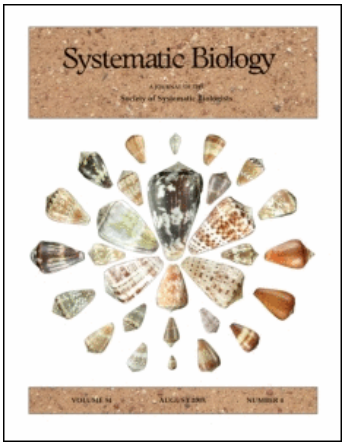

Systematic Biology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713658732

Does Choice in Model Selection Affect Maximum Likelihood Analysis?

Jennifer Ripplinger [a]; Jack Sullivan [ab]
[a] Bioinformatics and Computational Biology, University of Idaho, Moscow, Idaho, USA [b] Department of Biological Sciences, University of Idaho, Moscow, Idaho, USA

First Published on: 01 February 2008

To cite this Article Ripplinger, Jennifer and Sullivan, Jack(2008)'Does Choice in Model Selection Affect Maximum Likelihood Analysis?',Systematic Biology,57:1,76 — 85
To link to this Article: DOI: 10.1080/10635150801898920
URL: http://dx.doi.org/10.1080/10635150801898920

Syst. Biol. 57(1):76–85, 2008

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150801898920

Does Choice in Model Selection Affect Maximum Likelihood Analysis?

JENNIFER RIPPLINGER[1] AND JACK SULLIVAN[1,2]

[1]*Bioinformatics and Computational Biology, and* [2]*Department of Biological Sciences, University of Idaho, Moscow, Idaho 83844-3051, USA; E-mail: jripplinger@vandals.uidaho.edu (J.R.)*

Abstract.—In order to have confidence in model-based phylogenetic analysis, the model of nucleotide substitution adopted must be selected in a statistically rigorous manner. Several model-selection methods are applicable to maximum likelihood (ML) analysis, including the hierarchical likelihood-ratio test (hLRT), Akaike information criterion (AIC), Bayesian information criterion (BIC), and decision theory (DT), but their performance relative to empirical data has not been investigated thoroughly. In this study, we use 250 phylogenetic data sets obtained from TreeBASE to examine the effects that choice in model selection has on ML estimation of phylogeny, with an emphasis on optimal topology, bootstrap support, and hypothesis testing. We show that the use of different methods leads to the selection of two or more models for ~80% of the data sets and that the AIC typically selects more complex models than alternative approaches. Although ML estimation with different best-fit models results in incongruent tree topologies ~50% of the time, these differences are primarily attributable to alternative resolutions of poorly supported nodes. Furthermore, topologies and bootstrap values estimated with ML using alternative statistically supported models are more similar to each other than to topologies and bootstrap values estimated with ML under the Kimura two-parameter (K2P) model or maximum parsimony (MP). In addition, Swofford-Olsen-Waddell-Hillis (SOWH) tests indicate that ML trees estimated with alternative best-fit models are usually not significantly different from each other when evaluated with the same model. However, ML trees estimated with statistically supported models are often significantly suboptimal to ML trees made with the K2P model when both are evaluated with K2P, indicating that not all models perform in an equivalent manner. Nevertheless, the use of alternative statistically supported models generally does not affect tests of monophyletic relationships under either the Shimodaira-Hasegawa (S-H) or SOWH methods. Our results suggest that although choice in model selection has a strong impact on optimal tree topology, it rarely affects evolutionary inferences drawn from the data because differences are mainly confined to poorly supported nodes. Moreover, since ML with alternative best-fit models tends to produce more similar estimates of phylogeny than ML under the K2P model or MP, the use of any statistically based model-selection method is vastly preferable to forgoing the model-selection process altogether. [Akaike information criterion; Bayesian information criterion; decision theory; hypothesis tests; likelihood-ratio test; maximum likelihood; model selection; nonparametric bootstrap.]

Computational methods that utilize an explicit model of sequence evolution have come to dominate phylogenetics. It has been well established that the performance of model-based methods, such as maximum likelihood (ML) and Bayesian estimation, depends on the ability of the chosen model to capture the underlying evolutionary process adequately. If the model ignores particularly important parameters, it will underestimate the magnitude of evolutionary change, which may lead to inconsistent phylogenetic estimation (e.g., Gaut and Lewis, 1995; Huelsenbeck and Hillis, 1993; Sullivan and Swofford, 1997, 2001). Conversely, if an overly complex model is used, the additional parameters will largely capture stochastic signal and the decreased amount of information available for each calculation will lead to increased variation in parameter estimates and, in extreme cases, to nonidentifiable parameters (Rannala, 2002; Lemmon and Moriarty, 2004). Consequently, a model should be used that balances the trade-off between avoidance of bias associated with underparameterized models and increased variance generated by modeling superfluous parameters (Posada and Buckley, 2004; Steel, 2005; Sullivan and Joyce, 2005). Substitution models must therefore be selected in a statistically rigorous manner and several methods have been developed to select models under specific criteria. Four of these methods, the hierarchical likelihood-ratio test (hLRT), Akaike information criterion (AIC), Bayesian information criterion (BIC), and decision theory (DT), are relevant to ML analysis and will be addressed here. For more detailed reviews of these model-selection methods, see Posada and Buckley (2004) and Sullivan and Joyce (2005).

The hLRT was the first statistical method widely applied to phylogenetic model selection (Frati et al., 1997; Huelsenbeck and Crandall, 1997; Posada and Crandall, 1998; Sullivan et al., 1997) and remains the most commonly utilized model-selection method. The hLRT consists of a series of pairwise comparisons between nested models; the process is repeated until the method converges on the simplest model that cannot be rejected at a given significance level. Although the hLRT is simple to implement and may perform well in many cases (e.g., Posada and Crandall, 2001), there are drawbacks associated with its usage, including starting point and path dependence (Cunningham et al., 1998; Pol, 2004), reliance on an arbitrarily selected significance level, and lack of relevant theory to guide the traversal of model space by the requisite series of pair wise comparisons (reviewed by Posada and Buckley, 2004; Sullivan and Joyce, 2005). In addition, the hLRT cannot be used to weight candidate models in order to calculate model-averaged parameter estimates. Therefore, the use of alternative model-selection methods such as the AIC and BIC have been advocated as ways around the path dependence of the hLRT and as ways to incorporate model averaging into phylogenetic estimation (e.g., Alfaro and Huelsenbeck, 2006; Johnson and Omland, 2004; Kelchner and Thomas, 2007; Posada and Buckley, 2004; Sullivan and Joyce, 2005).

