Syst. Biol. 55(1):89–96, 2006
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500433565

# Comparative Performance of Bayesian and AIC-Based Measures of Phylogenetic Model Uncertainty

MICHAEL E. ALFARO[1] AND JOHN P. HUELSENBECK[2]

[1]*School of Biological Sciences, P.O. Box 644236, Pullman, Washington 99164-4236, USA; E-mail: alfaro@wsu.edu*
[2]*Section Ecology, Evolution, and Behavior, 9500 Gilman Drive, Muir Biology 0116, La Jolla, California 92093, USA; E-mail: johnh@biomail.ucsd.edu*

*Abstract.*— Reversible-jump Markov chain Monte Carlo (RJ-MCMC) is a technique for simultaneously evaluating multiple related (but not necessarily nested) statistical models that has recently been applied to the problem of phylogenetic model selection. Here we use a simulation approach to assess the performance of this method and compare it to Akaike weights, a measure of model uncertainty that is based on the Akaike information criterion. Under conditions where the assumptions of the candidate models matched the generating conditions, both Bayesian and AIC-based methods perform well. The 95% credible interval contained the generating model close to 95% of the time. However, the size of the credible interval differed with the Bayesian credible set containing approximately 25% to 50% fewer models than an AIC-based credible interval. The posterior probability was a better indicator of the correct model than the Akaike weight when all assumptions were met but both measures performed similarly when some model assumptions were violated. Models in the Bayesian posterior distribution were also more similar to the generating model in their number of parameters and were less biased in their complexity. In contrast, Akaike-weighted models were more distant from the generating model and biased towards slightly greater complexity. The AIC–based credible interval appeared to be more robust to the violation of the rate homogeneity assumption. Both AIC and Bayesian approaches suggest that substantial uncertainty can accompany the choice of model for phylogenetic analyses, suggesting that alternative candidate models should be examined in analysis of phylogenetic data. [AIC; Akaike weights; Bayesian phylogenetics; model averaging; model selection; model uncertainty; posterior probability; reversible jump.]

Estimation of all of the parameters of a phylogenetic model can be impaired by the choice of a poor model and this can lead to incorrect inference (e.g., Buckley and Cunningham, 2002; Buckley et al., 2001; Sullivan et al., 1995; Sullivan and Swofford, 1997). It has now become standard practice for phylogeneticists to devote some effort to the problem of choosing a good model of sequence evolution at the start of data analysis. Most commonly, this is accomplished by choosing from a pool of candidate models either by a series of nested likelihood ratio tests or the Akaike information criterion (Akaike, 1973). The program ModelTest (Posada and Crandall, 1998) works in conjunction with PAUP* (Swofford, 2003) to automate this procedure and is presently the most commonly used method of selecting a model for phylogenetic analysis. Other approaches have recently been developed including the use of decision theory to select models that minimize error in branch length estimation (Abdo et al., 2005; Sullivan, 2003).

Despite the emphasis that is placed on model selection and the proliferation of methods to evaluate models little attention has been focused on how much better the "best" model is from competing models. Assessment of model uncertainty is important because inferences may be affected by the choice of model. If the data for typical phylogenetic studies is decisive with respect to model choice—that is, if fit of the data to the best model is substantially greater than the fit to other candidate models—then the current practice of conditioning inference on a single model is justified. However, if the best model is only marginally better than an alternative model or models, then additional analyses under reasonable alternative models might be necessary to ensure robust phylogenetic inference. Results that are tied closely to the choice of a particular model might be judged less reliable than results that obtain over a range of reasonable models.

A second issue that remains to be addressed in the area of model selection is how to decide on an appropriate pool of candidate models. For example, ModelTest considers a pool of 56 candidate models. However, with respect to substitution classes, ModelTest examines only eight distinct models. This represents a small fraction of the 203 possible time-reversible models (Huelsenbeck et al., 2004). There would appear to be little justification for excluding all of these models from consideration. Analysis of a diverse range of data sets reveals that unnamed models often provide a better fit to the data than the named models (Huelsenbeck et al., 2004).

One reason that alternative time-reversible models are not considered is that likelihood-ratio tests have gained popularity as the method for selecting the best phylogenetic models (Posada and Buckley, 2004). Because likelihood-ratio tests rely on a chi-squared approximation of the expected distribution of log-likelihood differences among models, they require that candidate models be arranged in a nested hierarchy. As additional models are considered, it becomes difficult or impossible to develop a hierarchical scheme for them. Thus, examination of most of the possible time-reversible models is not feasible in a likelihood ratio test framework, without resorting to simulation under the null hypothesis. One solution to the problem of examining a large pool of reasonable candidate models is to adopt an approach that accommodates the comparison of non-nested models. The two most natural frameworks for these kinds of comparisons are AIC-based information theoretic and Bayesian.

