*Syst. Biol.* 53(5):793–808, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490522304

# Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests

DAVID POSADA[1] AND THOMAS R. BUCKLEY[2]

[1]*Departamento de Bioquímica, Genética e Inmunología, Facultad de Biología, Universidad de Vigo, Vigo 36200, Spain; E-mail: dposada@uvigo.es*
[2]*Landcare Research, Private Bag 92170, Auckland, New Zealand; E-mail: BuckleyT@LandcareResearch.co.nz*

*Abstract*.—Model selection is a topic of special relevance in molecular phylogenetics that affects many, if not all, stages of phylogenetic inference. Here we discuss some fundamental concepts and techniques of model selection in the context of phylogenetics. We start by reviewing different aspects of the selection of substitution models in phylogenetics from a theoretical, philosophical and practical point of view, and summarize this comparison in table format. We argue that the most commonly implemented model selection approach, the hierarchical likelihood ratio test, is not the optimal strategy for model selection in phylogenetics, and that approaches like the Akaike Information Criterion (AIC) and Bayesian methods offer important advantages. In particular, the latter two methods are able to simultaneously compare multiple nested or nonnested models, assess model selection uncertainty, and allow for the estimation of phylogenies and model parameters using all available models (model-averaged inference or multimodel inference). We also describe how the relative importance of the different parameters included in substitution models can be depicted. To illustrate some of these points, we have applied AIC-based model averaging to 37 mitochondrial DNA sequences from the subgenus *Ohomopterus* (genus *Carabus*) ground beetles described by Sota and Vogler (2001). [AIC; Bayes factors; BIC; likelihood ratio tests; model averaging; model uncertainty; model selection; multimodel inference.]

It is clear that models of nucleotide substitution (henceforth models of evolution) play a significant role in molecular phylogenetics, particularly in the context of distance, maximum likelihood (ML), and Bayesian estimation. We know that the use of one or other model affects many, if not all, stages of phylogenetic inference. For example, estimates of phylogeny, substitution rates, bootstrap values, posterior probabilities, or tests of the molecular clock are clearly influenced by the model of evolution used in the analysis (Buckley, 2002; Buckley and Cunningham, 2002; Buckley et al., 2001; Kelsey et al., 1999; Pupko et al., 2002; Sullivan and Swofford, 1997, 2001; Suzuki et al., 2002; Tamura, 1994; Yang et al., 1995; Zhang, 1999). We can argue, in general, that phylogenetic methods are less accurate (that is, they recover an incorrect phylogeny more often), or become inconsistent (converging to an incorrect tree with increasing number of characters) when the model of evolution assumed is wrong (Bruno and Halpern, 1999; Felsenstein, 1978; Huelsenbeck and Hillis, 1993; Penny et al., 1994). It is evident that the use of appropriate models is essential if we are to be confident in the results of a phylogenetic analysis, and indeed, several strategies for model choice have been proposed in the context of phylogenetics. We refer the reader to Johnson and Omland (2003), Posada and Crandall (2001b) and Posada (2001) for a detailed introduction, and for an evaluation of the performance of these methods to recover the model generating the data. Computer programs exist that implement these methods (Adachi and Hasegawa, 1996; Posada and Crandall, 1998). Among the available methods for model selection in phylogenetics, hierarchical likelihood ratio tests (hLRTs) are the most popular. However, here we argue that the hLRTs approach is not the optimal strategy for model selection in phylogenetics, and that approaches like the Akaike Information Criterion (AIC) and Bayesian methods offer important advantages. In particular, the latter two allow for assessment of model selection uncertainty and model averaging.

## MODEL SELECTION

Before proceeding further, it is worth reiterating the fact that any model of evolution we can construct is never going to be the "true model" that generated the data we observe. In other words, the set of models is misspecified. All models are wrong but some are useful (Box, 1976), and model selection is best seen as a way of approximating, rather than identifying, full reality (Burnham and Anderson, 2003, pp. 20–23). Statistical model selection is commonly based on William of Occam's (ca.1320) parsimony principle,[1] by which hypotheses should be kept as simple as possible. In statistical terms, this is a trade-off between bias (distance between the average estimate and truth) and variance (spread of the estimates around the truth) (Fig. 1). The idea is that by adding parameters to a model we obtain improvement in fit (see below) to some degree, but at the same time parameter estimates are "worse" because we have less data (i.e., information) per parameter. In addition, the computations typically require more time. So the question is how complex should the model be for a given problem.

## THE LIKELIHOOD FUNCTION

We referred above to the fit of a model to the data, but we have not yet explained how we measure this fit. In most cases, the fit of a model is measured by the likelihood function (see Edwards, 1972; Fisher, 1921), and in

[1]Occam's (ca. 1280–1349) parsimony principle or Occam's razor was stated as "*Pluralitas non est ponenda sine necessitate*," which translates literally into English as "plurality should not be posited without necessity."

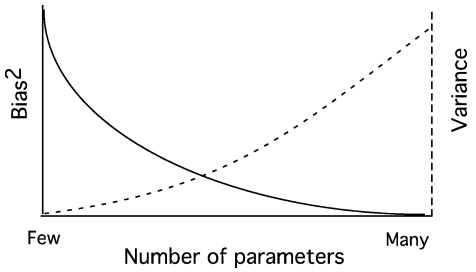


FIGURE 1. The principle of parsimony. Model selection is more or less based on the trade-off between bias and variance versus the number of estimable parameters in the model. The principle of parsimony tells us that as we increase the number of parameters in a model the bias decreases but the variance increases. This principle underlies all model selection approaches.

phylogenetics (see Felsenstein, 1981a; Goldman, 1990) we define the likelihood ($L$) as (proportional to) the probability of the data ($D$) given a model of evolution ($M$), a vector of $K$ model parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$, a tree topology ($\tau$), and a vector of $S$ branch lengths, $\nu = (\nu_1, \nu_2, \ldots, \nu_S)$:

$$L = P(D \mid M, \theta, \tau, \nu)$$

If the goal is to compute the likelihood of a given model, then $\theta$, $T$, and $\nu$ are *nuisance parameters*—they affect the likelihood calculation but they are not really what we want to infer—and they should somehow be eliminated from the inference. A common strategy to remove nuisance parameters is to assume that they take those values that maximize the overall likelihood, thus reducing the likelihood to a function of the parameters of interest. What is usually done in practice is to estimate a tree (topology and branch lengths) from the data and then—implicitly assuming that this tree is the maximum likelihood tree for every candidate model—calculate maximum likelihood estimates of all model parameters, including the branch lengths, for every model given this tree. In this way we obtain the *maximized* (log) likelihood under model $M$:

$$\ell = \ln P(D \mid M, \hat{\theta}, \hat{\tau}, \hat{\nu})$$

where $\hat{}$ means "estimate of" ($\hat{\theta}$ is an estimate of $\theta$). The strategy just described is sometimes called *joint estimation*. A different strategy to remove nuisance parameters is to assign them prior probabilities and integrate them out to obtain the *marginal probability* of the data given only the model, that is, the *model likelihood* (also called integrative, marginal, or predictive likelihood):

$$P(D \mid M) = \iiint P(D \mid M, \theta, \tau, \nu) P(\theta, \tau, \nu \mid M)\, d\theta d\tau d\nu$$