The AIC is based on the Kullback-Leibler distance, which measures the information loss associated with fitting a constrained model to the data (Akaike, 1973). AIC scores are a function of both the log-likelihood score, which measures the fit of the model to the data, and a term that penalizes additional parameters. Because the addition of parameters will always increase the likelihood score, the penalty term guards against the selection of overparameterized models. AIC scores are calculated simultaneously for all candidate models and the model with the lowest score (i.e., that minimizes the distance to the unconstrained model) is selected as optimal under the AIC. The relative support for each candidate model can be determined by calculating the rescaled (or Δ) AIC, which is the difference between the best AIC score and the score of the model in question. Although there is a correction for the AIC, the AIC_c (Hurvich and Tsai, 1989), which should be used when the ratio of sample size to free parameters is small, quantifying sample size can be problematic (e.g., deciding between sequence length, variable sites, number of taxa, etc.). Because the output of the AIC and AIC_c converge as sample size increases, it has been suggested that the AIC_c be used for all data sets (Burnham and Anderson, 2002, 2004; Posada and Buckley, 2004).

The BIC is used to approximate the model with the maximum posterior probability given the data and uniform priors across candidate models (Schwarz, 1978). The BIC superficially resembles the AIC and is calculated based on the maximized joint (not marginal) log-likelihood and a penalty term that penalizes additional parameters more strongly than the AIC. Consequently, the BIC tends to select simpler models than the AIC (Posada and Crandall, 2001; Abdo et al., 2005). Minin et al. (2003) developed a DT approach to model selection that uses a loss function to minimize expected branch-length error. This approach tends to select the simplest model that provides branch-length estimates similar to the best model under the BIC. DT appears to outperform the hLRT and AIC, even though it tends to select less complex models than both of these methods (Abdo et al., 2005; Minin et al., 2003). In addition, rescaled (Δ) scores can be calculated for candidate models under the BIC and DT in the same manner that they are estimated for the AIC.

Although it is widely known that the use of unsupported models can affect the outcome of phylogenetic analysis (e.g., Kelsey et al., 1999; Sullivan and Swofford, 1997), it is has also been demonstrated that alternative model-selection methods can select different models for the same data and that, in at least some cases, the use of alternative best-fit models (i.e., models selected under different selection criteria) can change the resulting tree topology. Abdo et al. (2005) simulated data from a rodent mtDNA data set and found that the hLRT, AIC, BIC, and DT methods often selected different models for the same replicate. Although use of alternative best-fit models influenced the resulting ML tree topology in many cases, the different model-selection methods did not differ significantly in their ability to recover the correct tree. When the effects of model selection on several hundred real data sets were examined, Lemmon and Moriarty (2004) found that the hLRT and AIC chose different models for 75% of the data sets. However, they did not include the BIC or DT methods in the model-selection tests and did not evaluate the effects of using alternative best-fit models on phylogenetic analysis. Similarly, Pol (2004) examined 18 empirical data sets and found that the AIC and alternative implementations of the hLRT selected different models for 16 data sets. Although use of alternative best-fit models changed the ML tree topology for two (out of eight) data sets, differences among trees were very slight (one or two nearest neighbor interchanges) and primarily due to nodes with low bootstrap support.

Even though it is known that the model-selection methods applicable to ML analysis can select alternative best-fit models, most systematists continue to select models via either the hLRT or AIC method (usually using ModelTest; Posada and Crandall, 1998). It is therefore important to determine how often the hLRT, AIC, BIC, and DT methods select different models for empirical data and assess how often the use of alternative best-fit models affects the outcome of phylogenetic analysis. In this article, we examine the effects of using different model-selection methods in a ML framework and compare the results to those obtained with ML using the Kimura two-parameter (K2P) model (Kimura, 1980), a common default model, as well as the maximum parsimony (MP) method. We first examine how often the four model-selection methods choose different models for the same data and subsequently conduct ML analyses to determine how the use of alternative models influences optimal topology and bootstrap support. We then test whether alternative ML trees are significantly different when evaluated under the same model (i.e., if differences among topologies could be ascribed to phylogenetic uncertainty). We conclude by investigating the effects of model selection on hypothesis tests of monophyletic relationships.

METHODS

Data Collection

In order to obtain a representative sample from the phylogenetic literature, we downloaded 250 prealigned DNA data sets from TreeBASE (http://www.treebase.org). Although the content of TreeBASE is biased towards multicellular eukaryotes and frequently sequenced genes, this bias reflects the large amount of phylogenetic data available for certain plants, animals, and fungi as well as commonly sequenced gene segments. In all, the culled data represented 1 viral, 1 bacterial, 13 unicellular eukaryote, 77 fungal, 73 plant, 83 animal, and 2 combined eukaryote data sets, ranging from species to domain level. The data included 136 nuclear, 49 mitochondrial, 36 chloroplast, and 29 multiple genome data sets, which represented 72 protein, 69 RNA, 13 noncoding, and 96 mixed product gene segments.

We prepared the data sets for analysis by first importing them into PAUP*4.0b10 (Swofford, 2002) and

removing all alignment regions labeled by the original authors as poor or ambiguous. Once these regions were discarded, we removed redundant haplotypes with gaps treated as fifth character states. Consequently, the number of taxa and characters included in our analysis was not always identical to those reported in the primary literature. Even after poor alignment regions and redundant haplotypes had been removed, the data sets exhibited a great deal of diversity. For example, the number of unique haplotypes ranged from 5 to 317 ($\bar{x} = 44.8$), sequence length varied from 256 to 9237 nucleotides (nt) ($\bar{x} = 1651.8$), and maximum p-distance ranged from 1.3% to 76.4% ($\bar{x} = 18.29\%$). Detailed information and citations for each data set, as well as all data collected as part of this study, are provided as online supplementary material (http://SystematicBiology.org).

Model Selection

We began our analysis by selecting best-fit models from among the 56 stationary, reversible Markov models included in the model-selection software Modeltest and DTModSel (Minin et al., 2003; i.e., the common GTR + I + Γ family models). We used PAUP* to calculate ML scores for each candidate model and then used Modeltest to choose best-fit models under the hLRT, AIC_c, and BIC. In order to calculate optimal models with DT, we reanalyzed the data sets in PAUP* and used the DTModSel script (Minin et al., 2003) to identify models with the lowest expected risk. Branch lengths were not included as parameters to be optimized during the model-selection process and sequence length was used a proxy for sample size in the AIC_c, BIC, and DT calculations (see Posada and Buckley [2004] for a discussion of sample size in phylogenetics).

