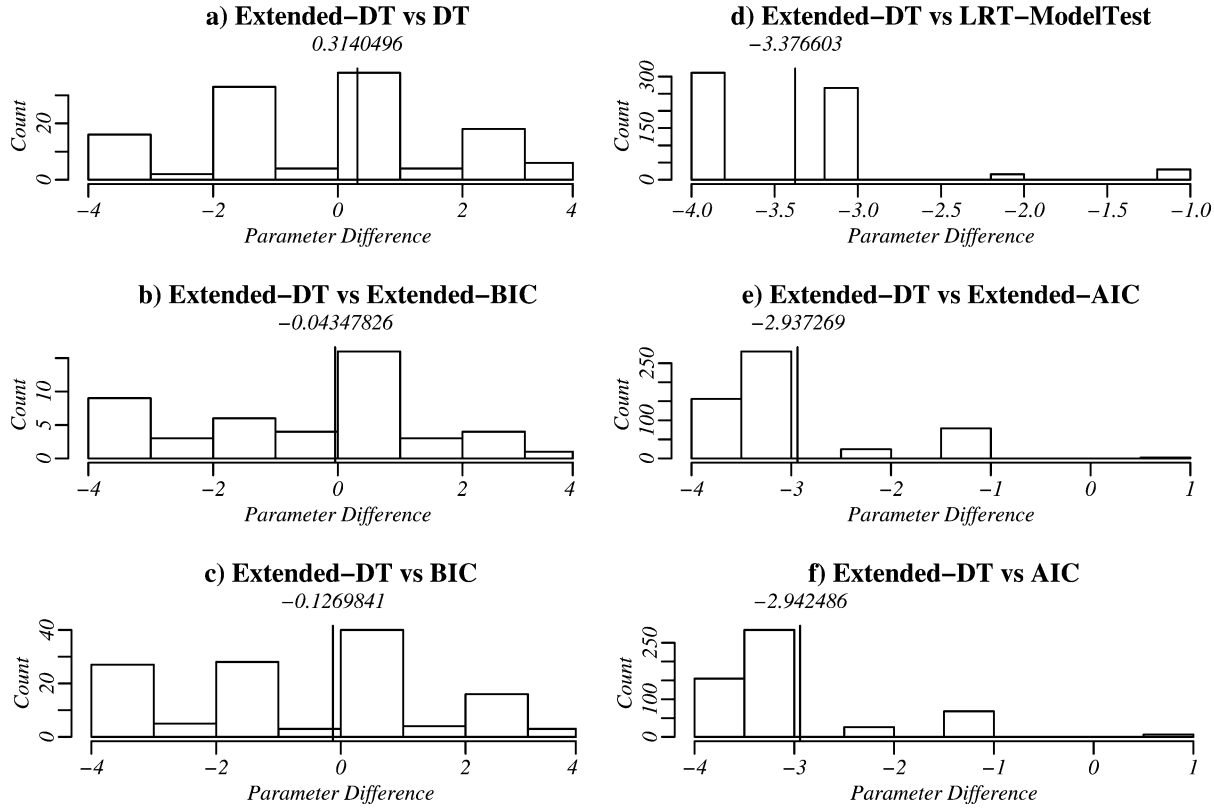

FIG. 1.—Distribution of the differences between model parameters of models chosen using Extended-DT as compared with models chosen using the other methods. These differences exclude the matched models. The vertical lines mark the mean of the distribution.

Table 3
Frequencies of Change Between the Different Models Chosen Using DT and Extended-DT on the Same Simulated Data Sets and the Parameter Differences Between these Models

Count	Difference in Number of Parameters	DT Model	DT Number of Parameters	Extended-DT Model	Extended-DT Number of Parameters
1	−4	GTR+I+Γ	10	HKY+I+Γ	6
1	−3	GTR+I+Γ	10	K81uf+I+Γ	7
1	−3	GTR+I+Γ	10	TIM+Γ	7
12	−3	GTR+I+Γ	10	TrN+I+Γ	7
1	−3	TVM+I+Γ	9	HKY+I+Γ	6
1	−2	GTR+I+Γ	10	TIM+I+Γ	8
1	−2	TVM+I+Γ	9	TrN+I+Γ	7
1	−1	GTR+I+Γ	10	TVM+I+Γ	9
4	−1	K81uf+I+Γ	7	HKY+I+Γ	6
1	−1	TIM+I+Γ	8	K81uf+I+Γ	7
5	−1	TIM+I+Γ	8	TrN+I+Γ	7
22	−1	TrN+I+Γ	7	HKY+I+Γ	6
4	0	TrN+I+Γ	7	K81uf+I+Γ	7
3	1	HKY+I+Γ	6	K81uf+I+Γ	7
27	1	HKY+I+Γ	6	TrN+I+Γ	7
1	1	TIM+I+Γ	8	TVM+I+Γ	6
4	1	TrN+I+Γ	7	TIM+I+Γ	6
3	1	TVM+I+Γ	9	GTR+I+Γ	10
3	2	HKY+I+Γ	6	TIM+I+Γ	8
1	2	TIM+I+Γ	8	GTR+I+Γ	10
3	3	HKY+I+Γ	6	TVM+I+Γ	9
15	3	TrN+I+Γ	7	GTR+I+Γ	10
6	4	HKY+I+Γ	6	GTR+I+Γ	10

NOTE.—This table is based on the 121 mismatched model choices between the two methods.