However, this multidimensional integral can be very difficult to compute, and it is typically approximated using computationally intensive techniques like Markov chain Monte Carlo (MCMC) (Gilks et al., 1996; Hastings, 1970; Metropolis et al., 1953). Steel and Penny (2000) and Holder and Lewis (2003) provide an instructive discussion on joint and marginal estimation in the context of phylogenetics.

## HIERARCHICAL LIKELIHOOD RATIO TESTS

The most popular strategy for model selection in phylogenetics are the hierarchical likelihood ratio tests (hLRTs) (Frati et al., 1997; Huelsenbeck and Crandall, 1997; Posada and Crandall, 1998) (Fig. 2). This method usually consists of performing pairwise likelihood ratio tests in a specific sequence until a final model is converged on that cannot be rejected. By means of the LRTs, we compare the maximized log-likelihoods of the null ($\ell_0$) and the alternative ($\ell_1$) models, and if the associated $P$-value is smaller than the predefined threshold (the *significance level*, usually 0.05), we say that alternative model fits the data significantly better than the null model (i.e., we reject the null model), and vice versa.

$$LRT = 2(\ell_1 - \ell_0)$$

The approximation of this $P$-value is straightforward for nested models, using a standard or mixed $\chi^2$ distribution (Goldman, 1993; Goldman and Whelan, 2000; Kendall and Stuart, 1979; Ota et al., 2000). Two models

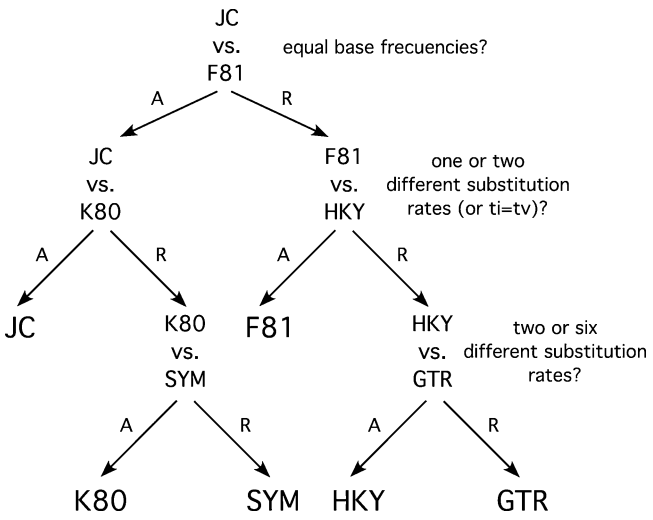


FIGURE 2. Hierarchical likelihood ratio tests (hLRTs). This figure illustrates an arbitrary hierarchy of LRTs for six different models. Within each LRT, the null model is depicted above the alternative model. When the LRT is not significant, the null model (above) is accepted (A), and it becomes the null model of the next LRT. When the LRT is significant, the null model is rejected (R) and the alternative model (below) becomes the null model of the next LRT. There are six possible paths depending on the outcome of the individual LRTs, and each path results in the selection of a different model. JC69: Jukes-Cantor model (Jukes and Cantor, 1969); K80: Kimura 1980 model (Kimura, 1980), also known as K2P; F81: Felsenstein 81 model (Felsenstein, 1981b); HKY85: Hasegawa-Kishino-Yano model (Hasegawa et al., 1985); SYM, symmetrical model (Zharkikh, 1994); GTR: general-time reversible model (Tavaré, 1986), also known as REV.

are nested when one of them, the null model, is a special case of the other, the alternative model. For example, the Jukes-Cantor model (Jukes and Cantor, 1969) (JC69) is nested within the Kimura two-parameter model (Kimura, 1980) (K80), because if we assume that transitions and transversions occur at the same rate (i.e., $\kappa = 1$), K80 collapses to JC69. However, obtaining *correct* $P$-values for the LRT statistics can be difficult. LRTs implicitly assume that at least one of the models compared is correct, and when the models are misspecified these tests can often be incorrect (Foutz and Srivastava, 1977; Golden, 1995; Kent, 1982). Although proper LRTs can be constructed when models are wrong (Vuong, 1989), standard LRTs in phylogenetics are not robust to model misspecification (Zhang, 1999). When the models are non-nested, the $\chi^2$ approximation is not longer valid, and more computationally intensive Monte Carlo methods are needed (Goldman, 1993; Whelan and Goldman, 1999). In addition, when sample size is small the usual asymptotic approximation on which $P$-values are based no longer applies.

Furthermore, LRTs were designed for hypothesis testing, and although classical hypothesis testing is commonly used as a model selection strategy, it has been argued that hypothesis testing and model selection are distinct issues (Burnham and Anderson, 2003, pp. 132–134). A stepwise procedure like the hLRTs, in which we sequentially decide whether to add (or remove) certain parameters, is analogous to forward and backward selections in best-subset linear regression (Miller, 2002, pp. 39–46), which do not guarantee finding the optimal model. As pointed out by Sanderson and Kim (2000), we can identify several potential problems with the use of hLRTs for model selection in phylogenetics. There exist situations in which an optimal model may not exist for the hLRTs procedure. This kind of situation occurs, for example, if the general time-reversible model (Tavaré, 1986) (GTR) is not significantly better than the Hasegawa et al. model (1985) (HKY85), HKY85 is not significantly better than JC69, but GTR is significantly better than JC69. Even if an optimal model exists, it will be always a function of the significance level, and the outcome of the model choice procedure may vary accordingly. In addition, the hLRTs approach performs multiple tests with the same data, and this will increase the rate of false positives (that is, to reject the null hypothesis when it is true): the probability of falsely rejecting the null hypothesis at least once in $n$ tests is $1-(1-\alpha)^n$. Although there are statistical procedures to correct for this effect—like the Bonferroni correction (see Hochberg, 1988)—here the tests are nonindependent, and the appropriate adjustment can be very complex (see also Shimodaira, 1998, 2001; Shimodaira and Hasegawa, 1999). The outcome of the hLRTs might also be affected by the starting model (for the hLRTs procedure we need to select a starting point, usually represented by the simplest or the most complex model in the set of candidate models). In addition, there are cases in which the hLRTs will not select the best model, according to its own criteria, among the candidate models.