The Akaike information criterion is a metric that describes the Kullback-Liebler distance of a model to the

Downloaded from http://sysbio.oxfordjournals.org/ by guest on March 18, 2012

true model (Akaike, 1973). The AIC score for each candidate in a pool of models can be calculated as

$$\mathrm{AIC} = -2\hat{l} + 2\mathrm{K}$$

where $\hat{l}$ is the log of the maximum likelihood of the model and K is the number of parameters. Akaike weights are based on the difference in AIC score between each model in the pool and the best model. To calculate the Akaike weight, one first calculates these differences ($\Delta$AIC):

$$\Delta \mathrm{AIC}_i = \mathrm{AIC}_i - \min \mathrm{AIC}$$

The relative likelihood of a model $g_i$ is a function of its Akaike difference $\Delta_i$ (Burnham and Anderson, 2003)

$$\mathrm{L}(g_i|\hat{\theta}) \propto \exp(-1/2\Delta_i)$$

where $\hat{\theta}$ is the data. By normalizing these relative likelihoods, the Akaike weight ($w_i$) of each model can be calculated.

$$w_i = \frac{\exp(-1/2\Delta_i)}{\sum_{r=1}^{R} \exp(-1/2\Delta_r)}$$

Here $R$ is the number of models in the pool of candidate models. Akaike weights sum to one and have the interpretation of describing the weight of the evidence favoring model $i$ as the model that minimizes the Kullback-Liebler distance to the true model (Burnham and Anderson, 2003). Akaike weights can be used to directly compare the weight of evidence for one model over another. Furthermore, a confidence set can be created by ranking models by their Akaike weight and adding them to the confidence set until their cumulative probability reaches 95%. Relatively little work has been performed exploring the nature of model uncertainty using Akaike weights in a phylogenetic context (but see Buckley et al., 2002; Posada and Buckley, 2004).

Assessment of model uncertainty in a Bayesian framework has recently been facilitated by the development of reversible jump algorithms for phylogenetic MCMC analysis (Huelsenbeck et al., 2004). This method allows the Markov chain to visit models with different numbers of parameters and traverse the entire space of time-reversible models. Posterior probabilities for any particular model can be calculated directly using the MCMC samples just as they would be used to determine the posterior for a parameter of the phylogenetic model in a traditional MCMC analysis.

Bayesian and information theoretic approaches were developed from distinct schools of statistical thought and have important differences (Burnham and Anderson, 2003), yet little is known about how these underlying differences translate to the analysis of phylogenetic data in the area of model selection. To examine the performance of AIC and Bayesian methods in assessing model uncertainty, we compared how these methods characterized uncertainty around simulated data sets. Simulation is a powerful tool for comparing alternative methods because the generating model is known and the differences in performance can thus be readily interpreted (Hillis, 1995; Hillis and Huelsenbeck, 1994).

We ask several questions related to the performance of these methods in phylogenetic model choice: how often is the generating model identified, how well does the 95% confidence interval succeed in capturing the generating model, what is the relative size of the confidence intervals, and are these intervals biased? We also explore how these methods perform under ideal conditions where all model assumptions are satisfied as well as under conditions where the model has been violated.

In this study, we focus on the problem of selecting the correct model of phylogenetic inference given a data set and use statistical consistency as the criterion for assessing the performance of Bayesian and AIC model selection methods with the principal goal of understanding how these methods differ. We focus on the parameterization of the model with respect to substitution types *only*. That is, we explore the frequency with which these methods recover models that match the generating model in the parameterization of the rate matrix without attention to the actual value of those parameters. Thus, we judge a method to be good if it recovers the generating model or models close to it (in terms of rate matrix parameterization) with high frequency. An understanding of method performance under such conditions (i.e., when the generating model is known and included in the pool of candidate models) is helpful in identifying inherent biases and properties of the methods themselves. Alternative simulation strategies, where the generating model is far more complex than the pool of candidate models (as advocated by Burnham and Anderson, 2003), are also useful for characterizing performance of methods in estimating "important" parameters of the model (e.g., Abdo et al., 2005) under "realistic" conditions. However, with more complex simulations, differences among methods may become closely tied to the exact choice of realistic generating conditions, making the identification of innate differences problematic. Thus, we feel that both strategies are useful, though we concentrate on relatively simple conditions in this study.

## Methods

### *Model Notation*

Throughout this study we make use of the restricted growth function notation of partitions to describe models, a scheme for model specification used in PAUP* (Swofford, 2003) and described in Huelsenbeck et al. (2004) to refer to the universe of time-reversible models. In it, we assign index values to each of the six substitution rates ($r_{AC}, r_{AG}, r_{AT}, r_{CG}, r_{CT}, r_{GT}$). If a model has the constraint that $r_i = r_j$, then the index value of those rates is the same. In addition, the index for the first rate is always 1 and indices are always labeled sequentially. Thus, the Jukes Cantor model (1969), where all rates are the same, is denoted "111111," and the GTR model,

Downloaded from http://sysbio.oxfordjournals.org/ by guest on March 18, 2012

with separate rates for all substitution types, is denoted "123456."

### Simulations

We employed a Bayesian simulation strategy, first described by Huelsenbeck and Rannala (2004). Unlike traditional simulations where parameters of the phylogenetic model are either fixed or systematically varied over a range, under this strategy model parameters are drawn from a prior distribution. This approach is consistent with the Bayesian model used in the analysis where values for all model parameters are assumed to be drawn from a specified prior.

We performed five sets of simulations. In the first three, we simulated data matrices of 100 and 1000 sites to examine how character number influenced the ability of Bayesian and AIC-based methods to identify the correct model of sequence evolution. For simulation sets 4 and 5, we examined only data sets of 1000 to focus on how the violation of the prior on models affected the two methods. In the first set of simulations, all assumptions of the Bayesian model where satisfied. We simulated data under time-reversible models with no rate variation across sites. In the second set of simulations, we explored how increasing the average length of the trees influenced model determination by changing $\lambda$, the exponential parameter for the branch length prior, from 10 to 3. All assumptions of the Bayesian RJ-MCMC model were satisfied in the second set of simulations as well (i.e., the branch length parameter for the RJ-MCMC analysis was also 3). In the third set of simulations, we violated the assumption of rate homogeneity by simulating data sets with gamma distributed rate variation and analyzing the data sets under the assumption of equal rates. For the fourth set of simulations, we violated the prior on substitution models for the Bayesian reversible-jump MCMC analysis (explained below) by selecting models from an informative prior for data set simulation (Table 1) and analyzing the data sets under a uniform prior on models. Simulating under an alternative prior on models did not represent a violation of the AIC analysis because the Akaike weights procedure does not require the assignment of prior probabilities to the pool of candidate models. Finally, in the fifth set of simulations we generated data with rate heterogeneity (as in the third set of simulations) under the informative prior from simulation set 4.