Table 4
Match Proportions Between Models Chosen Using DT, LRT-ModelTest, BIC, Extended-BIC, AIC, and Extended-AIC

Method 1	Method 2	% Match[a]	(*P* value)
DT	LRT-ModelTest	36.8% (1.15%)	0.0000
DT	BIC	95.9% (0.63%)	0.3171
DT	AIC	45.6% (1.58%)	0.0000
DT	Extended-BIC	88.2% (1.02%)	0.7881
DT	Extended-AIC	43.3% (1.57%)	0.0529
LRT-ModelTest	Extended-BIC	38.2% (1.54%)	0.0000
LRT-ModelTest	Extended-AIC	76.9% (1.33%)	0.0000
LRT-ModelTest	BIC	38.7% (1.54%)	1.0000
LRT-ModelTest	AIC	76.3% (1.34%)	0.9958
BIC	Extended-BIC	90.4% (0.93%)	0.5000
BIC	Extended-AIC	45.2% (1.57%)	0.0164
BIC	AIC	47.7% (1.58%)	0.0000
Extended-BIC	Extended-AIC	46.8% (1.58%)	0.0376
Extended-BIC	AIC	46.8% (1.58%)	0.0094
Extended-AIC	AIC	92.4% (0.84%)	0.5921

NOTE.—Models are compared to one another and to the proportion of the mismatched models, chosen by these different models, that had differences in their number of parameters greater than zero (*P* value).

[a] Number in parenthesis is equal to the standard error of the simulation.

2). Models chosen via these methods also tended to be significantly more complex than those chosen by the Extended-DT method, with *P* values of 0.0037 and 0.0111 for Extended-AIC and AIC, respectively (fig. 1*e* and *f* and table 2).

The simulation error associated with each of the match percentages emphasizes the similarity between the DT and the BIC and the AIC and Extended-AIC in their performance as compared with the Extended-DT. It also highlights the significance in the differences between these methods and the LRT and the Extended-BIC as compared with the Extended-DT (table 2).

Comparing All Other Methods

DT, BIC, and Extended-BIC did perform in a similar manner, with 95.6% match between models chosen using DT and models chosen using BIC, 88.2% match between models chosen using DT and models chosen using Extended-BIC, and 90.4% match between models chosen using BIC and models chosen using Extended-BIC (table 4). Our results suggest that there are no significant differences between DT and BIC. Between DT and Extended-BIC and between BIC and Extended-BIC with *p* values of 0.3171, 0.7881, and 0.5, respectively (table 4). These results highlight the close performance between DT and BIC, demonstrating again that under these conditions, model selection under DT is driven to a great extent by the posterior probability of the model. It is clear from figure 2*a* and *b* that when a mismatch occurs, DT tends to choose slightly simpler models, on average, than those chosen by the Extended-BIC and the BIC methods.

LRT-ModelTest, AIC, and Extended-AIC performed in a comparatively similar manner as well, with a 76.3% agreement between the resulting models from AIC and LRT, 76.9% match between Extended-AIC and LRT, and 92.4% match between models resulting from Extended-AIC and AIC. There was not much difference in performance between AIC and the Extended-AIC. On the other hand, the difference between the models resulting from LRT-ModelTest compared with those from Extended-AIC and AIC when a mismatch occurred was quite significant, with the zero having small probability of being within the distribution (*P* value = 0.9958, when compared with AIC, and *P* value = 0, when compared with Extended-AIC). In addition, figure 2*h* and *i* demonstrates that when a mismatch occurs, AIC and Extended-AIC chose, on average, much simpler models than did LRT.

Figure 2*c*–*e* shows that DT selected much simpler models than did AIC, Extended-AIC, and, especially, LRT-ModelTest. Comparisons between BIC, and Extended-BIC to AIC, Extended-AIC, and LRT also show that BIC and Extended-BIC chose much simpler models than did the others, in concordance with the closeness of these two methods to the DT. This is seen in the remaining portions of the graph and in the percent matches and *P* values listed in table 4.

It is clear from evaluating the simulation error in table 4 that the behavior of AIC and Extended-AIC is quite comparable to one another when contrasted to the other methods of model selection; the match proportions are always within 2 standard errors from one another. The BIC and the Extended-BIC on the other hand compare differently with the DT and Extended-DT, while comparing in a similar manner with the LRT and the AIC (the proportion of matches of the BIC and the Extended-BIC are not significantly different when compared with the