Indeed, these problems can have an impact on the analysis of real data sets, and we have analyzed a set of HIV sequences (Posada and Crandall, 2001a) for illustrative purposes (Fig. 3) (Pol, in press). In Figure 3a we can see a case in which an optimal model does not exist, as all of the three models are rejected when compared with one of the other two. However, we will select HKY85 as the best fit (because we did not compare HKY85 and GTR). Also, note that increasing the significance level (Fig. 3b) changes the outcome, as GTR now becomes the best fit model. With a different set of candidate models, and if we start with HKY85, the model selected will be HKY85 (Fig. 3c), which is a suboptimal choice, whereas if we start with GTR the model selected will be GTR (Fig. 3d), which is actually the optimal model. We cannot devise a hierarchy of hLRTs that overcomes all these problems at once, but better approaches exist than simply forward and backward selection (Miller, 2002).

## Bayesian Model Selection

Model selection is an integral part of Bayesian estimation (Gelfand, 1996; Raftery, 1996; Wasserman, 2000), and within this framework, different strategies exist to accomplish the same tasks.

### *Bayes Factors*

Bayes factors (Kass and Raftery, 1995) are the Bayesian analogue of the LRT (Suchard et al., 2003a). They contrast the evidence provided by the data for two competing models, $i$ and $j$, as:

$$B_{ij} = \frac{P(D \mid M_i)}{P(D \mid M_j)}$$

Evidence for $M_i$ is considered very strong if $B_{ij} > 150$, strong if $12 < B_{ij} < 150$, positive if $3 < B_{ij} < 12$, barely worth mentioning if $1 < B_{ij} < 3$, and negative (supports $M_j$) if $B_{ij} < 1$ (Raftery, 1996). It is important to note that Bayes factors compare *model likelihoods* or $P(D \mid M)$, which are calculated by integrating—not maximizing—over all possible parameter values (except in empirical Bayesian approaches, where maximum likelihood estimates can be used instead). Therefore we should not confound them with the log of the maximized likelihoods ($\ell$) used in the LRTs and AIC. Bayes factors are already being used in the context of phylogenetics, for example to infer the occurrence of recombination events (Suchard et al., 2002), to compare different phylogenetic hypothesis (Huelsenbeck and Imennov, 2002; Huelsenbeck et al., 2000; Suchard et al., 2003b) and for model selection (Aris-Brosou and Yang, 2002; Huelsenbeck et al., 2004; Nylander et al., 2004; Suchard et al., 2001).

### *Posterior Probabilities*

When multiple models are considered, the usual Bayesian solution is to choose the model with the highest posterior probability (Kass and Raftery, 1995; Raftery,

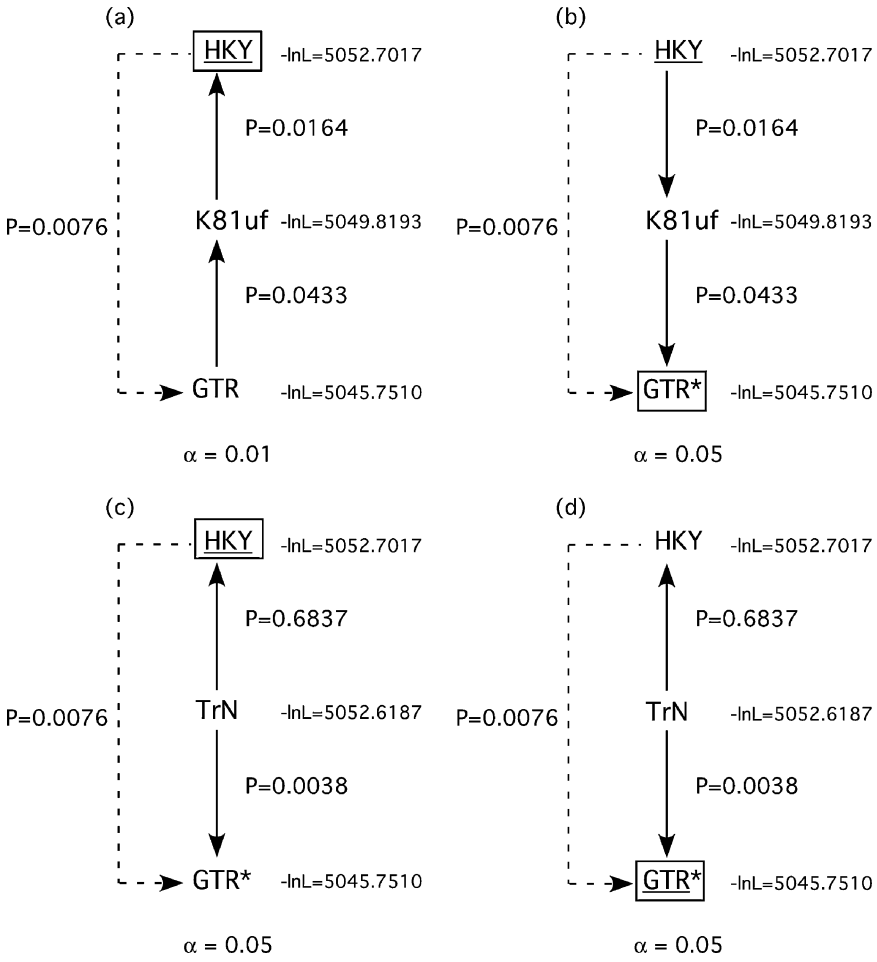


FIGURE 3. Problems of hLRTs with a real data set. See text for further details. The data set analyzed is an alignment of 12 HIV-1 subtype D sequences of a fragment of 1462 nucleotides from the *gag* region (Posada and Crandall, 2001a). K81uf is the Kimura 1981 model (Kimura, 1981) with unequal base frequencies. TN93 is the Tamura-Nei model (Tamura and Nei, 1993). Solid arrows indicate the outcome of the LRT performed, whereas discontinuous arrows indicate the outcome of a potential LRT not performed. P is the associated $P$-value of the LRTs. The underlined model is the starting point of the hLRT, the best model according to *all* LRTs is indicated with an asterisk, and the model selected is enclosed within a square.