Although we can examine many aspects of model selection by calculating the best model under alternative criteria, this strategy does not allow us to investigate how well these methods differentiate among alternative models. The model ranks and rescaled (Δ) scores calculated as part of AIC_c, BIC, and DT model averaging can be used to assess support for alternative models under each selection criterion. We used ModelTest to obtain model ranks and scores for alternative best-fit models (i.e., those selected using the hLRT, BIC, and DT methods) under the ΔAIC_c as well as for alternative models (i.e., those selected using the hLRT, AIC_c, and DT) under the ΔBIC. In addition, model ranks and delta values for models selected by the hLRT, AIC_c, and BIC were compared under DT using a version of DTModSel that had been modified to produce rescaled DT scores.

Because simulating DNA sequence data using Seq-Gen (Rambaut and Grassly, 1997) precludes inclusion of gaps and ambiguous characters, we removed these characters from each alignment using PAUP*. We then excluded 25 data sets with 50 or fewer remaining characters from further analysis, leaving 225 condensed data sets from which to draw inferences on the effects of model selection. After removing characters with missing data and or/ambiguity, we collapsed redundant haplotypes and repeated model selection as described above.

Phylogenetic Analysis

We conducted ML analyses in PAUP* for all 225 condensed data sets using alternative best-fit models as well as the K2P model. We estimated initial model parameters from a neighbor-joining tree constructed with the LogDet distance correction and used these parameter estimates as starting values for the first ML heuristic search iteration, which was conducted using tree-bisection-reconnection (TBR) branch swapping on 10 random-addition starting trees. We subsequently performed two additional ML iterations while re-optimizing parameter estimates after each iteration (following Sullivan et al., 2005). Three data sets failed to reach an optimal topology after 2 months of continuous heuristic search and were subsequently excluded from the analysis. In addition to ML, we conducted MP analyses for all applicable data sets by first obtaining 100 starting trees via random stepwise addition and carrying out heuristic searches by TBR branch swapping. In order to quantify topological differences among trees, we calculated symmetric-distance differences (SDDs; Robinson and Foulds, 1981) among all pairs of alternative ML trees as well as between ML and MP trees. Because the magnitude of SDD values depends on the number of taxa, we obtained rescaled values by dividing each SDD by the respective number of sequences.

Although SDD values allow us to quantify overall differences among topologies, they do not let us assess support for individual nodes (i.e., support for bipartitions in the data). We conducted nonparametric bootstrap analyses in PAUP* for 40 data sets using ML with alternative models as well as MP to quantify support for both nodes that were present in all trees regardless of method (which we will refer to as invariable nodes) and nodes that were present or absent depending on the model-selection method (variable nodes). For each ML analysis, we used previously optimized model-parameter estimates and calculated 1000 bootstrap replicates using a heuristic search with random-addition starting trees and TBR branch swapping. We carried out MP bootstrap analyses in a similar manner.

Although the use of different model-selection methods may lead to the inference of incongruent phylogenies, it is not known if these differences could be due to uncertainty in estimating the ML tree. We therefore performed a series of Swofford-Olsen-Waddell-Hillis (SOWH) tests (based on the methods of Goldman et al., 2000) to determine whether alternative ML trees were significantly different when evaluated under the same model. We limited our analysis to the most and least complex models selected for 60 data sets as well as the K2P model. For each analysis, we first evaluated the test statistic

$$\delta = \ln L(D|M_A, T_A) - \ln L(D|M_A, T_B),$$

where the first component is the maximum log-likelihood of the data given model A and the ML tree (topology and branch lengths) generated under model A and the second component is the maximum

log-likelihood of the data given model A and the ML tree estimated with model B. The first term of the test statistic had already been calculated; we calculated the second term in PAUP* by conducting an iterative ML search under model A constrained to the tree estimated with model B. Because the generation of a parametric null distribution requires a fully bifurcating topology, we resolved all polytomies by inserting nodes with branch lengths set to zero. We produced the null distribution by simulating two sets of 100 replicates in Seq-Gen, with one set constructed under model A and tree A and the other under model A and tree B. We then imported all simulated data into PAUP* and conducted a single heuristic ML search for each replicate. We calculated the null distribution following the methods outlined in Sullivan (2005), compared the rank of the test statistic to the null distribution, and assessed the test statistic at $\alpha = 0.05$. We then calculated the reciprocal test statistic (i.e., the difference in log-likelihood between the data given model B and tree B versus model B and tree A) and evaluated it against the appropriate null distribution.

Hypothesis Testing

In addition to altering optimal tree topology, the use of alternative model-selection methods has the potential to change statistical inferences drawn from ML hypothesis tests. Consequently, we conducted tests of monophyletic relationships in PAUP* using the Shimodaira-Hasegawa (S-H; Shimodaira and Hasegawa, 1999) and SOWH methods under alternative models to determine whether choice in model selection influenced the outcome of these tests. Tests were conducted for 27 a priori hypotheses (distributed among 6 data sets) posed by the original authors of the data; the scope of hypotheses ranged from intraspecific biogeographic relationships to the evolutionary history of animal phyla. Because we had previously calculated unconstrained ML trees with alternative models, we began each analysis by calculating optimal ML trees constrained to each hypothesis using the iterative search strategy. Although the entire set of possible topologies should be included in S-H tests (Goldman et al., 2000), the test is typically conducted using only the ML tree and one or more alternative topologies (e.g., Bos and Posada, 2005). Consequently, we only included the ML and constraint trees in our analyses and carried out each test with 1000 replicates analyzed via the RELL method. The resulting test statistics were evaluated at $\alpha = 0.05$. For each SOWH test, we used Seq-Gen to simulate 100 replicates on the constraint tree and subsequently performed single heuristic ML searches in PAUP* to find the optimal log-likelihood scores constrained and unconstrained to the given hypothesis.

RESULTS

Model Selection

For the 250 full data sets, we found that the hLRT, AIC_c, BIC, and DT criteria favored the same model 51 times (20.4%), two models 123 times (49.2%), three models 70 times (28.0%), and four models 6 times (2.4%). There was a significant difference in the average number of model parameters selected by each method (one-way ANOVA; d.f. = 3; $P < 0.001$); the AIC_c selected an average of 8.4 ± 1.8 parameters per data set, whereas the hLRT selected 6.9 ± 2.2 parameters, the BIC selected 6.7 ± 2.5, and DT selected 6.7 ± 2.4. Rate heterogeneity parameters were almost always included in selected models, while most variation occurred in the rate matrix; this is consistent with results obtained by Abdo et al. (2005) and Kelchner and Thomas (2007). Although model-selection methods often selected models with a similar number of parameters, they occasionally chose models that differed by up to 9 (out of 10) parameters (Fig. 1). Consequently, there is no guarantee that the use of different approaches will result in the selection of models with similar complexity.