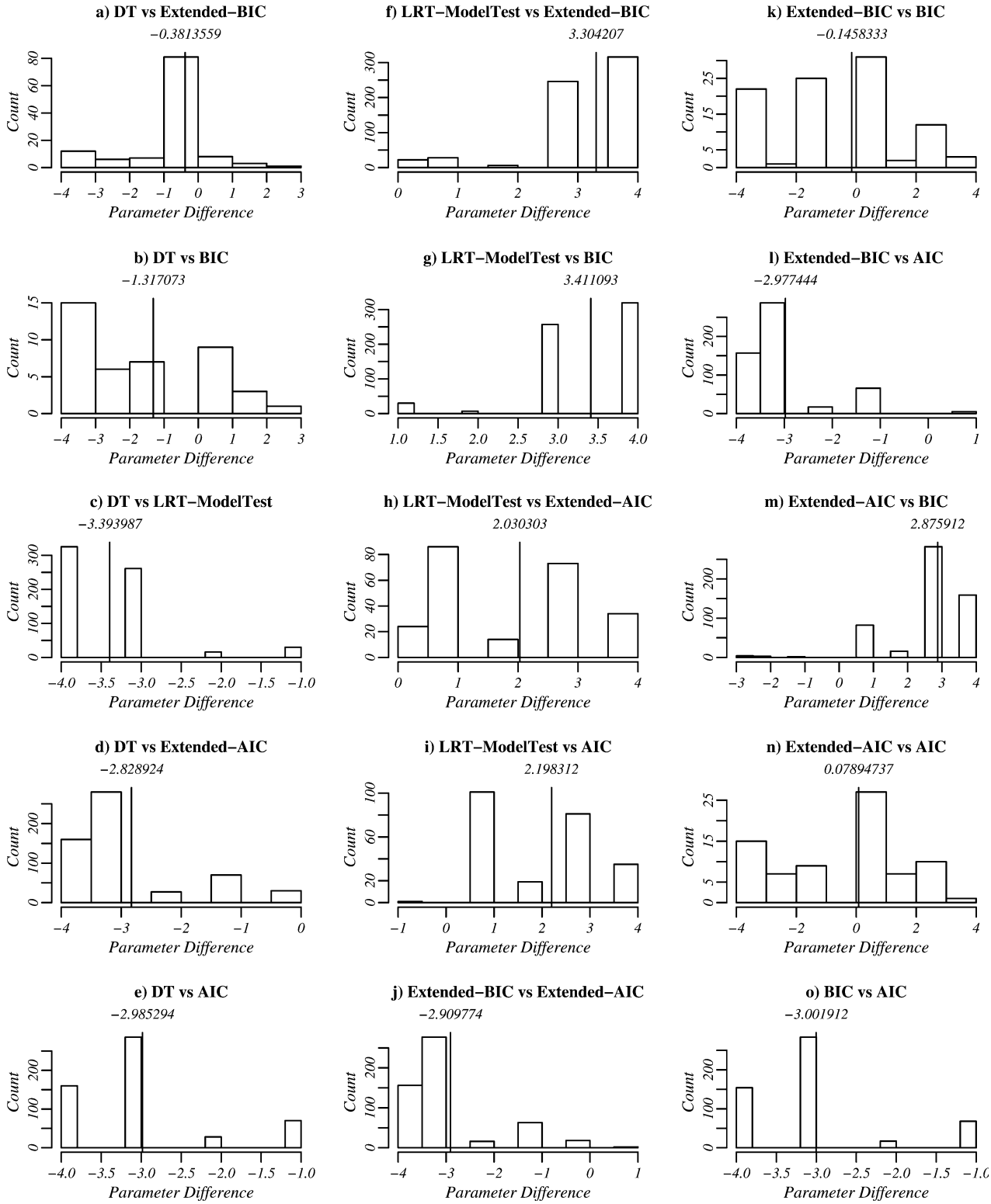

FIG. 2.—Distribution of the differences between model parameters of models chosen using each of the methods (other than Extended-DT) as compared with one another. These differences exclude the matched models and the Extended-DT. The vertical lines mark the mean of the distribution.

LRT and the AIC). The performance of the Extended-BIC compared with Extended-DT (table 2) is similar to that of BIC when compared with DT (table 4), and visa versa, further highlighting the similarity of these methods. DT showed no significant difference when compared with AIC and Extended-AIC. LRT showed a consistent behavior when compared with DT, BIC, and Extended-BIC (table 4); the percent matches where all within 2 standard errors from one another. LRT also showed similarity in behavior when compared with AIC and Extended-AIC (table 4).

Table 5
Comparing the Topologies Resulting Under Models Chosen Using the Different Methods to the True Topology

Method	% Exact Match[a]	Median	Mean	Standard Deviation
Extended-DT	7.2% (0.82%)	6	7.232	5.36
Extended-BIC	7.3% (0.82%)	6	7.295	5.41
Extended-AIC	7.5% (0.83%)	6	7.167	5.31
DT	7.6% (0.84%)	6	7.480	5.55
LRT-ModelTest	7.3% (0.82%)	6	7.291	5.36
BIC	7.4% (0.83%)	6	7.508	5.53
AIC	7.6% (0.84%)	6	7.290	5.43

NOTE.—Comparisons are made via the proportions of the exact match and the median, mean, and standard deviation of symmetric distances

[a] Number in parenthesis is equal to the standard error of the simulation.