We created an informative prior using the 95% credible interval of models for 16 empirical data sets analyzed in Huelsenbeck et al. (2004). Ninty-five percent of the weight of our informative prior was assigned to this pool of 41 models derived from the empirical data sets, weighted by the average posterior probability that those models received (Table 1). The remaining 5% of the prior weight was distributed evenly over the 162 models that fell outside the credible set of all of the empirical data sets.

TABLE 1. Prior probability of time reversible models derived from empirical data sets in Huelsenbeck et al. (2004) (see Methods and Materials). "Model" denotes sequence models following notation described in Huelsenbeck et al. (2004). "Prior" is the prior probability the model received. Models not listed received a prior weight of 0.0003.

| Model | Notation | Prior | Model | Notation | Prior |
|---|---|---|---|---|---|
| 25 | 112212 | 0.094 | 147 | 123424 | 0.011 |
| 15[a] | 121121 | 0.090 | 102 | 112213 | 0.011 |
| 193 | 123345 | 0.073 | 90 | 121321 | 0.011 |
| 50 | 123323 | 0.051 | 201 | 121345 | 0.010 |
| 112 | 121323 | 0.048 | 152 | 123423 | 0.010 |
| 40[b] | 121131 | 0.047 | 97 | 112313 | 0.009 |
| 162 | 121343 | 0.041 | 95 | 121123 | 0.009 |
| 203[c] | 123456 | 0.040 | 177 | 123421 | 0.009 |
| 60 | 123313 | 0.039 | 164 | 123413 | 0.008 |
| 166 | 123143 | 0.038 | 198 | 123451 | 0.008 |
| 64 | 112232 | 0.031 | 85 | 123121 | 0.007 |
| 189 | 123454 | 0.030 | 117 | 123123 | 0.007 |
| 125 | 123343 | 0.029 | 169 | 123314 | 0.007 |
| 168[d] | 123341 | 0.028 | 159 | 123414 | 0.006 |
| 136 | 121341 | 0.028 | 183 | 121324 | 0.006 |
| 200 | 123145 | 0.018 | 171 | 112343 | 0.006 |
| 138 | 121134 | 0.017 | 122[e] | 123321 | 0.005 |
| 134 | 123141 | 0.016 | 173 | 112342 | 0.005 |
| 157 | 123324 | 0.015 | 175 | 112234 | 0.004 |
| 100 | 112312 | 0.014 | 140 | 112314 | 0.002 |
| 191 | 123453 | 0.012 | 162 others | — | 0.050 |

[a] K80 (Kimura, 1980); [b] TrN (Tamura and Nei, 1993); [c] GTR (Tavare, 1986); [d] TVM (Posada, 2003); [e] K3P (Kimura, 1981).

As an illustration of the simulation procedure, we describe the steps in our analysis for creating a single replicate data set. First we selected a time-reversible model by drawing either from a uniform distribution over all 203 models or the informative prior described in Table 1. Second, we assigned values to the parameter(s) of the rate matrix for the model by generating an appropriate number (from 1 to 6, depending on the specific time-reversible model selected in step 1) of independent, exponentially distributed random variables with parameter 1. Third, we drew a tree of 10 taxa randomly from the pool of 2,027,025 possible topologies. Fourth, we assigned the length of each branch on the tree by independently drawing the length from an exponential distribution with parameter $\lambda = 3$ or $\lambda = 10$, depending on the experiment. The average expected branch length is equal to the mean on the exponential, $1/\lambda$, so branches in this study had average lengths of either 0.1 or 0.3 expected changes per site. Fifth, we chose the nucleotide frequencies by drawing from a Dirichlet (5,5,5,5) distribution. Sixth, for the model violation simulations, we chose the gamma-shape parameter for rate heterogeneity from an exponential distribution with parameter 2. This produced an average $\alpha$ value of 0.5 over the simulations. Finally, we simulated the evolution of 100 or 1000 sites under the model with the chosen parameter values. This process was repeated 1000 times for each of the simulation sets.

### Analysis

Simulated data sets were analyzed using a program written by JPH that implements the reversible jump method for model exploration (described in Huelsenbeck et al., 2004). Preliminary analysis revealed that 25,000 generations were sufficient for the chain to reach stationarity for the data sets in this study. We ran one cold

Downloaded from http://sysbio.oxfordjournals.org/ by guest on March 18, 2012

and four heated Markov chain for 500,000 generations, sampling every 100 generations. All of the priors for the Bayesian analysis matched those used in the simulation of the data except in simulation sets 3, 4, and 5, where the assumptions of rate constancy, model prior, or both, were violated.

We assessed performance in a number of ways. We calculated the posterior probability of each of the 203 models, estimated from the MCMC samples, and examined how often the generating model equaled the model that received the highest posterior probability. We assembled the 95% credible set of models by ranking models by their posterior probability and adding them to the credible set until the cumulative posterior probability of all models in the set was 95%. We then calculated the average probability that the credible set contained the generating model as well as the number of models that the set contained.