Out of the 199 data sets where different methods selected two or more models, there was a significant difference in the average number of times method pairs selected the same model (one-way ANOVA; d.f. = 5; $P < 0.001$). The BIC/DT methods selected the same model most often while the hLRT/AIC_c methods agreed for only a small number of data sets, (Table 1). These results vary somewhat from the simulation results of Abdo et al. (2005), especially in regards to comparisons involving the AIC; we found that the AIC_c and hLRT selected the same model for 15.6% of our data sets, whereas Abdo et al. (2005) found that they agreed in 76.3% of their replicates. There was also a significant difference in the average number of times each method selected the most and least parameter-rich models for each data set (one-way ANOVA; d.f. = 3; $P < 0.001$ in both instances). The AIC_c (rather than the hLRT) selected the most complex models for the majority of data sets, whereas the BIC selected the least parameter-rich model most often

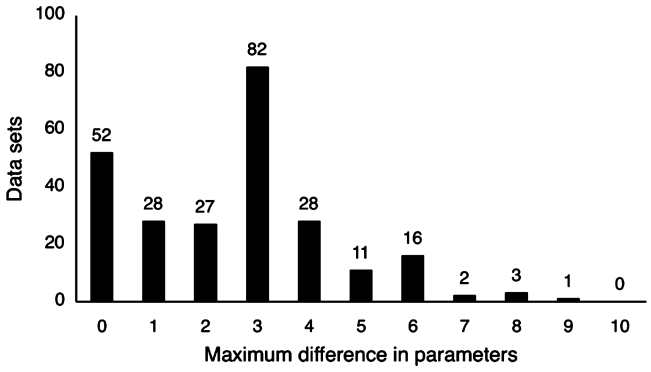

FIGURE 1. Maximum difference in the number of model parameters selected by the hierarchical likelihood-ratio test (hLRT), corrected Akaike information criterion (AIC_c), Bayesian information criterion (BIC), and decision theory (DT) methods for 250 data sets with gaps and ambiguous characters included. The frequencies of alternative best-fit models with different numbers of parameters are listed above each bar. Although model-selection methods often selected models with a similar number of parameters, model complexity varied by as much as 9 (out of 10) parameters. The large number of data sets with alternative models that differed by three parameters is due to models that parameterized base frequencies as either equal (no free parameters) or unequal (three free parameters).

TABLE 1. The frequencies with which the hierarchical likelihood-ratio test (hLRT), corrected Akaike information criterion (AIC_c), Bayesian information criterion (BIC), and decision theory (DT) methods selected the same models. Calculations were based on 199 data sets where methods had selected two or more models; the convergence frequencies among method pairs differed from those of Abdo et al. (2005). *P*-values were calculated based on a binomial distribution; the number of matches significantly deviated from random for each method pair.

Comparison	Matches	*P*-values
hLRT & AIC_c	31 (15.6%)	<0.001
hLRT & BIC	65 (32.7%)	<0.001
hLRT & DT	59 (29.6%)	<0.001
AIC_c & BIC	44 (22.1%)	<0.001
AIC_c & DT	42 (21.1%)	<0.001
BIC & DT	172 (86.4%)	<0.001

(Table 2). In addition, the hLRT did not select the most complex GTR+I+Γ model for most of the data sets, as was observed by Minin et al. (2003). The AIC_c selected a more complex model than the hLRT, BIC, and DT for the majority of data sets, whereas it rarely chose a less complex model. The hLRT normally selected less complex models than the AIC_c and performed fairly similar to the BIC and DT methods. The BIC and DT also tended to select less complex models than the AIC_c and virtually always selected the same models as each other (Table 3).

We found that model-selection methods tended to select a greater number of models and more divergent parameters for small data sets (i.e., those with relatively short sequence length or a small number of taxa), but the correlation did not explain much of the variation in the data. There was a weak negative correlation between the number of haplotypes and number of selected models (d.f. = 248; $P = 0.004$; $r^2 = 0.033$), sequence length and number of selected models (d.f. = 248; $P < 0.001$; $r^2 = 0.058$), and sequence length and difference in number of parameters (d.f. = 248; $P = 0.025$; $r^2 = 0.020$). There was a suggestion of a relationship between the number of sequences and difference in parameters (d.f. = 248; $P = 0.073$), but there was no discernable relationship between maximum p-distance and either the number of selected models (d.f. = 248; $P = 0.863$) or spread in parameters (d.f. = 248; $P = 0.652$).

Although there were some discernable patterns in model weighting, the poor correlation between model rank and delta values made it impossible to interpret strictly the degree to which the rescaled AIC_c, BIC, and DT criteria supported models chosen by other methods. The majority of models selected by the hLRT, BIC, and DT appeared to have some support under the rescaled AIC_c, following the guidelines suggested by Burnham and Anderson (2002, 2004) for interpreting rescaled AIC scores (i.e., they had $\Delta\ AIC_c < 10$). There was no significant difference across model-selection methods in the number of strongly supported alternative models ($\Delta \leq 2$) under the AIC_c; however, a difference in the number of supported models emerged as the delta values increased (Table 4). The guidelines suggested by Burnham and Anderson (2002, 2004) provide an approximate framework for interpreting rescaled AIC values; however, these guidelines may not hold when the data are not independently distributed (Burnham and Anderson, 2002). Although the AIC_c provided similar support for the hLRT, BIC, and DT model ranks, the rescaled AIC_c provided substantially less support for the models chosen by the hLRT than by the BIC or DT. Overall, it appears that the ΔAIC_c, ΔBIC, and ΔDT methods all provide the least support for models selected via the hLRT and that ΔDT provides fairly strong support for models selected under the BIC (Table 5).

TABLE 2. The rates with which each model-selection method selected the most and least complex models for each data set. Calculations were based on 195 data sets with multiple supported models that varied in number of parameters. The hLRT did not select the most parameter-rich GTR+I+Γ model for the majority of data sets and normally favored a less complex model than the AIC_c.

Method	Most complex	Least complex
hLRT	48 (24.6%)	117 (60.0%)
AIC_c	182 (93.3%)	4 (2.1%)
BIC	36 (18.5%)	147 (75.4%)
DT	40 (20.5%)	135 (69.2%)

TABLE 3. The frequencies with which each model-selection method selected more (A > B), less (A < B), or the same (A = B) number of parameters as alternative methods. Counts were made based on 199 data sets where methods had selected two or more models. The AIC_c normally selected more complex models than alternative methods.

		Number of parameters		
Method A	Method B	A > B	A = B	A < B
hLRT	AIC_c	13	40	146
hLRT	BIC	74	78	47
hLRT	DT	74	71	54
AIC_c	BIC	155	44	0
AIC_c	DT	152	46	1
BIC	DT	4	180	15

Phylogenetic Analysis

The use of alternative best-fit models for phylogeny estimation resulted in different ML topologies for 93 of the 188 condensed data sets where methods had selected more than one model (~50% of the data sets). The hLRT and AIC_c selected different models the most often and

TABLE 4. ΔAIC_c values for models selected by the alternative hLRT, BIC, and DT methods. Only models with $\Delta AIC_c > 0$ were included. The guidelines proposed by Burnham and Anderson (2002, 2004) suggest that models with $0 < \Delta \leq 2$ are strongly supported by the AIC, those with $4 \leq \Delta \leq 7$ have moderate support, and models with $\Delta > 10$ have almost no support. These results suggest that alternative models selected by the hLRT, BIC, and DT are normally not strongly supported by the AIC_c.