Table 6
Comparing the Branch Lengths Resulting Under Models Chosen Using the Different Methods to the True Topology

Method	% Exact Match	Median	Mean	Standard Deviation
Extended-DT	0.0%	6.501	6.462	0.393
Extended-BIC	0.0%	6.504	6.462	0.395
Extended-AIC	0.0%	6.658	6.552	0.405
DT	0.0%	6.502	6.461	0.392
LRT-ModelTest	0.0%	6.686	6.583	0.392
BIC	0.0%	6.506	6.466	0.395
AIC	0.0%	6.661	6.559	0.399

NOTE.—Comparisons are made via the proportions of the exact match and the median, mean, and standard deviation of the square root of the squared distances

Comparing Topologies and Branch Lengths

The above analysis does not describe the comparative performance of the different methods in recovering the closest tree to the truth (i.e., phylogenetic accuracy). Accuracy is measured by the distance between the true and estimated branch-length vectors and by the ability to recover the true topology under the selected model. To evaluate such performance, we used the symmetric difference distance and the square root of the squared distance (Sqrt-Distance) from the true tree. We also measured the symmetric relative distances between topologies resulting from the different methods to assess the agreement between these methods.

Phylogenetic Accuracy

All methods were comparable in finding the true topology, with a slight tilt toward DT and AIC (7.6% exact match for both [table 5]); DT chose much simpler models than did AIC. On average, only 7.4% of the time the true topology was recovered, as can be seen from the symmetric distances. The simulation standard error indicates that there are no significant differences between the proportions of times that the true tree was recovered; each one of these match proportions is within 2 standard errors from the others. Fifty percent of the estimated trees were six or fewer steps away from the true tree, a result that could not have occurred by chance alone (Penny, Foulds, and Hendy, 1982). Extended-AIC chose slightly closer trees to the truth than all other methods, with a mean distance of 7.167 followed by Extended-DT (mean distance of 7.232). AIC (mean distance of 7.29), LRT-ModelTest (mean distance of 7.291) and Extended-BIC (mean distance of 7.295) were very close. DT and BIC came last with mean distances of 7.48 and 7.508, respectively. The standard deviation was smallest for Extended-AIC (5.31) followed by Extended-DT and LRT. It was largest for DT and BIC (5.55 and 5.53, respectively). Given these measures, none of these differences are noteworthy; in effect, the distance from the true tree is about the same for all. Again, we have to emphasize here that Extended-DT, DT, Extended-BIC, and BIC chose much simpler models than did LRT-ModelTest, Extended-AIC, and AIC. Despite the fact that the DT and BIC methods chose simpler models, phylogenies estimated using those models are as accurate as those estimated by the more complex models chosen with the alternative model-selection methods.

Extended-DT and DT achieved a minimum median Sqrt-Distance of 6.501 and 6.502, respectively (table 6). This is quite expected as Sqrt-Distance is used to penalize model selection in Extended-DT and a similar measure is used in DT. This is also clear in comparing the means; although DT did slightly better than Extended-DT with mean distance of 6.461 and 6.462, respectively. Extended-BIC and BIC were close third and fourth, with median distances of 6.504 and 6.506. The means show that Extended-BIC did as well as Extended-DT, on average, in estimating the branch lengths with a mean distance of 6.462. The other methods did not do as well. The worst performer was LRT-ModelTest (median distance 6.686 and mean 6.583), followed by AIC (median 6.661 and mean 6.559). Table 6 introduces the Sqrt-Distance, their medians, means, standard deviations, and the percent match. The standard deviations for the distances are quite comparable; with the minimum attained for DT and LRT and the maximum for Extended-AIC. Figures 3 and 4 show the distribution of the absolute branch-length differences (symmetric and squared, respectively).

Figure 5 shows the signed distribution of the estimate of the branch lengths. It is quite clear from the figure that the estimated branch-length vectors always significantly underestimate the true branch length (The zero does not belong to the distribution and is far away in the lower tail).

Taken together, the results of topological accuracy and accuracy of branch lengths suggest that the insignificant differences in accuracy at recovering nodes (as measured by the symmetric difference distance) dealt with very short internal branches. Conversely, the improved branch-length accuracy of Extended-DT and DT relative to other model selection methods results from better estimates of the lengths of the longer branches. This suggests that nodal support values estimated under the simpler models chosen by the two decision-theory approaches might be more reliable. This will require extensive simulations to assess.

Relative Symmetric Distance

The relative symmetric distance from one tree (chosen by certain method) to another (chosen by another