To compare Bayesian and AIC approaches, we used PAUP* (Swofford, 2003) to calculate the AIC score (Akaike, 1973) of each of the 203 time-reversible models for each simulation replicate. For example, for the first simulation replicate, we used the rclass command to define a model of sequence evolution, starting with the Jukes-Cantor model (rclass = (aaaaaa)). We then used PAUP* (Swofford, 2003) to obtain the likelihood score of this model on the generating tree, estimating the matrix parameters and base frequencies (rmatrix = estimate, basefreq = est). After recording the score to a log file, we iterated the model (rclass = aaaaab) and repeated the process until we obtained likelihood scores for the simulated data set under all 203 models. We used a simple computer program written in C to convert the likelihood scores to AIC scores and to transform the AIC scores into Akaike weights (Burnham and Anderson, 2002). In an analogous fashion to the Bayesian analysis above, we asked how often the generating model equaled the model with the best AIC score and how often the generating model fell within a 95% credible interval constructed with Akaike weights and recorded the average size of credible interval.

We also measured the distance and bias of the posterior distribution of models to the generating model. To measure distance, we used a metric developed by Gusfield (2002) to describe the similarity between two partitions called the "partition distance." Using our notation for substitution models (Huelsenbeck et al., 2004), partition distance can be thought of as the minimum number of elements that must be exchanged between partitions so that two models are identical (Fig. 1). We computed the partition distance for each of the 203 models to the generating model and weighted this value by the posterior probability or the Akaike weight that model received. The mean of this weighted sum is the average partition distance. This measure describes the amount of error associated with model estimation in the Bayesian and AIC procedures used in this study. The average partition distance should not be confused with the Robinson-Foulds partition distance (Robinson and Foulds, 1981), which is commonly used to compare phylogenetic trees.

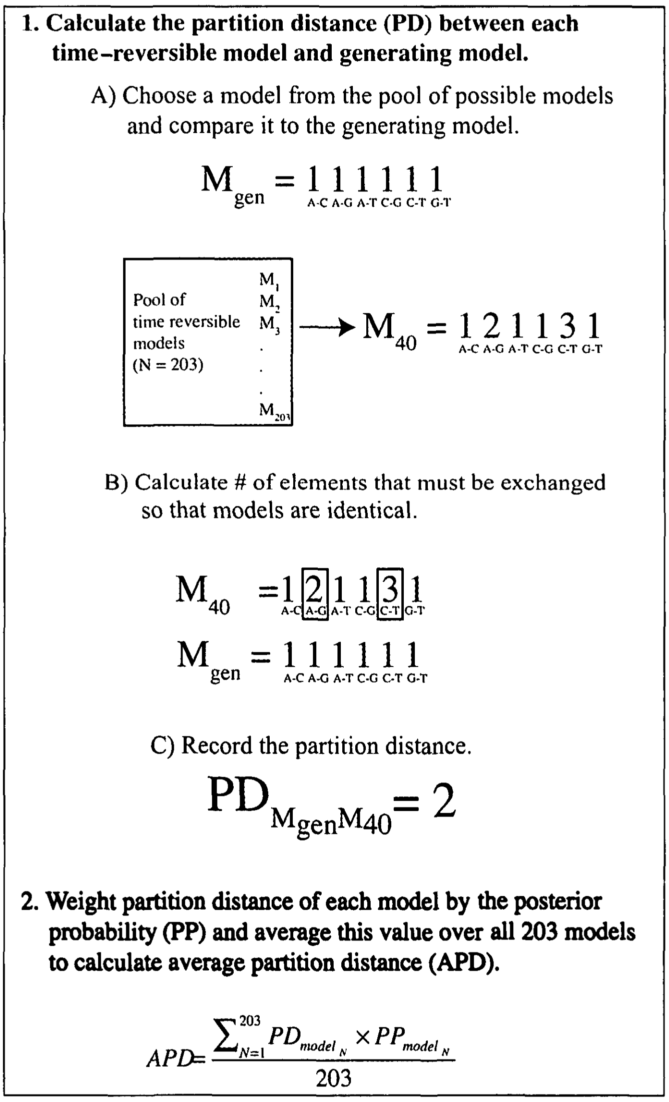
1. Calculate the partition distance (PD) between each time-reversible model and generating model.
A) Choose a model from the pool of possible models and compare it to the generating model.
M gen = 1 1 1 1 1 1
A-C A-G A-T C-G C-T G-T
Pool of time reversible models (N = 203)
M1 M2 M3 . . . M203
M40 = 1 2 1 1 3 1
A-C A-G A-T C-G C-T G-T
B) Calculate # of elements that must be exchanged so that models are identical.
M40 = 1 2 1 1 3 1
A-C A-G A-T C-G C-T G-T
M gen = 1 1 1 1 1 1
A-C A-G A-T C-G C-T G-T
C) Record the partition distance.
PD MgenM40 = 2
2. Weight partition distance of each model by the posterior probability (PP) and average this value over all 203 models to calculate average partition distance (APD).
APD = Σ N=1 203 PD model N × PP model N / 203

FIGURE 1. Calculation of average partiton distance. The partition distance (Gusfield, 2002) between the generating model and a candidate model is determined by the number of elements (rate parameters) in a model that must change partitions in order for the two models to be identical. The average partition distance is the weighted average of all partition distances for the 203 possible time-reversible models.