	ΔAIC_c scores				
Method	0–2	2–4	4–7	7–10	>10
hLRT	37	24	42	17	71
BIC	35	24	31	32	34
DT	40	25	27	33	32

TABLE 5. The average rank and rescaled (Δ) scores for optimal models selected by the hLRT, AIC_c, BIC, and DT methods when evaluated under alternative criteria. Although guidelines for interpreting rescaled BIC and DT scores are not available, the results suggest that best-fit models under one criterion are not necessarily strongly supported by an alternative method.

	Support measure (Δ)		
Selection method	AIC_c	BIC	DT
hLRT	6.44/19.95	4.84/19.02	10.55 0.14
AIC_c	—	6.04/7.7	10.58/0.10
BIC	6.65/7.29	—	8.08/0.00
DT	6.35/9.38	3.59/14.92	—

led to the largest proportion of divergent topologies. The BIC and DT selected the same model most often and when they did select different models, use of these models often produced the same tree topology. Furthermore, default use of the K2P model instead of a model-selection method changed the resulting ML topology for ~72% of the data sets, whereas use of MP leads to different optimal trees for almost all of the data sets (~90%). Rescaled SDD values indicate that ML trees estimated with alternative best-fit models were much more similar to each other than to ML trees estimated with the K2P model or MP trees (Table 6).

Nonparametric bootstrap analyses indicate that variable nodes tended to be weakly supported, while invariable nodes had fairly constant bootstrap support across trees estimated with alternative methods (Fig. 2). All nodes that varied across ML trees estimated with alternative best-fit models had bootstrap values ≤75% ($\bar{x}$ 56.6%), whereas nodes that varied between ML trees estimated with best-fit models and the K2P model had bootstrap values ≤90% ($\bar{x}$ 58.8%) and those that varied between ML and MP trees had bootstrap values ≤95% ($\bar{x}$ 61.5%). Similarly, invariable nodes that were present in all ML trees made with statistically supported models had bootstrap values that differed by ≤19% ($\bar{x}$ 2.0%); nodes that were present in ML trees made with both statistically supported models and the K2P model had bootstrap values that varied by ≤35% ($\bar{x}$ 3.9%), whereas nodes that were present in both ML and MP trees had support values that differed by ≤41% ($\bar{x}$ 5.2%).

The results of the SOWH tests demonstrate that, in most cases, ML trees constructed with alternative best-fit models were not significantly different (Fig. 3). Only one set of ML trees estimated with best-fit models was found to have significant differences. Although it is not clear why this was the only set of significantly different ML trees, it is worth noting that this data set had relatively high divergence (~25% uncorrected p-distance) and that the ML tree topologies differed strongly (standardized SDD 0.60). Although ML trees estimated from best-fit models were normally not significantly different according to the SOWH test, ML trees estimated with the K2P model often differed from ML trees generated with statistically supported models. We found that ML trees estimated with the K2P model were not compatible with trees estimated under best-fit models for 42 of the 60 data sets (70%) when evaluated with the SOWH test. Consequently, our results indicate that not all models behave in an equivalent fashion and that the K2P model lacks parameters that contribute to the shape of alternative ML trees.

Despite these results, the use of alternative model-selection methods, as well as default use of the K2P model, did not seem to influence tests of a priori hypotheses (Table 7). The use of alternative best-fit models did not change the outcome of S-H tests of monophyly, although use of the K2P model led to the rejection of an otherwise supported hypothesis for one data set. Conversely, the use of different best-fit models changed the interpretation of SOWH tests for two hypotheses, whereas use of the K2P model did not change the outcome of any test. As expected, the S-H test produced more conservative results than the SOWH test and choice of test changed the statistical interpretation of eight hypotheses (~30% of the data). It appears that the use of a particular model-selection strategy may be less important than choice of an appropriate test and, consequently, computationally intensive methods such as model-averaged hypothesis testing (Posada and Buckley, 2004) may not be necessary for this set of substitution models (i.e., the GTR + I + Γ family).

TABLE 6. Results of model selection and phylogenetic analysis for condensed data sets. Comparisons between ML trees estimated with alternative model-selection methods were based on 188 data sets with multiple supported models, whereas comparisons between alternative ML trees and both ML trees estimated with the Kimura two-parameter (K2P) model and maximum-parsimony (MP) trees were made based on 222 data sets. Symmetric-distance differences (SDDs) were rescaled by the number of taxa and averaged across all applicable data sets.

Comparison	Matches (models)	Matches (topologies)	Average rescaled SDD
hLRT & AIC_c	24 (12.8%)	117 (62.2%)	0.07
hLRT & BIC	63 (33.5%)	133 (70.7%)	0.09
hLRT & DT	59 (31.4%)	125 (66.5%)	0.08
AIC_c & BIC	28 (14.9%)	119 (63.3%)	0.07
AIC_c & DT	29 (15.4%)	118 (62.8%)	0.07
BIC & DT	156 (83.0%)	177 (94.1%)	0.11
hLRT & K2P	—	65 (29.3%)	0.22
AIC_c & K2P	—	60 (27.0%)	0.22
BIC & K2P	—	64 (28.8%)	0.22
DT & K2P	—	63 (28.4%)	0.22
hLRT & MP	—	20 (9.0%)	0.36
AIC_c & MP	—	22 (9.9%)	0.36
BIC & MP	—	21 (9.5%)	0.36
DT/MP	—	22 (9.9%)	0.36

DISCUSSION

In order to have confidence in the accuracy of model-based methods, one must use a model that adequately parameterizes the data without undue loss of analytical power. Although several model-selection methods exist, empirical users of model-based methods frequently select models via either the hLRT or AIC. We have found that the use of different model-selection methods leads to the selection of alternative models for ~80% of our data sets and that the use of different best-fit models changes