1996; Wasserman, 2000). For $R$ models, the posterior probability of the $i$th model is:

$$P(M_i \mid D) = \frac{P(D \mid M_i)P(M_i)}{\sum_{r=1}^{R} P(D \mid M_r)P(M_r)}$$

A word is needed about model prior probabilities $P(M_i)$. Although models are commonly assigned equal prior probabilities, in phylogenetics we may have prior beliefs stating that some models are more probable than others. For example, we have enough information about the process of mitochondrial sequence evolution to believe that the JC69 model is less probable in this case than the HKY85 model with a gamma distribution for rates among sites (see Yang, 1996a). Ideally, this information should be reflected in the model priors, and although considerable Bayesian research exists on eliciting prior information (Kadane and Wolfson, 1998; Madigan et al., 1995), it still seems be very difficult to quantify. Fortunately, if the signal in the data, conveyed through the likelihood, is strong enough, then the prior distributions should not have a large influence on the posterior distribution. Indeed, posterior probabilities of trees are already being used to estimate phylogenies (Holder and Lewis, 2003; Huelsenbeck et al., 2001, 2002; Larget and Simon, 1999; Mau and Newton, 1997; Mau et al., 1999; Yang and Rannala, 1997).

When the priors for the parameters in the complex model are very diffuse, Bayesian approaches tend to support the null model in contradiction to significance tests (e.g., LRTs) as sample size increases—the so called Jeffreys-Lindley's paradox (Bartlett, 1957; Jeffreys, 1939; Lindley, 1957; Shafer, 1982). If the diffuseness of these priors arises because of mere ignorance of the values these parameters can take, this conflict highlights a disadvantage of Bayesian approaches, especially in the case of Bayesian Information Criterion (BIC) (see below), which assume flat, improper priors. In any case, Jeffreys-Lindley's paradox illustrates the relevance, for good or

for bad, of the priors we choose for the model parameters (Huelsenbeck et al., 2002). Moreover, in some situations Bayesian approaches and standard significance tests can also be irreconcilable when testing point (or sharp) null hypotheses, for example, $H_0$: $ti/tv = 0.5$ versus $H_1$: $ti/tv \neq 0.5$ (Berger and Sellke, 1987) ($ti/tv$ is the transition/transversion ratio).

### *Bayesian Information Criterion*

In order to calculate model likelihoods, Bayesian methods often require computationally intensive techniques like Markov chain Monte Carlo (Gilks et al., 1996; Hastings, 1970; Metropolis et al., 1953). Although easy to implement, Bayes factor calculations do exist for some nested models via the Savage-Dickey ratio (Suchard et al., 2001; Verdinelli and Wasserman, 1995). However, there is a computationally more tractable approach, the Bayesian Information Criterion (BIC) (Schwarz, 1978):

$$\text{BIC} = -2\ell + K \log n$$

where $K$ is the number of estimable parameters, and $n$ is the sample size (for now we assume that $n$ can be approximated by the total number of characters in the alignment). The BIC was developed as an approximation to the log marginal likelihood of a model, and therefore, the difference between two BIC estimates may be a good approximation to the natural log of the Bayes factor (Kass and Wasserman, 1995). Given equal priors for all competing models, choosing the model with the smallest BIC is equivalent to selecting the model with the maximum posterior probability. The BIC assumes that the (parameters) prior is the *unit information prior* (i.e., a multivariate normal prior with mean at the maximum likelihood estimate and variance equal to the expected information matrix for one observation) (Kass and Wasserman, 1995), which can be thought of as a prior distribution that contains the same amount of information as a single, typical observation. This prior is quite diffuse, so the BIC tends to select models that are less complex than Bayes factors (for discussion see Raftery, 1999; Weakliem, 1999), and if $n > 8$, the BIC selects simpler models than the AIC (Forster and Sober, 2004). However, Burnham and Anderson (2003, pp. 302–305) suggest that the BIC can be used more generally with any prior.

A collection of BIC statistics contains the same information as a collection of pairwise Bayes factors. However, when choosing among several models, the BIC statistics are easier to interpret by visual inspection, as they allow for the simultaneous comparison of multiple models, so the best-fit models can be immediately identified. On the other hand, selecting the best-fit model from a collection of multiple pairwise Bayes factors could be more burdensome, and such procedure might suffer from some of the problems described above for the hLRTs. Nevertheless, the BIC approximation might not be appropriate when the posterior mode occurs at the boundary of the parameter space (Hsiao, 1997; Ota et al., 2000).

### *Decision Theoretic Approaches*

Recently, Minin et al. (2003) applied decision theory (Bernardo and Smith, 1994) to develop a novel model selection strategy (the DT method) that extends the BIC. Minin et al. (2003) argue that there is no guarantee that the best-fit models will produce the best estimates of phylogeny, and therefore propose a model selection method that incorporates some measure of phylogenetic performance. They assess models through a penalty or loss function, related to how dissimilar the branch length estimates are across models, and pick the model with the minimum posterior loss. As expected, simulations suggested that models selected with this criterion result in slightly more accurate branch length estimates than those obtained under models selected by the hLRTs.

### *Model Selection Uncertainty*

Once we have selected a model it is very important that we are able to assess how confident we are in that selection (see Chatfield, 1995). We would like to be able to rank the models and to know whether the model selected is much better than the other candidate models. At the same time, we should be interested to learn whether we would select the same model if several other independent samples were available. The assessment of model selection uncertainty has a long tradition within the Bayesian community and posterior probabilities can be naturally used to take account of model uncertainty (Kass and Raftery, 1995; Madigan and Raftery, 1994). For example, models can be ranked according to their posterior probabilities and 95% credible intervals (Occam's Window) can easily be constructed by summing these probabilities (Madigan and Raftery, 1994). Although computing posterior probabilities can be hard and time consuming, in theory we could approximate those probabilities with the BIC. Furthermore, we could also use the BIC values or posterior risks of the DT method (Minin et al., 2003) in the same way that we use the AIC below above to assess model selection uncertainty, although this could be considered ad hoc (see Hoeting et al., 1999).

### *Model Averaging*

Although in general model selection is concerned with the selection of just the best fit model, Bayesian approaches allow us to make inferences based on the entire set of candidate models, or *model averaging* (Hoeting et al., 1999; Madigan and Raftery, 1994; Raftery, 1996; Wasserman, 2000). Indeed, obtaining model averaged phylogenetic estimates is straightforward (Posada, 2003). If we consider, for example, $G$ models that include the gamma distribution for rate variation among sites (Yang, 1996a), the overall posterior mean of the shape of the gamma distribution ($\alpha$) would be:

$$E(\alpha \mid D) = \sum_{i=1}^{G} \hat{\alpha}_i P(M_i \mid D)$$

where $\hat{\alpha}_i$ is the estimate of $\alpha$ for model $i$.