To measure the bias of Bayesian and AIC-based methods in characterizing model uncertainty, we defined average model bias using a procedure similar to that used in calculating the average partition distance above. We calculated the difference in model parameters between the generating model and each time-reversible model and weighted this value by the posterior probability or Akaike weight of the model. The average of these values over all models is the bias of the method and reflects the average difference in complexity (parameter number) between the generating model and those models judged to be good by Bayesian and Akaike-based approaches.

In this study, we measure the distance among models in terms of the number of rate categories only. Thus, for example, all K2P models (121121) have the same partition

Downloaded from http://sysbio.oxfordjournals.org/ by guest on March 18, 2012

TABLE 2. Probability model with best score equaled correct model. PP = posterior probability; AIC = Akaike weight; $\lambda$ = branch length parameter; $\alpha$ = alpha parameter of gamma-distributed rate heterogeneity. Italics indicate simulations generated under a prior on models generated from empirical data sets. Boldface indicates simulations where the assumption of rate constancy was violated.

| | | 100 sites | | 1000 sites | |
|---|---|---|---|---|---|
| | | PP | AIC | PP | AIC |
| $\lambda = 10$ | $\alpha = \infty$ | 19.6 | 16.2 | 53.5 | 40.1 |
| $\lambda = 3$ | $\alpha = \infty$ | 21.9 | 20.0 | 56.0 | 41.1 |
| $\boldsymbol{\lambda = 10}$ | $\boldsymbol{\alpha = 0.5}$ | **10.2** | **9.6** | **31.9** | **30.1** |
| $\lambda = \mathit{10}$ | $\alpha = \infty$ | | | *51.7* | *38.5* |
| $\boldsymbol{\lambda = 10}$ | $\boldsymbol{\alpha = \mathit{0.5}}$ | | | ***26.9*** | ***25.4*** |

TABLE 4. Probability generating model fell within 95% credible interval. PP = posterior probability; AIC = Akaike weight; $\lambda$ = branch length parameter; $\alpha$ = alpha parameter of gamma-distributed rate heterogeneity. Italics indicate simulations generated under a prior on models generated from empirical data sets. Boldface indicates simulations where the assumption of rate constancy was violated.

| | | 100 sites | | 1000 sites | |
|---|---|---|---|---|---|
| | | PP | AIC | PP | AIC |
| $\lambda = 10$ | $\alpha = \infty$ | 92.6 | 92.6 | 94.3 | 96.1 |
| $\lambda = 3$ | $\alpha = \infty$ | 93.8 | 94.0 | 95.3 | 95.6 |
| $\boldsymbol{\lambda = 10}$ | $\boldsymbol{\alpha = 0.5}$ | **88.3** | **89.5** | **84.8** | **93.8** |
| $\lambda = \mathit{10}$ | $\alpha = \infty$ | | | *91.6* | *95.6* |
| $\boldsymbol{\lambda = 10}$ | $\boldsymbol{\alpha = \mathit{0.5}}$ | | | ***81.7*** | ***91.2*** |

distance (2) to the Jukes-Cantor model (111111) whether kappa is 1.01 or 100. Although a finer scale distance measure might be desirable, we believe that the partition distance captures an important aspect of phylogenetic model selection where workers are typically faced with the problem of choosing the number an appropriate number of parameters for their analysis.

## RESULTS

The probability that the model with the highest posterior probability or Akaike weight equaled the generating model was low under all conditions examined in this study, ranging from approximately 10% when the data set was small; the assumption of among-site rate constancy was violated to 56% when the data set was large and the average branch length was long (Table 2). Increasing character number from 100 to 1000 sites doubled or tripled the probability of selecting the correct model. The posterior probability was always more successful in indicating the generating model than the Akaike weight, although when the assumptions of the model were violated, the difference between these methods was usually negligible. Violation of the assumption of rate constancy decreased the probability of identifying the correct model.

Support for the single best model (the Bayesian maximum a posteriori or MAP estimate) was also generally low, ranging from approximately 13% under the most challenging conditions (small data set size with rate heterogeneity) to approximately 55% for many of the simulation sets with 1000 sites (Table 3). The posterior probability of the MAP estimate was always higher than the Akaike weight, and usually twice as great. Increasing data set size from 100 to 1000 sites more than doubled the posterior probability or Akaike weight. Violation of the assumption of rate constancy lowered the posterior probability and the Akaike weight of the MAP estimate.

TABLE 3. Score of best model. PP = posterior probability; AIC = Akaike weight; $\lambda$ = branch length parameter; $\alpha$ = alpha parameter of gamma-distributed rate heterogeneity. Italics indicate simulations generated under a prior on models generated from empirical data sets. Boldface indicates simulations where the assumption of rate constancy was violated.

| | | 100 sites | | 1000 sites | |
|---|---|---|---|---|---|
| | | PP | AIC | PP | AIC |
| $\lambda = 10$ | $\alpha = \infty$ | 0.22 ± 0.13 | 0.12 ± 0.07 | 0.54 ± 0.19 | 0.23 ± 0.12 |
| $\lambda = 3$ | $\alpha = \infty$ | 0.24 ± 0.15 | 0.13 ± 0.07 | 0.55 ± 0.18 | 0.23 ± 0.12 |
| $\boldsymbol{\lambda = 10}$ | $\boldsymbol{\alpha = 0.5}$ | **0.13 ± 0.09** | **0.08 ± 0.05** | **0.40 ± 0.19** | **0.18 ± 0.11** |
| $\lambda = \mathit{10}$ | $\alpha = \infty$ | | | *0.55 ± 0.19* | *0.25 ± 0.14* |
| $\boldsymbol{\lambda = 10}$ | $\boldsymbol{\alpha = \mathit{0.5}}$ | | | ***0.39 ± 0.18*** | ***0.19 ± 0.11*** |