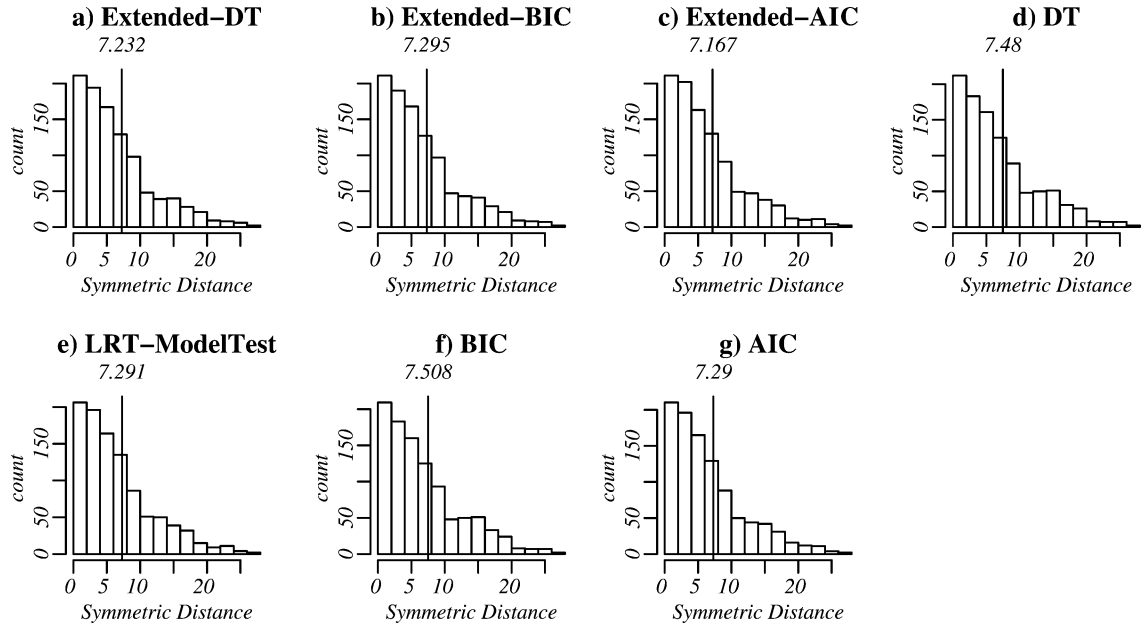

FIG. 3.—Distribution of the symmetric distances measured from the optimal tree resulting under the chosen model using each of the methods to the true tree. Vertical line represents the location of the mean.

method) gives an indication about the variation in the performance of these methods in retrieving the same topology under different models. Table 7 indicates that Extended-DT and Extended-BIC have high correspondence in choosing the topology (97.3% match). This correspondence stands also in respect to DT and BIC with 97.8% topology match. Extended-AIC and AIC matched 95.1% of the time, Extended-BIC and BIC matched 92.5% of the time, and Extended-DT and DT matched 91.2% of the time, indicating that the extensions of these methods did not result in much difference in inferred topology. This suggests that using the simpler methods is sufficient to select well-performing models. Also of interest is that Extended-DT and BIC and Extended-BIC and DT did have a high matching rate in terms of the resulting topologies. This, again, highlights the impact that the posterior probabilities play in choosing the model and the fact that the performance between the extended and simple methods do not result in much difference. The standard errors associated with these mismatch proportions indicate that the simulation error is small, lending more evidence to the above introduced results.

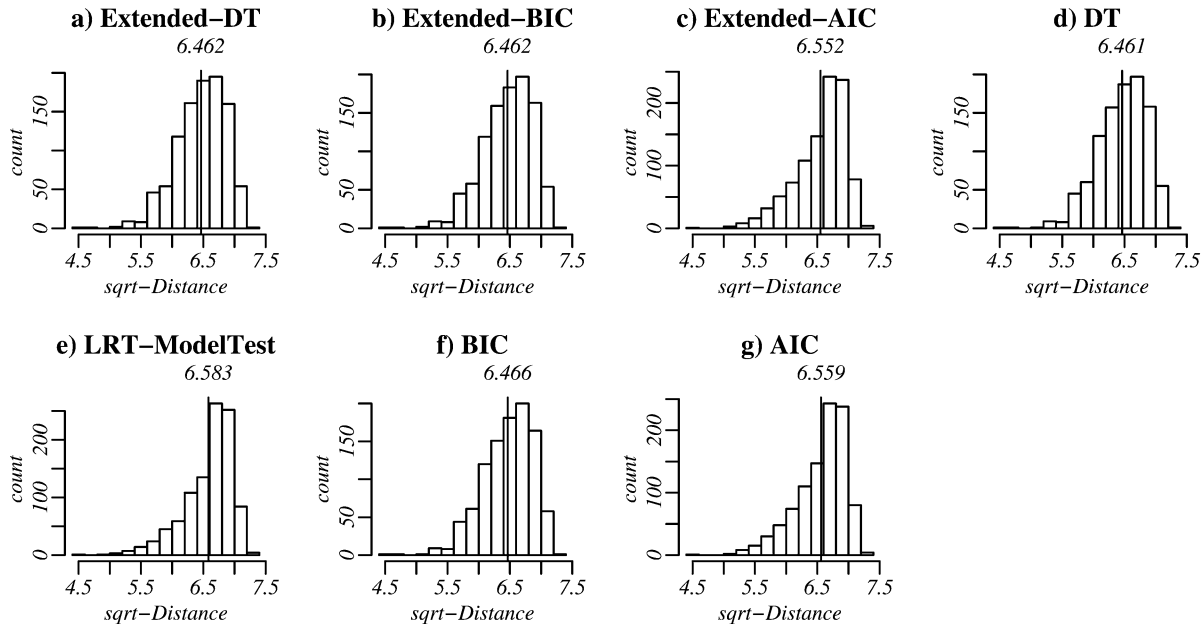

FIG. 4.—Distribution of the square root of the squared distance measured from the optimal tree resulting under the chosen model using each of the methods to the true tree. Vertical line represents the location of the mean.