Because not all parameters have the same interpretation across models, we should be careful when calculating and interpreting model-averaged parameter estimates. For example, the gamma shape parameter describing among-site rate variation has a different interpretation depending on whether the model also includes a proportion of invariable sites, because in such a case only the rates at variable sites, and not at all sites, are gamma-distributed. To facilitate a correct interpretation we could obtain two separate model-averaged estimates of the gamma shape parameter, one from models that include a proportion of invariable sites, and another from models that do not include a proportion of invariable sites. Moreover, from the above formulation we can see that it would be easy to estimate the *relative importance* of any parameter by summing the posterior probabilities across all models that included the parameters we are interested in. For example, the relative importance ($w_+$) for the shape of the gamma distribution across *all* candidate models is simply:

$$w_+(\alpha) = \sum_{i=1}^{R} P(M_i \mid D) I_\alpha(M_i)$$

where

$$I_\alpha(M_i) = \begin{cases} 1 \text{ if } \alpha \text{ is in model } M_i \\ 0 \text{ otherwise} \end{cases}$$

We also need to be careful when interpreting the relative importance of parameters. When the number of candidate models is less than the number of possible combinations of parameters, the presence-absence of some pairs of parameters can be correlated, and so their relative importances. In other words, if parameter $\varepsilon$ actually has a high relative importance, then a second parameter $\eta$ might yield a high relative importance simply because the presence-absence of parameters $\varepsilon$ and $\eta$ among models is positively correlated. For the 56 models in Table 1, the presence of the different base frequencies parameters ($\pi$) is completely correlated, whereas the presence of several substitution rates ($\varphi$) show complete or high levels of correlation. The presence of parameter $\kappa$ is inversely correlated with that of several substitution rate parameters (e.g., $\varphi_{A-G}$). The presence of $\alpha$, the shape of the gamma distribution for rate variation among sites, or $p_{inv}$, the proportion of invariable sites, is not correlated with that of any other parameter.

Indeed, the averaged parameter could be the topology itself, so we could construct a model-averaged estimate of phylogeny. We will come back to this later.

TABLE 1. $AIC_c$ values, $AIC_c$ differences ($\Delta$), and Akaike weights ($w$) for the carabid beetles *Ohomopterus* mitochondrial DNA data set from Sota and Vogler (2001). Because branch lengths were estimated for each candidate model, the number of branches was included in the penalty parameter $K$ (= number of parameters). $\ell$ are the maximized log likelihoods and Cum($w$) are the cumulative Akaike weights.

| Model | $\ell$ | $K$ | $AIC_c$ | $\Delta AIC_c$ | $w$ | Cum($w$) |
|---|---|---|---|---|---|---|
| TN93+I+Γ | 5441.4600 | 78 | 11045.5888 | 0.0000 | 0.5221 | 0.5221 |
| TIM+I+Γ | 5441.3765 | 79 | 11047.5965 | 2.0077 | 0.1913 | 0.7134 |
| HKY85+I+Γ | 5443.6729 | 77 | 11047.8422 | 2.2534 | 0.1692 | 0.8826 |
| K81uf+I+Γ | 5443.5566 | 78 | 11049.7821 | 4.1934 | 0.0641 | 0.9468 |
| GTR+I+Γ | 5440.9150 | 81 | 11051.0301 | 5.4413 | 0.0344 | 0.9811 |
| TVM+I+Γ | 5442.7393 | 80 | 11052.4991 | 6.9103 | 0.0165 | 0.9976 |
| TN93+Γ | 5448.6792 | 77 | 11057.8549 | 12.2661 | 0.0011 | 0.9988 |
| HKY85+Γ | 5450.5068 | 76 | 11059.3402 | 13.7514 | 0.0005 | 0.9993 |
| TIM+Γ | 5448.6577 | 78 | 11059.9843 | 14.3955 | 0.0004 | 0.9997 |
| K81uf+Γ | 5450.4883 | 77 | 11061.4730 | 15.8843 | 0.0002 | 0.9999 |
| GTR+Γ | 5448.0298 | 80 | 11063.0802 | 17.4914 | 0.0001 | 1.0000 |
| TVM+Γ | 5449.6685 | 79 | 11064.1804 | 18.5917 | 0.0000 | 1.0000 |
| TN93+I | 5470.7568 | 77 | 11102.0102 | 56.4214 | 0.0000 | 1.0000 |
| TIM+I | 5470.7417 | 78 | 11104.1522 | 58.5635 | 0.0000 | 1.0000 |
| GTR+I | 5470.3452 | 80 | 11107.7110 | 62.1223 | 0.0000 | 1.0000 |
| HKY85+I | 5476.8496 | 76 | 11112.0257 | 66.4370 | 0.0000 | 1.0000 |
| K81uf+I | 5476.8208 | 77 | 11114.1381 | 68.5493 | 0.0000 | 1.0000 |
| TVM+I | 5476.1650 | 79 | 11117.1736 | 71.5849 | 0.0000 | 1.0000 |
| F81+I+Γ | 5769.1118 | 76 | 11696.5501 | 650.9614 | 0.0000 | 1.0000 |
| F81+Γ | 5782.0566 | 75 | 11720.2721 | 674.6834 | 0.0000 | 1.0000 |
| F81+I | 5807.4927 | 75 | 11771.1442 | 725.5554 | 0.0000 | 1.0000 |
| GTR | 5805.0576 | 79 | 11774.9588 | 729.3700 | 0.0000 | 1.0000 |
| TVM | 5808.4727 | 78 | 11779.6141 | 734.0254 | 0.0000 | 1.0000 |
| TIM | 5810.4102 | 77 | 11781.3168 | 735.7280 | 0.0000 | 1.0000 |
| TN93 | 5813.4780 | 76 | 11785.2825 | 739.6938 | 0.0000 | 1.0000 |
| K81uf | 5813.5190 | 76 | 11785.3646 | 739.7758 | 0.0000 | 1.0000 |
| HKY85 | 5816.5894 | 75 | 11789.3375 | 743.7488 | 0.0000 | 1.0000 |
| SYM+I+Γ | 5861.0859 | 78 | 11884.8407 | 839.2520 | 0.0000 | 1.0000 |
| TVMef+I+Γ | 5867.6128 | 77 | 11895.7221 | 850.1333 | 0.0000 | 1.0000 |
| SYM+Γ | 5876.7803 | 77 | 11914.0570 | 868.4683 | 0.0000 | 1.0000 |
| TVMef+Γ | 5884.4272 | 76 | 11927.1810 | 881.5922 | 0.0000 | 1.0000 |
| TIMef+I+Γ | 5885.0684 | 76 | 11928.4632 | 882.8745 | 0.0000 | 1.0000 |
| K81+I+Γ | 5893.7642 | 75 | 11943.6872 | 898.0984 | 0.0000 | 1.0000 |
| TN93ef+I+Γ | 5897.7529 | 75 | 11951.6647 | 906.0759 | 0.0000 | 1.0000 |
| TIMef+Γ | 5899.2588 | 75 | 11954.6764 | 909.0877 | 0.0000 | 1.0000 |
| K80+I+Γ | 5906.2329 | 74 | 11966.4593 | 920.8706 | 0.0000 | 1.0000 |
| K81+Γ | 5908.7876 | 74 | 11971.5687 | 925.9800 | 0.0000 | 1.0000 |
| TN93ef+Γ | 5911.5659 | 74 | 11977.1254 | 931.5366 | 0.0000 | 1.0000 |
| SYM+I | 5908.7021 | 77 | 11977.9008 | 932.3120 | 0.0000 | 1.0000 |
| TVMef+I | 5917.6128 | 76 | 11993.5521 | 947.9633 | 0.0000 | 1.0000 |
| K80+Γ | 5920.9038 | 73 | 11993.6382 | 948.0494 | 0.0000 | 1.0000 |
| TIMef+I | 5928.9629 | 75 | 12014.0846 | 968.4959 | 0.0000 | 1.0000 |
| K81+I | 5938.0137 | 74 | 12030.0209 | 984.4321 | 0.0000 | 1.0000 |
| TN93ef+I | 5940.7383 | 74 | 12035.4701 | 989.8813 | 0.0000 | 1.0000 |
| K80+I | 5949.5186 | 73 | 12050.8677 | 1005.2789 | 0.0000 | 1.0000 |
| F81 | 6088.2227 | 74 | 12330.4388 | 1284.8501 | 0.0000 | 1.0000 |
| JC69+I+Γ | 6101.2656 | 73 | 12354.3618 | 1308.7730 | 0.0000 | 1.0000 |
| JC69+Γ | 6114.8408 | 72 | 12379.3515 | 1333.7628 | 0.0000 | 1.0000 |
| JC69+I | 6142.1719 | 72 | 12434.0137 | 1388.4249 | 0.0000 | 1.0000 |
| SYM | 6170.8916 | 76 | 12500.1097 | 1454.5209 | 0.0000 | 1.0000 |
| TVMef | 6190.3394 | 75 | 12536.8375 | 1491.2488 | 0.0000 | 1.0000 |
| TIMef | 6194.5806 | 74 | 12543.1547 | 1497.5659 | 0.0000 | 1.0000 |
| TN93ef | 6210.6353 | 73 | 12573.1011 | 1527.5123 | 0.0000 | 1.0000 |
| K81 | 6214.1152 | 73 | 12580.0610 | 1534.4723 | 0.0000 | 1.0000 |
| K80 | 6230.2100 | 72 | 12610.0898 | 1564.5011 | 0.0000 | 1.0000 |
| JC69 | 6411.5161 | 71 | 12970.5438 | 1924.9551 | 0.0000 | 1.0000 |