For data sets of 100 sites, the probability that the generating model fell within the credible interval was slightly under 95% (Table 4). Coverage probabilities increased to approximately 95% with 1000 sites when no model assumptions were violated. The AIC credible interval was always equally, or more likely, to contain the generating model than the Bayesian credible interval. Both Bayesian and AIC coverage probabilities fell when assumptions of the model were violated although the Bayesian credible interval generally fell more. Violation of the assumption of rate constancy degraded the performance of the AIC credible interval but had a more profound effect on the Bayesian credible set. Although increasing data set size from 100 to 1000 sites improved AIC coverage probability, additional data *decreased* the reliability of the Bayesian credible set.

The credible interval contained a large number of models (Table 5) on average. With small data sets, uncertainty was spread out over almost 80 models. Increasing data set size decreased the size of the credible sets, although under the most favorable conditions in our simulations, the credible interval still contained 10 to 11 models. The Bayesian credible set was always smaller than the AIC interval, containing up to 75% fewer models. The addition of rate heterogeneity substantially increased the size of the credible interval while the violation of the prior on models appeared to have a minor or no effect on credible set size.

TABLE 5. Number of models in the 95% credible set. PP = posterior probability; AIC = Akaike weight; $\lambda$ = branch length parameter; $\alpha$ = alpha parameter of gamma-distributed rate heterogeneity. Italics indicate simulations generated under a prior on models generated from empirical data sets. Boldface indicates simulations where the assumption of rate constancy was violated.

| | | 100 sites | | 1000 sites | |
|---|---|---|---|---|---|
| | | PP | AIC | PP | AIC |
| $\lambda = 10$ | $\alpha = \infty$ | 42.60 ± 37.68 | 47.33 ± 40.51 | 11.61 ± 14.52 | 19.42 ± 26.26 |
| $\lambda = 3$ | $\alpha = \infty$ | 39.09 ± 37.46 | 43.10 ± 40.54 | 11.40 ± 13.76 | 18.97 ± 24.81 |
| $\boldsymbol{\lambda = 10}$ | $\boldsymbol{\alpha = 0.5}$ | **78.80 ± 51.59** | **79.04 ± 49.40** | **22.80 ± 27.51** | **31.51 ± 37.78** |
| $\lambda = \mathit{10}$ | $\alpha = \infty$ | | | *9.92 ± 11.30* | *16.28 ± 20.30* |
| $\boldsymbol{\lambda = 10}$ | $\boldsymbol{\alpha = \mathit{0.5}}$ | | | ***24.53 ± 31.02*** | ***30.85 ± 37.87*** |

Downloaded from http://sysbio.oxfordjournals.org/ by guest on March 18, 2012

TABLE 6. Average partition distance to generating model. PP = posterior probability; AIC = Akaike weight; $\lambda$ = branch length parameter; $\alpha$ = alpha parameter of gamma-distributed rate heterogeneity. Italics indicate simulations generated under a prior on models generated from empirical data sets. Boldface indicates simulations where the assumption of rate constancy was violated.

| | | 100 sites | | 1000 sites | |
|---|---|---|---|---|---|
| | | PP | AIC | PP | AIC |
| $\lambda = 10$ | $\alpha = \infty$ | 1.70 ± 0.49 | 1.82 ± 0.39 | 0.93 ± 0.58 | 1.38 ± 0.46 |
| $\lambda = 3$ | $\alpha = \infty$ | 1.63 ± 0.51 | 1.77 ± 0.40 | 0.89 ± 0.56 | 1.38 ± 0.46 |
| **$\lambda = 10$** | **$\alpha = 0.5$** | **2.03 ± 0.45** | **2.07 ± 0.38** | **1.36 ± 0.65** | **1.58 ± 0.49** |
| *$\lambda = 10$* | *$\alpha = \infty$* | | | *0.93 ± 0.58* | *1.31 ± 0.46* |
| ***$\lambda = 10$*** | ***$\alpha = 0.5$*** | | | ***1.46 ± 0.67*** | ***1.60 ± 0.50*** |

The average partition distance to the generating model ranged from approximately 2.0 when data set size was small to just less than 0.9 under simulation conditions with long branches and 1000 sites (Table 6). Bayesian partition distances were always lower than AIC partition distances. Violation of the rate constancy assumption increased the average partition distance while violation of the model prior did not appear to affect this measure.

Average model bias for the AIC was always slightly positive (meaning that, on average, models more complex than the generating model were preferred), ranging from 0.15 with 100 sites and rate heterogeneity to 0.80 with 1000 sites and long branch lengths (Table 7). In contrast, Bayesian average model bias was close to zero except with data sets of 1000 sites where the prior on models was violated. In these cases, Bayesian model bias was slightly negative, meaning that models less complex than the generating model were preferred. Simulation under the informative model prior also decreased AIC bias, though it remained above 0.4 under the most severe test in our simulations.

## DISCUSSION

Our simulation results underscore the point made by recent empirical studies (Buckley et al., 2002): that substantial uncertainty can surround model identification in phylogenetic analyses. Under all of the conditions examined in this study, identification of the single best model for analysis was problematic. The model with the highest AIC score or posterior probability was, frequently, not the generating model. The best models often received low support and reasonable confidence could be spread out over a large number of possible models.