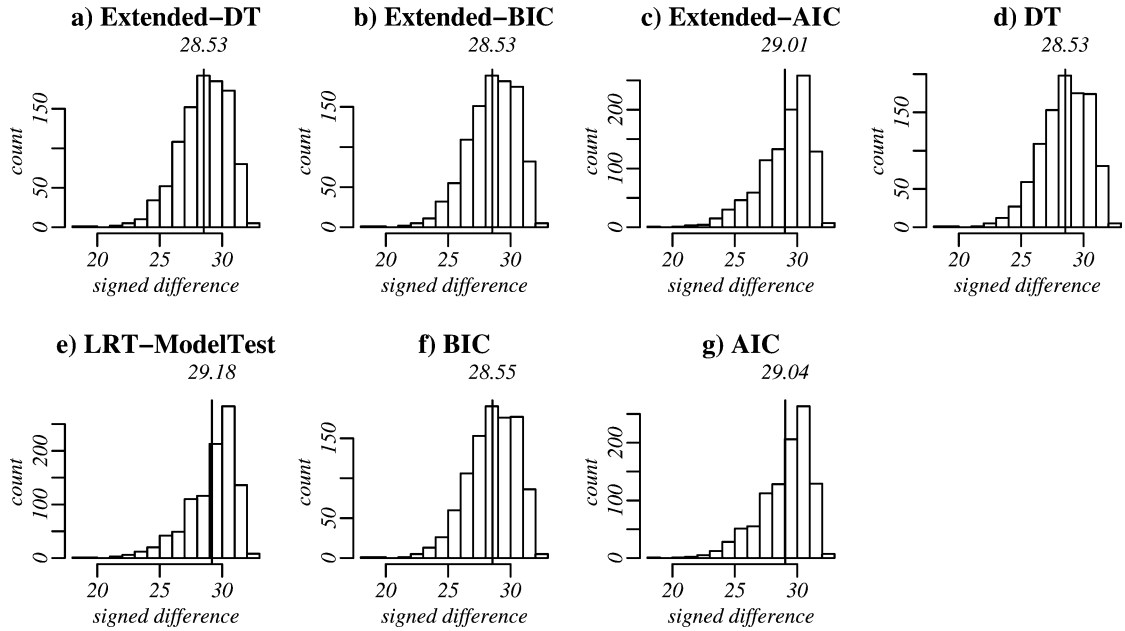

FIG. 5.—Distribution of the signed difference measured from the optimal tree resulting under the chosen model using each of the methods to the true tree. Vertical line represents the location of the mean.

The closest match occurs between the Extended-DT and the Extended-BIC, with mean difference of 0.151 (table 7). This results in the tight distribution of these differences with standard deviation of 1.16, minimum as compared with all other differences. On the other hand, the highest difference is between the Extended-DT and the LRT-ModelTest with mean of 2.541, resulting from a spread distribution with standard deviation of 4.75 (highest among all other differences). The match between these two models is also at a minimum (63.3%). Figure 6 introduces the distribution of the relative symmetric distances. Only mismatched model symmetric distances were plotted. Results of the relative-squared distances are similar to the above and, hence, are not shown.

Time

Finding the optimal trees under each of the models before selecting the best model can be computationally prohibitive, especially as the number of taxa involved increases. Searches for the optimal tree under each of the models took about 5 hours per simulated data set for the second extension of Minin et al. (2003). It took about 22 hours per data set to setup the files before choosing a model using the first extension. These simulations were distributed across the University of Idaho's Beowulf cluster, which contained 64 nodes with 0.9 to 1.2 GHz CPUs. Processing the 1,000 data sets took about 30 days, including cluster down time. For the real data set, a more rigorous search process took 189 days of CPU time for a sequential search under all available models, following the strategy stated in the methods section, on machines with CPU speed of 2.4 GHz. This is equivalent to about 378 days if run on a 1.2-GHz processor. Not having to go through such an extensive search reduces the decision-making time frame to about an hour of initial-tree and score calculations and a few minutes for choosing the model. As shown in the previous section, the resulting model choices were not significantly changed by allowing topology to change during the model-selection process, hence, it is more computationally feasible to maintain the strategy of fixing the tree.

The chosen model for the real data using the new extension was HKY+I+Γ, identical to that chosen by DT, BIC, and the Extended-BIC. The model chosen by LRT-ModelTest was GTR+I+Γ, and that chosen by Extended-AIC and AIC was TVM+I+Γ.

Discussion and Conclusions

In these analyses, we focused on a single condition, derived from a real data set. The performance of phylogenetic methods differs across different true tree shapes (e.g., Sullivan and Swofford 2001) and different true substitution processes (e.g., Gaut and Lewis 1995). Similarly, there are some conditions under which model selection is robust to this approach (Minin et al. 2003), specifically when divergences levels are low or when the substitution process is relatively simple (i.e., several choose the same simple model). However, it is precisely with the difficult-to-estimate tree shapes (i.e., trees with long terminal branches and short internal branches) that model choice becomes particularly critical (Sullivan and Swofford 2003). Therefore, we focused on simulating such conditions and chose a data set that appears to exhibit the

→

FIG. 6.—Distribution of the symmetric distance measured between the optimal trees resulting under the chosen model using each of the methods understudy. These graphs exclude comparisons of the trees resulting from the matched models.