## AKAIKE INFORMATION CRITERION

A different approach to model selection is the Akaike Information Criterion (AIC) (Akaike, 1973, 1974; and see Sakamoto et al., 1986). The AIC is an asymptotically unbiased estimator of the *expected* relative Kullback-Leibler information quantity or distance (K-L) (Kullback and Leibler, 1951), which represents the amount of information lost when we use model $g$ to approximate model $f$

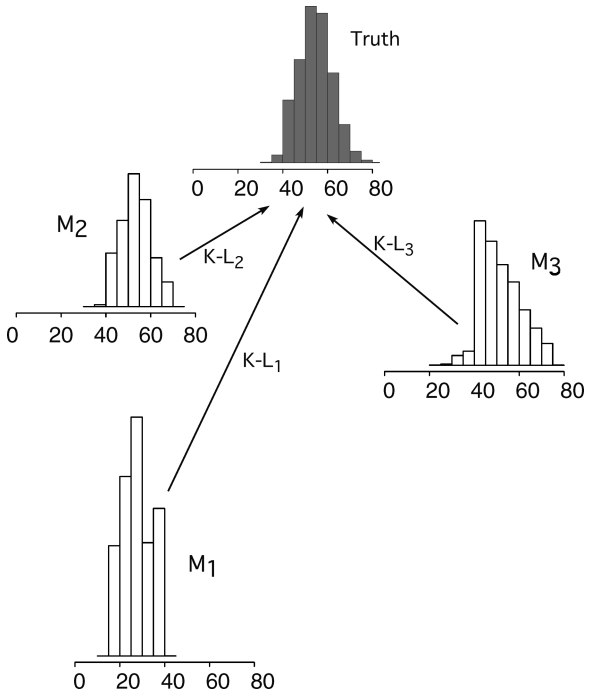


FIGURE 4. The Kullback-Leibler distance. The K-L distance aims to represent how close a model is to the truth. Here, $M_2$ is the candidate model that best approximates truth and therefore it is the model with the smallest K-L distance. The AIC chooses the candidate model with the smallest *expected* K-L distance.

(Fig. 4):

$$K\text{-}L = I(f,g) = \int f(x)\log\left(\frac{f(x)}{g(x \mid \theta)}\right)dx \,^2$$

The AIC for a given model is a function of its maximized log-likelihood ($\ell$) and the number of estimable parameters ($K$):

$$\mathrm{AIC} = -2\ell + 2K$$

In the context of phylogenetics we can think of the AIC as the amount of information lost when we use, say HKY85, to approximate the real process of nucleotide substitution. Hence, we prefer the model with the smallest AIC. The second term $K$ includes the parameters from the substitution model, like base frequencies, substitution rates, proportion of invariable sites, or rate variation among sites. If branch lengths are estimated de novo for every model, $K$ should also include the number of branches (for an unrooted bifurcated tree, twice the number of taxa minus three). Although the inclusion of the number of branches, constant for all models, does not change the order of the AIC values, it will change their relative magnitude.