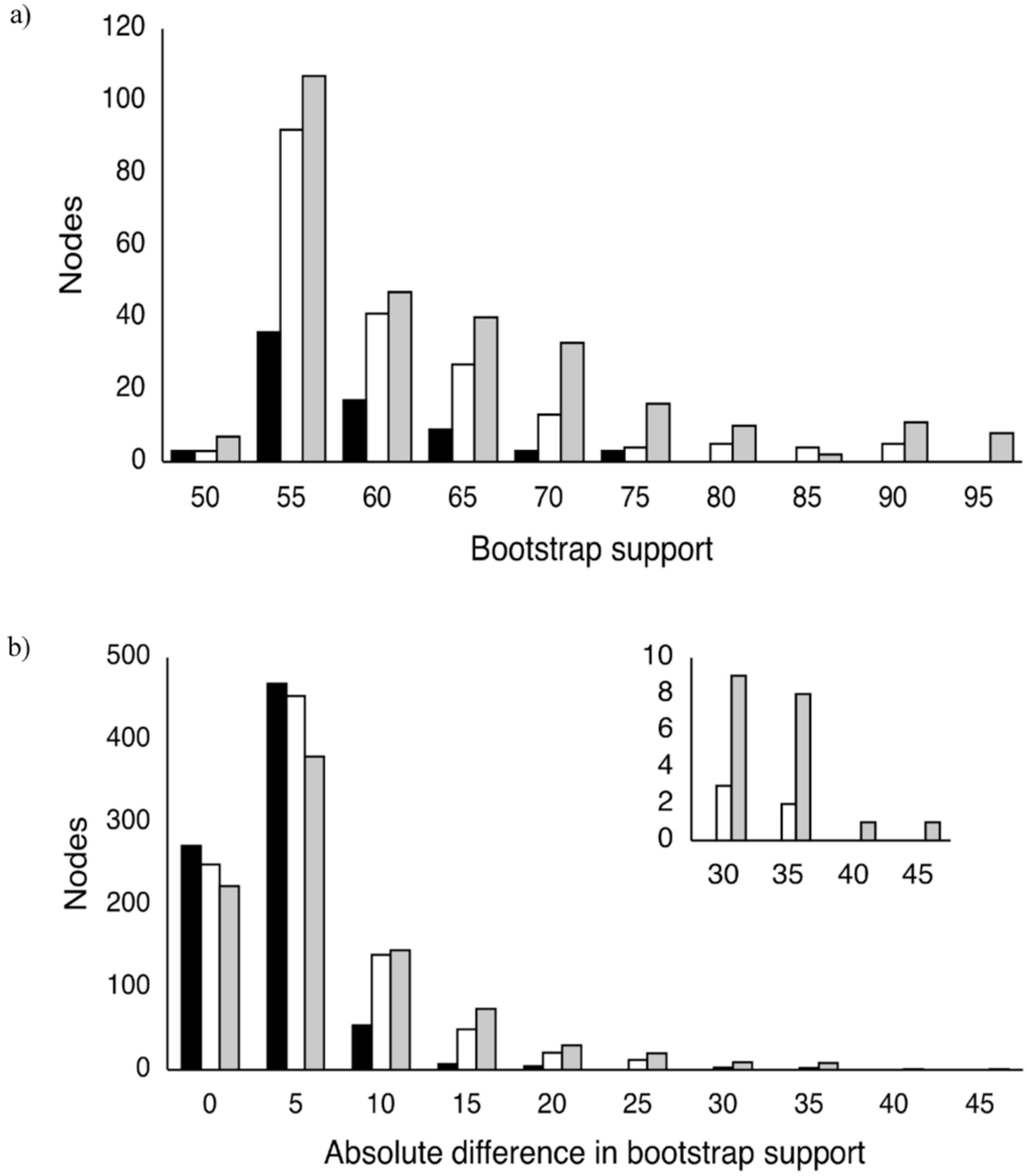

FIGURE 2. The outcome of nonparametric bootstrap analyses conducted to assess support for variable nodes (those that changed across optimal trees estimated with different methods) as well as invariable nodes (those that were present in all trees regardless of method). (a) Nodes that varied among maximum likelihood (ML) trees estimated with statistically supported models (black) had lower bootstrap values ($\bar{x}$ 56.6%) than nodes that varied between ML trees made with best-fit models and either ML trees estimated with the Kimura two-parameter (K2P) model (white) or maximum parsimony (MP) trees (grey; $\bar{x}$ 58.8% and 61.5%, respectively). (b) Nodes that were invariable among ML trees estimated with statistically supported models (black) tended to have similar bootstrap values across trees made with alternative best-fit models ($\bar{x}$ 2.0%). The difference in bootstrap support values was higher when ML trees estimated with supported models were compared to ML trees estimated with the K2P model (white) or MP trees (grey; $\bar{x}$ 3.9% and 5.2%, respectively).

the optimum tree topology in ∼50% of cases. Because our data include a wide range of phylogenetic data sets, this result is probably very general. Consequently, many researchers who have used the hLRT or AIC to select a model for their analyses may have selected a different model and generated an alternative ML tree if they had used a different model-selection method. However, ML trees calculated with alternative best-fit models are normally not significantly different from each other and, consequently, use of any statistically supported model may provide a statistically equivalent estimate of the phylogeny.

The results of this study are similar to those obtained by Abdo et al. (2005), who found that the BIC

TABLE 7. Results of a priori tests of monophyletic relationships evaluated with both the Shimodaira-Hasegawa (S-H) and Swofford-Olsen-Waddell-Hillis (SOWH) tests. Twenty-seven hypotheses were divided among six diverse data sets; full hypotheses and references are given in the Supplementary Material. Use of alternative model-selection methods had little influence on the outcome of hypothesis tests.