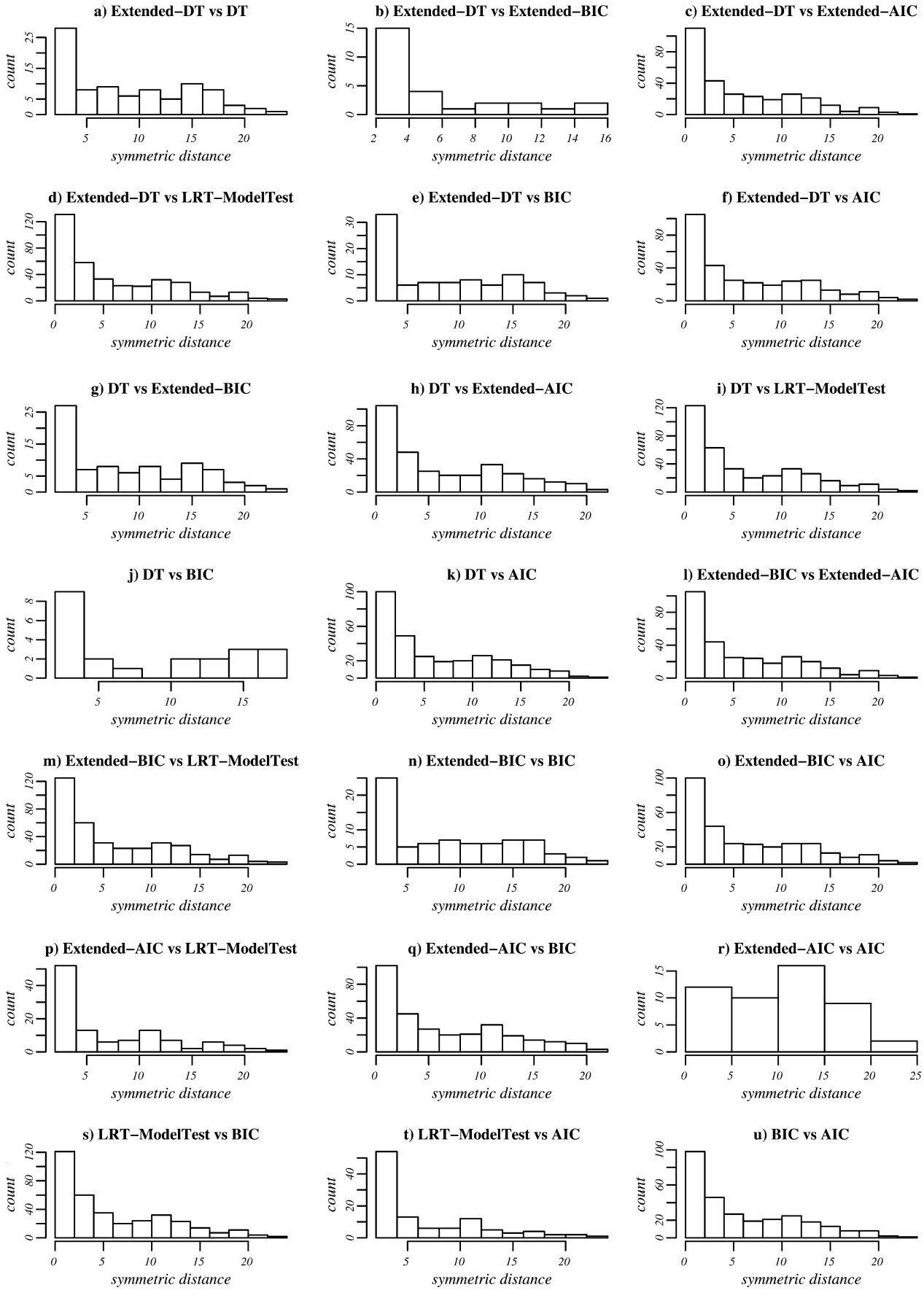
a) Extended−DT vs DT
b) Extended−DT vs Extended−BIC
c) Extended−DT vs Extended−AIC
d) Extended−DT vs LRT−ModelTest
e) Extended−DT vs BIC
f) Extended−DT vs AIC
g) DT vs Extended−BIC
h) DT vs Extended−AIC
i) DT vs LRT−ModelTest
j) DT vs BIC
k) DT vs AIC
l) Extended−BIC vs Extended−AIC
m) Extended−BIC vs LRT−ModelTest
n) Extended−BIC vs BIC
o) Extended−BIC vs AIC
p) Extended−AIC vs LRT−ModelTest
q) Extended−AIC vs BIC
r) Extended−AIC vs AIC
s) LRT−ModelTest vs BIC
t) LRT−ModelTest vs AIC
u) BIC vs AIC
count
symmetric distance

Table 7
Relative Comparison of the Resulting Topologies Under Models Chosen Using the Different Methods (Other Than Extended-DT) via Symmetric Distances