In the AIC, as more parameters are added to the model the first term becomes smaller, representing an increased fit, whereas the second component, or penalty term, becomes larger. Indeed, when the sample is large, the number of adjustable parameters makes a negligible difference, and more complex models will be favored (Forster and Sober, 1994). It is important to note that although the AIC formula appears to be superficially very simple, its derivation is well founded on information theory (deLeeuw, 1992), and the so called "penalty term" $2K$ is not an arbitrary value (Burnham and Anderson, 2003, pp. 64). When sample size ($n$) is small compared to the number of parameters (say, $n/K < 40$) the use of a second-order AIC, $\mathrm{AIC}_c$ (Hurvich and Tsai, 1989; Sugiura, 1978), is recommended:

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2K(K+1)}{n-K-1}$$

where sample size is approximated by the total number of characters in the alignment (see below for discussion). Note that in this case the inclusion of branch lengths as estimated parameters can change the order of the $\mathrm{AIC}_c$ values, and therefore, the selected model.

Because the AIC is on a relative scale, it is critical to compute and present the *AIC differences* ($\Delta$AIC), rather than actual AIC values, over all candidate models (Buckley and Cunningham, 2002; Burnham and Anderson, 2003, pp. 70–72). For the $i$th model, the AIC difference is:

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

where min AIC is the smallest AIC value among all candidate models.

The AIC is designed to estimate the *predictive accuracy* of competing hypotheses (Forster, 2002; Sober, 2002b), which is the expected performance of a model when predicting new data. The prediction of new data is a common application in phylogenetics, for example in parametric bootstrapping or simulation studies. It seems that the AIC was first applied in the context of phylogenetics by Hasegawa and collaborators (1990a; 1990b; Kishino and Hasegawa, 1989), and although several phylogenetics programs implement the AIC, like MOLPHY (Adachi and Hasegawa, 1996) and MODELTEST (Posada, 2003; Posada and Crandall, 1998), the use of the AIC is much less common than that of the hLRTs.

The AIC makes several assumptions. First, there is the assumption of "uniformity of nature" (Forster and Sober, 1994), that is, that all data sets (future and past) are drawn from the same underlying process. Second, the AIC assumes that the sample size is large enough to ensure that the likelihood function will approximate its asymptotic properties. Finally the AIC assumes that the true distribution of parameter estimates, when the number of data $n$ is sufficiently large, follows a multivariate normal distribution. In principle, these assumptions (on the other hand, common in statistical phylogenetics) should not be unduly restrictive (Forster and Sober, 1994, 2004), but the implications of potential violations need to be studied. It has been argued that constraining parameters at

[2]For continuous functions.

their boundaries, for example setting the proportion of invariable sites to be zero, might violate the derivation of the AIC (and the BIC) (Ota et al., 2000).

### *Model Selection Uncertainty with the AIC*

The AIC differences allow for an immediate ranking of the candidate models. The larger the AIC difference for a model, the less probable that it is the best K-L model. As a rough rule of thumb, Burnham and Anderson (2003, p. 70) propose that models for which $\Delta_i \leq 2$ receive substantial support and are considered when making inferences, models having $4 \leq \Delta_i \leq 7$ have considerably less support, and models having $\Delta_i > 10$ receive no support. However, they also warn that these guidelines are not expected to hold when observations are not independent but are assumed so, as is usually the case in phylogenetics.

Akaike (1983) also suggested that the $\exp(-1/2\Delta_i)$ approximates the relative *likelihood of the models given the data*$L(M_i \mid D)$, which are then normalized to obtain a positive set of *Akaike weights* ($w$). The Akaike weight for the $i$th model in a set of $R$ candidate models is:

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

Akaike weights are very useful for assessing model-selection uncertainty without having to use computer intensive methods like Monte Carlo simulation or bootstrapping (Buckland et al., 1997; see Buckley et al., 2002, for an example). We can establish a 95% confidence set of models for the best K-L model by summing the Akaike weights from largest to smallest until the sum is just 0.95; the corresponding subset of models is a type of confidence set on the best K-L model (Burnham and Anderson, 1998, pp. 169–171; 2003). We can also assess the relative likelihoods of model $i$ versus model $j$ as simply the ratio of the two Akaike weights, which are called *evidence ratios* (Anderson et al., 2000; Burnham and Anderson, 2003, pp. 77–79). Techniques exist to compare whether two AICs differ significantly (Linhart, 1988; Shimodaira, 1997; Vuong, 1989), and multiple comparison techniques can be used to construct a confidence set of models that minimize the sampling error of the AIC (Shimodaira, 1998). Such techniques have already been proposed to construct a confidence sets of trees (Shimodaira, 2001; Shimodaira and Hasegawa, 1999).

There is a Bayesian basis for interpreting the Akaike weights as being the probability that a model is the expected best K-L model (Akaike, 1981). In fact, the Akaike weights can be generalized to also include prior information ($\rho_i$):

$$w_i = \frac{\mathcal{L}(M_i \mid D)\rho_i}{\sum_{r=1}^{R} \mathcal{L}(M_r \mid D)\rho_r}$$

(Burnham and Anderson, 2003, p. 76). However, the above is not a true Bayesian approach, because these priors only refer to the model, and not to the prior probability distribution of the parameters of the model. Neither do these priors refer to the belief that $M_i$ is the true model, but rather to the belief that model $M_i$ is the best K-L model for the data (Burnham and Anderson, 1998, 2003). Usually $\rho_i$ is set to $1/R$ for every model.

### *Model Averaging with the AIC*

Within the AIC framework, it is straightforward to obtain a *model-averaged estimate* of any parameter (Posada, 2003). For example, a model-averaged estimate of the substitution rate between adenine and cytosine ($\varphi_{A-C}$) using the Akaike weights ($w$) for $R$ candidate models would be:

$$\hat{\bar{\varphi}}_{A-C} = \frac{\sum_{i=1}^{R} w_i I_{\varphi_{A-C}}(M_i)\varphi_{A-Ci}}{w_+(\varphi_{A-C})}$$

where

$$w_+(\varphi_{A-C}) = \sum_{i=1}^{R} w_i I_{\varphi_{A-C}}(M_i)$$

and

$$I_{\varphi_{A-C}}(M_i) = \begin{cases} 1 \text{ if } \varphi_{A-C} \text{ is in model } M_i \\ 0 \text{ otherwise} \end{cases}$$

Again, the caveats described above about interpreting model-averaged parameter estimates apply. Likewise, it is again easy to estimate the relative importance of any parameter by summing the Akaike weights across all models that include the parameters we are interested in. For example, the relative importance of the substitution rate between adenine and cytosine across all candidate models is simply the denominator above, $w_+(\varphi_{A-C})$.