Best-fit models	S-H test	SOWH test	Best-fit models	S-H test	SOWH test
			Hypotheses 16–18		
			HKY+I+Γ	$P < 0.01$	$P < 0.01$
Hypothesis 1			K2P+I+Γ	$P < 0.01$	$P < 0.01$
GTR+I+Γ	$P = 0.45$	$P = 0.08$	K2P	$P < 0.01$	$P < 0.01$
TrN+I+Γ	$P = 0.42$	$P = 0.01$	Hypothesis 19		
F81+I+Γ	$P = 0.42$	$P < 0.01$	HKY+I+Γ	$P < 0.01$	$P < 0.01$
K2P	$P = 0.34$	$P < 0.01$	K2P+I+Γ	$P < 0.01$	$P < 0.01$
Hypothesis 2			K2P	$P = 0.05$	$P < 0.01$
GTR+I+Γ	$P = 0.46$	$P = 0.03$	Hypotheses 20, 21		
TrN+I+Γ	$P = 0.45$	$P = 0.05$	HKY+I+Γ	$P < 0.01$	$P < 0.01$
F81+I+Γ	$P = 0.47$	$P = 0.11$	K2P+I+Γ	$P < 0.01$	$P < 0.01$
K2P	$P = 0.22$	$P < 0.01$	K2P	$P < 0.01$	$P < 0.01$
Hypothesis 3			Hypothesis 22		
GTR+I+Γ	$P = 0.25$	$P < 0.01$	GTR+I	$P = 0.16$	$P = 0.44$
TrN+I+Γ	$P = 0.25$	$P < 0.01$	HKY+Γ	$P = 0.24$	$P = 0.61$
F81+I+Γ	$P = 0.25$	$P < 0.01$	HKY+I	$P = 0.25$	$P = 0.31$
K2P	$P = 0.07$	$P < 0.01$	K2P	$P = 0.26$	$P = 0.31$
Hypotheses 4, 5			Hypothesis 23		
GTR+Γ	$P > 0.99$	$P > 0.99$	GTR+I	$P = 0.09$	$P < 0.01$
TVM+Γ	$P > 0.99$	$P > 0.99$	HKY+Γ	$P = 0.09, 0.08$	$P = 0.04$
K2P	$P > 0.99$	$P > 0.99$	HKY+I	$P = 0.09, 0.08$	$P < 0.01$
Hypothesis 6			K2P	$P = 0.09$	$P < 0.01$
GTR+Γ	$P = 0.29$	$P = 0.03$	Hypothesis 24		
TVM+Γ	$P = 0.30$	$P = 0.03$	GTR+I+Γ	$P < 0.01$	$P < 0.01$
K2P	$P = 0.28$	$P < 0.01$	TIM+I+Γ	$P < 0.01$	$P < 0.01$
Hypotheses 7–12			K2P	$P < 0.01$	$P < 0.01$
GTR+Γ	$P > 0.99$	$P > 0.99$	Hypothesis 25		
TVM+Γ	$P > 0.99$	$P > 0.99$	GTR+I+Γ	$P = 0.02$	$P < 0.01$
K2P	$P > 0.99$	$P > 0.99$	TIM+I+Γ	$P = 0.02$	$P < 0.01$
Hypotheses 13, 14			K2P	$P = 0.02$	$P < 0.01$
HKY+I+Γ	$P < 0.01$	$P < 0.01$	Hypothesis 26		
K2P+I+Γ	$P < 0.01$	$P < 0.01$	GTR+I+Γ	$P = 0.28$	$P < 0.01$
K2P	$P < 0.01$	$P < 0.01$	TrN+I+Γ	$P = 0.28$	$P < 0.01$
Hypothesis 15			K2P	$P = 0.08$	$P < 0.01$
HKY+I+Γ	$P = 0.18$	$P < 0.01$	Hypothesis 27		
K2P+I+Γ	$P = 0.13$	$P < 0.01$	GTR+I+Γ	$P = 0.11$	$P < 0.01$
K2P	$P = 0.32$	$P < 0.01$	TrN+I+Γ	$P = 0.12$	$P < 0.01$
			K2P	$P = 0.02$	$P < 0.01$

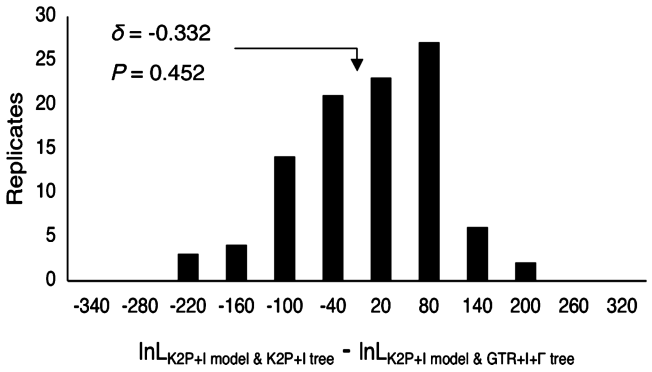

FIGURE 3. An example of a Swofford-Olsen-Waddell-Hillis (SOWH) test conducted to assess whether ML trees generated with different models are significantly different when evaluated under the same model. In this example, the AIC_c had selected the GTR+I+Γ model (10 free parameters) whereas the BIC and DT had picked K2P+I (2 free parameters). A SOWH test was used to assess whether the ML tree estimated with the GTR+I+Γ model was significantly suboptimal to the ML tree generated with the K2P+I model when evaluated under the K2P+I model. The test failed to reject the null hypothesis, indicating that although the trees had disparate topologies and branch lengths, they were not significantly different.

and DT consistently selected simpler models than the hLRT and AIC and that these models performed at least as well as more complex alternatives. In addition, both studies found that differences among alternative best-fit models were primarily confined to differences in the transition/transversion rate (or other aspects of the R-matrix). The minor differences between the results of this study and those obtained by Abdo et al. (2005) are due to discrepancies in summary statistics; methods tended to pick the same model and infer the same ML tree more often for the simulated data of Abdo et al. (2005) than for the empirical data used in this study. These differences are expected and most likely due to the simplified conditions used to simulate data.

The hLRT, AIC, BIC, and DT approaches are examples of relative model-selection methods that choose an optimal model regardless of the absolute fit between candidate models and the data. There are currently two methods that evaluate the absolute fit between the model and data: the Goldman-Cox test (Goldman, 1993; Whelan et al., 2001) and posterior predictive simulations (Huelsenbeck et al., 2001; Bollback, 2002). These methods have not been used as standard model-selection methods

because of their computational complexity but have been used to assess the absolute fit of models selected by other methods (e.g., Sullivan et al. 2000; Althoff et al., 2006; Bos and Posada, 2005; Carstens et al., 2004, 2005; Foster, 2004). Despite situations where the Goldman-Cox test has failed to reject models as inadequate (e.g., Demboski and Sullivan, 2003), there have also been cases where statistically supported models did not provide a sufficient fit to the data (e.g., Foster, 2004), leading to speculation that the current set of models is inadequate and that there is a need to develop more complex models (e.g., Kelchner and Thomas, 2007; Sanderson and Kim, 2000; Whelan et al., 2001). An evaluation of the set of 56 commonly used substitution models with both the Goldman-Cox test and posterior predictive simulations would help assess the need for more complex models as well as clarify the relationship between model-fit and model-selection methods.

Our results suggest that there may be substantial variance when estimating the optimal ML tree. Although model-selection uncertainty can be accounted for by averaging candidate trees by an objective statistical criterion (Posada and Buckley, 2004), it is not clear how to weight candidate phylogenies. Posada and Buckley (2004) state that trees could be weighted by their respective AIC model score; however, there may not be a direct relationship between a model's AIC score and the variation among resulting ML trees. Consequently, it would be particularly informative to extend our application of the SOWH test to comparisons among ML trees made with all 56 common substitution models. The results could be compared with various tree-weighting schemes, as well as the results of model selection and the Goldman-Cox test, to help determine an appropriate model-selection method. In addition, the results will allow us to determine if default use of the most complex substitution model is a viable alternative to model selection for this set of substitution models.