TABLE 7. Average model bias. PP = posterior probability; AIC = Akaike weight; $\lambda$ = branch length parameter; $\alpha$ = alpha parameter of gamma-distributed rate heterogeneity. Italics indicate simulations generated under a prior on models generated from empirical data sets. Boldface indicates simulations where the assumption of rate constancy was violated.

| | | 100 sites | | 1000 sites | |
|---|---|---|---|---|---|
| | | PP | AIC | PP | AIC |
| $\lambda = 10$ | $\alpha = \infty$ | 0.02 ± 0.77 | 0.39 ± 0.78 | −0.01 ± 0.61 | 0.76 ± 0.64 |
| $\lambda = 3$ | $\alpha = \infty$ | 0.00 ± 0.77 | 0.42 ± 0.78 | 0.01 ± 0.57 | 0.80 ± 0.59 |
| **$\lambda = 10$** | **$\alpha = 0.5$** | **−0.06 ± 0.82** | **0.15 ± 0.84** | **−0.03 ± 0.75** | **0.60 ± 0.72** |
| *$\lambda = 10$* | *$\alpha = \infty$* | | | *−0.11 ± 0.69* | *0.64 ± 0.76* |
| ***$\lambda = 10$*** | ***$\alpha = 0.5$*** | | | ***−0.19 ± 0.88*** | ***0.44 ± 0.90*** |

Together, these results cast doubt on the adequacy of some standard practices in phylogenetics regarding model selection. Perhaps most importantly, simply selecting a single model on the basis of likelihood-ratio tests or AIC score may not ensure the choice of the best phylogenetic model. There are two reasons for this. First, the number of models considered with respect to substitution type by ModelTest is a small fraction of the universe of possible models, or even of those models that are reasonable for empirical data sets (Table 1). Thus, the optimal model for any particular analysis might not ever be evaluated. In addition, even when the best model is among the candidate models, the best-supported model may often receive a relatively small proportion of overall support (Table 3) and may not correspond to the best (true) model (Table 2).

### *AIC versus Bayesian Measures of Model Uncertainty*

The Akaike information criterion and Bayesian inference provide natural frameworks for characterizing uncertainty surrounding model choice although their statistical assumptions differ in many important respects. Both measures appear to perform similarly with very small data sets and reflect high uncertainty in model identification under these conditions. However, at more meaningful data set sizes, the confidence intervals described by these measures differ. The Bayesian posterior probability appears to be more tightly distributed around the generating model, with a smaller credible interval, higher support for the best model, and higher correspondence between the best supported model and the generating model (Tables 2, 3, 5). Models in the Bayesian posterior distribution also more closely match the generating model in number of parameters (Table 6). In contrast, the distribution of Akaike-weighted models is more diffuse with lower support for any particular model, more models in the 95% credible interval, and a higher average partition distance. Notably, the distribution of Akaike-weighted models is biased slightly towards models of greater complexity, whereas the Bayesian posterior distribution of models is unbiased, or slightly biased towards less complex models (Table 7).

From these considerations, the Bayesian approach appears to recommend itself over the use of Akaike weights to characterize model uncertainty. However, the increased fidelity of the Bayesian posterior distribution to the generating model appears to come at the cost of robustness to violation of the model assumptions. Under the most severe violation of model assumptions in this study, the 95% Akaike weights interval still contained the generating model over 90% of the time while performance of the smaller Bayesian credible interval fell to 82% (Table 4). Also worrisome is the differential effect on these methods of adding data under conditions where the assumptions of the model have been violated. More data increase the performance of the Akaike weights 95%

Downloaded from http://sysbio.oxfordjournals.org/ by guest on March 18, 2012

credible set but *decrease* the probability of the generating model falling within the Bayesian 95% credible interval (Table 4). This behavior suggests that under some conditions, model misspecification can cause a Bayesian analysis to be positively misled (e.g., Buckley, 2002) and underscores the need for models of adequate complexity in Bayesian phylogenetics (Huelsenbeck and Rannala, 2004).

### *Why Should Uncertainty in Model Choice Be Accommodated?*

Phylogenetic inference depends critically on the underlying assumptions of the statistical models used in analysis (Goldman, 1993). Models that poorly fit the data can lead to inconsistent behavior of likelihood methods (Gaut and Lewis, 1995; Sullivan and Swofford, 1997) and produce poor or misleading estimates of all of the parameters of the phylogenetic model including topology (Sullivan et al., 1995), branch lengths (Minin et al., 2003), and substitution rates (Wakeley, 1994), as well as influence bootstrap values and posterior probabilities (Buckley and Cunningham, 2002; Buckley et al., 2001; Sullivan and Swofford, 1997). To mitigate against these effects, phylogeneticists use likelihood ratio tests or AIC scores to choose the best model from a pool of candidate models for analysis. The weakness of this approach is that inference is still conditioned upon a single model that may be only marginally better than competing models. Theoretically, the failure to accommodate uncertainty in model choice will lead to an underestimate of the variance in the parameters of the model and, ultimately, to overconfident inferences (Hoeting et al., 1999; Madigan and Raftery, 1994). In practice, little work has been done investigating how phylogenetic model parameters change when model uncertainty is accommodated. Recently, Posada and Buckley (2004) showed that distantly related models within a credible set produced different estimates of tree topology, although the topology averaged over all models was very similar to the topology produced by the best model. The difference between parameters estimated by the best model and by averaging over candidate models was minor except when the amount of data relevant to parameter estimation was small.