Method 1	Method 2	% Exact Match[a]	Median	Mean	Standard Deviation
Extended-DT	DT	91.2% (0.90%)	0	0.820	3.21
Extended-DT	Extended-BIC	97.3% (0.51%)	0	0.151	1.16
Extended-DT	Extended-AIC	70.3% (1.44%)	0	1.991	4.22
Extended-DT	LRT-ModelTest	63.3% (1.52%)	0	2.541	4.75
Extended-DT	BIC	91.0% (0.90%)	0	0.814	3.18
Extended-DT	AIC	69.9% (1.45%)	0	2.170	4.55
DT	Extended-BIC	91.8% (0.87%)	0	0.763	3.10
DT	Extended-AIC	68.7% (1.47%)	0	2.269	4.57
DT	LRT-ModelTest	63.7% (1.52%)	0	2.529	4.71
DT	BIC	97.8% (0.46%)	0	0.192	1.57
DT	AIC	70.4% (1.44%)	0	2.078	4.37
Extended-BIC	Extended-AIC	70.9% (1.44%)	0	1.962	4.19
Extended-BIC	LRT-ModelTest	63.9% (1.52%)	0	2.522	4.74
Extended-BIC	BIC	92.5% (0.83%)	0	0.717	3.03
Extended-BIC	AIC	70.3% (1.44%)	0	2.159	4.54
Extended-AIC	LRT-ModelTest	88.7% (1.00%)	0	0.842	3.08
Extended-AIC	BIC	69.5% (1.46%)	0	2.193	4.49
Extended-AIC	AIC	95.1% (0.68%)	0	0.525	2.67
LRT-ModelTest	BIC	64.7% (1.51%)	0	2.417	4.60
LRT-ModelTest	AIC	89.2% (0.98%)	0	0.743	2.81
BIC	AIC	71.4% (1.43%)	0	1.966	4.23

NOTE.—Included are proportion of exact match and median, mean, and standard deviation of the symmetric distance differences.
[a] Number in parenthesis is equal to the standard error of the simulation.

problematic features of short internal branches, long terminal branches that vary in length, and a very complex substitution process. Our conclusions are, therefore, restricted to these difficult-to-estimate scenarios.

The two extensions of the DT approach resulted in identical choices of evolutionary models for each of the simulated data sets and for the real data. This is a direct outcome of having only a small number of models with sufficiently high posterior probabilities to have an impact on model choice. Searches under these most probable models yielded trees with similar branch-length vectors and similar topologies. Hence, fixing the tree in the more extensive extension (equation (8)) did not change the loss function enough to make a difference. This lack of difference will result in a computational time save of 17 hours, which is quite significant, compared with the simulated data searches (22 hours), yet not so significant, compared with the search associated with the actual data (378 days).

Extended-DT and DT had an 88% match in model choice. The remaining mismatches did not have a significant pattern. The differences in these mismatches, measured by the difference in their number of parameters, were not significant. Moreover, the symmetric difference distance between the resulting trees when applying the different models were much alike, with 91.2% match in topology and about 88% match in Sqrt-Distance. Accordingly, we do not have sufficient evidence to indicate that these two methods are different. This is also reflected in the identical model choice for the actual data. Based on this, we conclude that adding the topology to the decision criterion will not make any significant difference in selecting a model for data analysis. However, it will add to the computational expense in a way that might make the process of selecting a model computationally prohibitive and quite infeasible.

A surprising result was the closeness of the outcomes of the BIC and DT (extended and simple). This comes back to the fact that only a few models are considered in the choice process because of the small probability of the other models in the set examined. It is clear from table 1 that there is a strong bias toward complex models. This is quite reasonable, as the data were simulated under a more complex model than any of the models we evaluated. Models that did not compensate for rate heterogeneity among sites had very low posterior probabilities (0 in the case when the invariable sites and the rate heterogeneities were not accounted for). This closeness in the results does not undermine the DT approach introduced in Minin et al. (2003); the strength of the DT approach is the flexibility in choosing the loss function that highlights the biological process that one wants to study. Both BIC and the DT results are listed in the output of the DT-ModSel program we developed.

Complex models resulting from the LRT-ModelTest, Extended-AIC, and AIC did not perform better than the simpler models coming from Extended-DT, DT, Extended-BIC, and BIC. This was quite clear in comparing the estimated trees under these models with the true tree. On the other hand, models from the latter keep within the parsimony criterion of model selection (selecting the simplest among a set of equivalent models) and, hence, have less tendency of having higher variation in the parameter estimates than do the more complicated models (Burnham and Anderson 2002). In addition, using the simpler models result in savings in the computational time. There was about 70 hours difference between the search using GTR+I+Γ and that using HKY+I+Γ and about 105 hours difference between the search using TVM+I+Γ and that using HKY+I+Γ. HKY+I+Γ was chosen using DT, Extended-DT, BIC, and Extended-BIC, whereas

GTR+I+Γ and TVM+I+Γ were chosen using LRT-ModelTest, and Extended-AIC and AIC, respectively.

The resulting trees always underestimated the true branch lengths. This suggests that the models considered here need to be improved to be able to capture important features of sequence evolution. An obvious next extension is to include partitioned-likelihood models in the set evaluated.

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 NCCR grant 1P20PR016448-01 (to IBEST). We thank the associate editor A. Von Haeseler for his patient handling of our manuscript and an anonymous reviewer for his helpful comments. We thank T. Reinhardt, 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.W.). We thank K. Blair, our Unix System Administrator, for maintaining the University of Idaho Beowulf cluster and for his assistance and contribution to this project. Special thanks to all the IBEST group, who made this research possible.

Literature Cited

Burnham K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. 2nd edition. Springer-Verlag, New York.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Mass.

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.

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

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

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

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

Penny, D., L. R. Foulds, and M. D. Hendy. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature **297**:197–200.

Posada, D., and K. A. Crandall. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics **14**:817–818.

———. 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.

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.

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

Ardnt von Haeseler, Associate Editor

Accepted November 9, 2004