Although there is much to still be learned about the effects of model selection on phylogenetic analysis, our results provide guidelines for systematists who utilize model-based phylogenetic methods. First, because default use of the K2P model appears to be inadequate, researchers should always use a statistically rigorous model-selection method. Second, because alternative best-fit models appear to produce statistically equivalent estimates of phylogeny, it may be beneficial to use the simplest supported model to reduce the computational burden associated with analysis. Lastly, because it seems that support for some nodes may vary depending on the substitution model, systematists should always assess support for ML trees with the nonparametric bootstrap and carefully evaluate their results.

CONCLUSIONS

Although we have found that the use of the hLRT, AIC_c, BIC, and DT methods often leads to the selection of alternative best-fit models and recovery of different ML trees, the differences among these trees are primarily due to poorly supported nodes and the trees that are optimal under different statistically supported models are usually not significantly different from each other. Furthermore, ML trees constructed with alternative best-fit models are typically more similar to each other than they are to ML trees generated with the K2P model or MP trees. In addition, use of models supported by alternative approaches does not seem to have a large influence on the outcome of hypothesis tests. Therefore, it appears that the approach used for model selection does not affect ML analysis and that the use of any model-selection method is preferable to using an unsupported default model or MP. Moreover, because alternative best-fit models seem to behave in a similar manner, it may be preferable to select the simplest supported model (often selected by the BIC or DT) for ML analysis.

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 and NIH NCCR grant NIH NCCR 1P20PR016448-01 (both to IBEST). We would like to thank Zaid Abdo, Jason Evans, and Paul Joyce for their comments on an earlier version of this manuscript. We would also like to thank Celeste Brown for her help in using the Beowulf clusters and our IBEST systems administrators for their assistance during the data analysis process. Thomas Buckley, Rod Page, and two anonymous reviewers also provided suggestions that helped improve the manuscript.

REFERENCES

Abdo, Z., V. N. Minin, P. Joyce, and J. Sullivan. 2005. Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation. Mol. Biol. Evol. 22:691–703.

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267–281 *in* Proceedings of the Second International Symposium on Information Theory. Akademiai Kiado, Budapest.

Alfaro, M. E., and J. P. Huelsenbeck. 2006. Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. Syst. Biol. 55:89–96.

Althoff, D. M., K. A. Segraves, J. Leebens-Mack, and O. Pellmyr. 2006. Patterns of speciation in the yucca moths: Parallel species radiations within the *Tegeticula yuccasella* species complex. Syst. Biol. 55:398–410.

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

Bos, D. H., and D. Posada. 2005. Using models of nucleotide evolution to build phylogenetic trees. Dev. Comp. Immunol. 29:211–227.

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

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference: Understanding AIC and BIC in model selection. Sociol. Method Res. 33:261–304.

Carstens, B. C., J. D. Degenhardt, A. L. Stevenson, and J. Sullivan. 2005. Accounting for coalescent stochasticity in testing phylogeographical hypotheses: Modeling Pleistocene population structure in the Idaho giant salamander *Dicamptodon aterrimus*. Mol. Ecol. 14:255–265.

Carstens, B. C., A. L. Stevenson, J. D. Degenhardt, and J. Sullivan. 2004. Testing nested phylogenetic and phylogeographic hypotheses in the *Plethodon vandykei* species group. Syst. Biol. 53:781–792.

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

Demboski, J., and J. Sullivan. 2003. Extensive mtDNA variation within the yellow-pine chipmunk, *Tamias amoenus* (Rodentia: Sciuridae),

and phylogeographic inferences for northwest North America. Mol. Phylogenet. Evol. 26:389–408.
Foster, P. G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.
Frati, F., C. Simon, J. Sullivan, and D. L. Swofford. 1997. Evolution of the mitochondrial cytochrome oxidase II 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.
Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652–670.
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 D. M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.
Huelsenbeck, J. P., F. Ronquist, R. Nielson, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.
Hurvich, C. M., and C.-L. Tsai. 1989. Regression and time series model selection in small samples. Biometrika 76:297–307.
Johnson, J. B., and K. S. Omland. 2004. Model selection in ecology and evolution. Trends Ecol. Evol. 19:101–108.
Kelchner, S. A., and M. A. Thomas. 2007. Model use in phylogenetics: Nine key questions. Trends Ecol. Evol. 22:87–94.
Kelsey, C. R., K. A. Crandall, and A. F. Voevodin. 1999. Different models, different trees: The geographic origin of PTLV-I. Mol. Phylogenet. Evol. 13:336–347.
Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions between homologous nucleotide sequences. J. Mol. Evol. 16:111–120.
Lemmon, A. R., and E. C. Moriarty. 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53:265–277.
Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:1–10.
Pol, D. 2004. Empirical problems of the hierarchical likelihood ratio test for model selection. Syst. Biol. 53:949–962.
Posada, D., and T. R. Buckley. 2004. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53:793–808.
Posada, D., and K. A. Crandall. 1998. ModelTest: Testing the model of DNA substitution. Bioinformatics 14:817–818.
Posada, D., and K. A. Crandall. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50:580–601.
Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.
Rannala, B. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. Syst. Biol. 51:754–760.
Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
Schwarz, G. 1978. Estimating the dimensions of a model. Ann. Stat. 6:461–464.
Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.
Steel, M. 2005. Should phylogenetic models be trying to fit "fit an elephant"? Trends Genet. 21:307–309.
Sullivan, J. 2005. Maximum-likelihood methods for phylogeny estimation. Methods Enzymol. 395:757–779.
Sullivan, J., Z. Abdo, P. Joyce, and D. L. Swofford. 2005. Evaluating the performance of a successive-approximations approach to parameter optimization in maximum-likelihood phylogeny estimation. Mol. Biol. Evol. 22:1386–1392.
Sullivan, J., E. Arellano, and D. S. Rogers. 2000. Comparative phylogeography of Mesoamerican highland rodents: Concerted versus independent response to past climate fluctuations. Am. Nat. 155:755–768.
Sullivan, J., and P. Joyce. 2005. Model selection in phylogenetics. Ann. Rev. Ecol. Evol. Syst. 36:445–466.
Sullivan, J., J. A. Markert, and C. W. Kilpatrick. 1997. Phylogeography and molecularsystematics of the *Peromyscus aztecus* species group (Rodentia: Muridae) inferred using parsimony and likelihood. Syst. Biol. 46:426–440.
Sullivan, J., and D. L. Swofford. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mamm. Evol. 4:77–86.
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.
Swofford, D. L. 2002. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.
Whelan, S., P. Lio, and N. Goldman. 2001. Molecular phylogenetics: State-of-the-art methods for looking into the past. Trends Genet. 17:262–272.

First submitted 26 June 2007; reviews returned 10 September 2007; final acceptance 23 October 2007
Editor in Chief: Rod Page and Thomas Buckley