On the basis of many studies that have shown topology to be relatively insensitive to the choice of phylogenetic model (Posada and Buckley, 2004; Posada and Crandall, 2001; Sullivan and Swofford, 2001; Yang et al., 1994), we predict that in most cases alternative models within the credible set will not produce substantially different estimates of the tree. However, the implications of topological differences, even if they involve a small number of clades, might be important in a particular analysis. Other parameters of the model may be more sensitive to the choice of model than tree topology, particularly in cases where little data are available upon which to infer the estimate. The danger for any particular analysis is that without an examination of model uncertainty, an investigator cannot know which, if any, of the parameters for the model at hand are sensitive to the choice of model. In addition, bootstrap values and posterior probabilities, which have both been shown to be sensitive to model choice, are likely to change when model uncertainty is accommodated. Adequate reflection of model uncertainty might especially improve posterior probability estimates by mitigating against the effects of model misspecification.

For these reasons, we strongly recommend that workers examine the effects of model choice on all of the parameters of interest in a particular analysis. Akaike weights are now reported as part of the output of ModelTest, facilitating the identification of reasonable alternatives to the best model. Model choice can also be performed within a Bayesian framework using reversible jump (Huelsenbeck at al., 2004). One advantage of the reversible-jump MCMC approach for those who are primarily interested in tree topology and support is that model-averaged clade posterior probabilities can be easily calculated from the Monte Carlo samples. The proportion of reversible-jump MCMC samples that contain the clade of interest is an estimate of the posterior probability of that clade. The only difference between the posterior from a traditional Bayesian analysis and one determined from reversible-jump samples is that the reversible-jump posterior probability reflects uncertainty across all possible time reversible models. Calculating the analogous model-averaged bootstrap proportion in an information-theoretic framework would require a multistep process of calculating bootstrap values or posterior probabilities for each credible model followed by an averaging of those support values according to model weight.

## Conclusions

Our study suggests that the Bayesian approach provides a more precise estimate of model uncertainty than an Akaike information criterion–based approach when model assumptions are satisfied but may also be less accurate in situations where model assumptions have been violated. More importantly, for data set sizes common in phylogenetic studies, it is not likely that the data decisively support a single model, underscoring a greater need for sensitivity analysis of phylogenetic results to model choice. Future study should focus on the behavior of these methods with increasing data set size and complexity. In addition, a more thorough examination of the effects of Bayesian and Akaike-based model averaging on tree topology and support estimation might reveal other important differences between these approaches.

## Acknowledgments

We thank Jon Bollback for helpful comments on many aspects of this research. We also thank Rod Page, Thomas Buckley, Peter Foster, and an anonymous reviewer for helpful comments on an earlier version of the manuscript. Funding was provided by NSF EF-0331654, NSF DEB-0445453, and NIH GM-069801.

Downloaded from http://sysbio.oxfordjournals.org/ by guest on March 18, 2012

## 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* Second Annual Symposium on Information Theory (B. N. Petrov, and F. Csaki, eds.). Akademi Kiado, Budapest.

Buckley, T. R. 2002. Model misspecification and probabilistic tests of topology: Evidence from empirical data sets. Syst. Biol. 51:509–523.

Buckley, T. R., P. Arensburger, C. Simon, and G. K. Chambers. 2002. Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. Syst. Biol. 51:4–18.

Buckley, T. R., and W. W. Cunningham. 2002. The effects of nucleotide substitution model assumptions of estimates of nonparametric bootstrap support. Mol. Biol. Evol. 19:394–405.

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

Burnham, K. P., and D. R. Anderson. 2003. Model selection and multimodel inference, a practical information-theoretic approach. Springer, New York.

Gaut, B., and P. 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.

Gusfield, D. 2002. Partition-distance: A problem and class of perfect graphs arising in clustering. Inform. Process. Lett. 82:159–164.

Hillis, D. M. 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. 44:3–16.

Hillis, D. M., and J. P. Huelsenbeck. 1994. To tree the truth: Biological and numerical simulation of phylogeny. Soc. Gen. Physiol. Ser. 49:55–67.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: A tutorial. Stat. Sci. 14:382–417.

Huelsenbeck, J. P., B. Larget, and M. E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 2004:1123–1133.

Huelsenbeck, J. P., and B. Rannala. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53:904–913.

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

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

Kimura, M. 1981. Estimation of evolutionary distance between homologous nucleotide sequences. Proc. Natl. Acad. Sci. 78:454–458.

Madigan, D., and A. E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. J. Amer. Statistical Assoc. 89:1535–1546.

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

Posada, D. 2003. Using ModelTest and PAUP* to select a model of nucleotide substitution. Pages 6.5.1–6.5.14 *in* Current protocols in bioinformatics (A. D. Baxevanis, ed.). John Wiley & Sons, New York.

Posada, D., and T. R. Buckley. 2004. Model selection and model averaging in phylogenetics: Advantages of AIC 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.

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

Swofford, D. L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Sullivan, J. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:1–10.

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

Sullivan, J., and D. L. Swofford. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mammal. Evol. 2:77–86.

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.

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526.

Tavare, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lec. Math. Life Sci. 17:57–86.

Wakeley, J. 1994. Substitution-rate variation and the estimation of transition bias. Mol. Biol. Evol. 11:436–442.

Yang, Z., N. Goldman, and A. Friday. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11:316–24.

*First submitted 11 January 2005; reviews returned 3 May 2005;*
*final acceptance 8 August 2005*
*Associate Editor: Thomas Buckley*

Downloaded from http://sysbio.oxfordjournals.org/ by guest on March 18, 2012