## MODEL-AVERAGED ESTIMATION OF PHYLOGENIES

As discussed above, model averaging can also be applied to the estimation of phylogenetic trees (Posada, 2003). This can be easily accomplished in programs like PAUP* (Swofford, 1998), and perhaps the only limitation is the time we want to dedicate to the analysis. We start by estimating a tree for each candidate model and then build a consensus tree using model weights as tree weights (these model weights can be Akaike weights, BIC weights, or model likelihoods from a Bayesian analysis) (see Jermiin et al., 1997). In a Bayesian framework one could also directly obtain a model-averaged estimate of phylogeny by using reversible-jump MCMC, an algorithm that moves through both parameter and model space (Green, 1995), and very recently implemented by Huelsenbeck et al. (2004), for phylogenetic model selection. It is also interesting to note that the AIC and Bayesian approaches allow for the direct comparison of trees estimated under different models because

likelihoods calculated on different trees and on different models are comparable (e.g., ML-JC69 versus ML-HKY). In this sense, the AIC has already been used as an extension of the likelihood optimality criterion for phylogenetic estimation (Kishino and Hasegawa, 1989; Ogishima et al., 2000; Sober, 2002b; Sober and Steel, 2002; Tanaka et al., 1999), and nothing prevents the BIC from also being considered as another phylogenetic criterion. Posterior probabilities for different trees inferred under different models are also directly comparable if they fall under the same posterior distribution.

We have applied AIC-based model averaging to 37 mitochondrial DNA sequences from the subgenus *Ohomopterus* (genus *Carabus*) ground beetles described by Sota and Vogler (2001). This alignment contains 1927 sites, 301 of which are variable. We took three approaches to selecting the best-fit model. First, we optimized the likelihood and model parameters for the 56 substitution models currently implemented in the program MODELTEST (Posada and Crandall, 1998) on a neighbor-joining tree estimated from Jukes and Cantor (1969) distances. We then used the AIC and $AIC_c$ to select the best-fit model from these likelihoods. Second, we took these model parameters and performed a tree search under each of the 56 models so as to find the tree with the highest likelihood under each of these optimized models. Again, the AIC and $AIC_c$ was used to chose the best-fit model. The second approach is superior to the first approach because it involves a more thorough search for the maximum likelihood under each model; however, the computational burden is much greater. Third, we also used the specific hLRT strategy implemented in MODELTEST (Posada and Crandall, 1998). From the likelihood values we calculated $AIC_c$ values, Akaike weights, the relative importance of different parameters, and model averaged estimates of parameters and topology. In addition, we performed a bootstrap analysis on the data using the best $AIC_c$ model with 500 replicates. All tree searches used five random addition replicates followed by TBR branch swapping. All likelihood calculations and tree searches were performed using PAUP*4.0b10 (Swofford, 2000).

Examining the $AIC_c$ values and Akaike weights for the models optimized on the NJ tree we immediately observe that only 11 out of the 56 models received noticeable support from the data (Table 1). Importantly, this confidence set of models, and the ranking of models within this set is almost identical to that obtained from optimizing the topology (data not shown) (see also Nylander, 2004). All of the supported models incorporated the gamma distribution for among site rate variation and the best-supported models also included a proportion of invariable sites. Models that assumed equal base frequencies fitted the data poorly and received essentially no support (i.e., their Akaike weights are close to zero). The TN93+I+$\Gamma$ model had the smallest $AIC_c$ value, but there was considerable uncertainty in identifying the most appropriate number of different substitution rates between nucleotides. The Akaike weights calculated from the $AIC_c$ values were very similar to those calculated from the AIC. This is because the $n/K$ ratio, 37.14, is close to the value of 40, which Burnham and Anderson (2003, p. 66) recommend as the cut-off for preferring $AIC_c$. Indeed, when $n/K$ is relatively large the $AIC_c$ converges back to the AIC, and so it is still appropriate to use the $AIC_c$ instead of the AIC. The hLRT approach led to selection of the HKY+I + $\Gamma$ model, which only received an $AIC_c$ weight of 0.1692 (Table 1), but was contained within the 95% AIC confidence set of models. The ML tree under the HKY+I + $\Gamma$ model differs by a symmetrical distance (Foulds et al., 1979) of 4 and 5 from the two trees estimated under the TN93+I + $\Gamma$ model.

In total 23 unique tree topologies were estimated from all of the models; however, only 8 unique topologies were contained in the set of trees that were estimated from models that received greater than or equal to 0.00001 support from the $AIC_c$ weights. Some tree searches under the among-site rate variation models recovered two topologies, where one of these topologies had an internal branch collapsed to zero length. The weighted $AIC_c$ consensus topology (Fig. 5A) was almost identical to the topology estimated under the best $AIC_c$ model (TN93+I + $\Gamma$) (Fig. 5B), but due to the model selection uncertainty there is considerable ambiguity in selecting the best point estimate of topology for these data. The bootstrap analysis under the best $AIC_c$ model indicates that the nodes that are not supported under all of the models also have low bootstrap support (Fig. 5). This observation is important because it suggests that in this case if we had ignored model selection uncertainty our conclusion as to what hypotheses were well supported by the data would be the same. It is worth mentioning that the numbers above branches in Figure 5A describe the uncertainty of branches due to uncertainty on the models of molecular evolution. This is in contrast with the bootstrap values in Figure 5B, which describe uncertainty due to the stochasticity of molecular evolution. The former numbers can be regarded as "bootstrap proportions" obtained by resampling models with probabilities proportional to the Akaike weights. The phylogenetic relationships among the *Ohomopterus* carabid beetles are very similar to those estimated by Sota and Vogler (2001) using maximum parsimony.

We examined the association between pairwise $AIC_c$ differences and pairwise tree distances (Foulds et al., 1979) for the 11 models included in the 99% confidence set (Fig. 6). This relationship shows a weak but significant correlation ($r^2 = 0.2394$; $P = 0.00015$) between the improvement of fit of a model to the data and differences in topology. This graph supports, to a limited extent, the intuition that models with similar fits to the data tend to support similar trees.

The model averaged parameter estimates are very similar to the maximum likelihood estimates under the best-fit models (Table 2) because models with similar likelihoods, and thus low AIC differences tend to result in similar parameter estimates. The variability between the model averaged and best-fit model parameter estimates is unlikely to have a large effect on estimation of

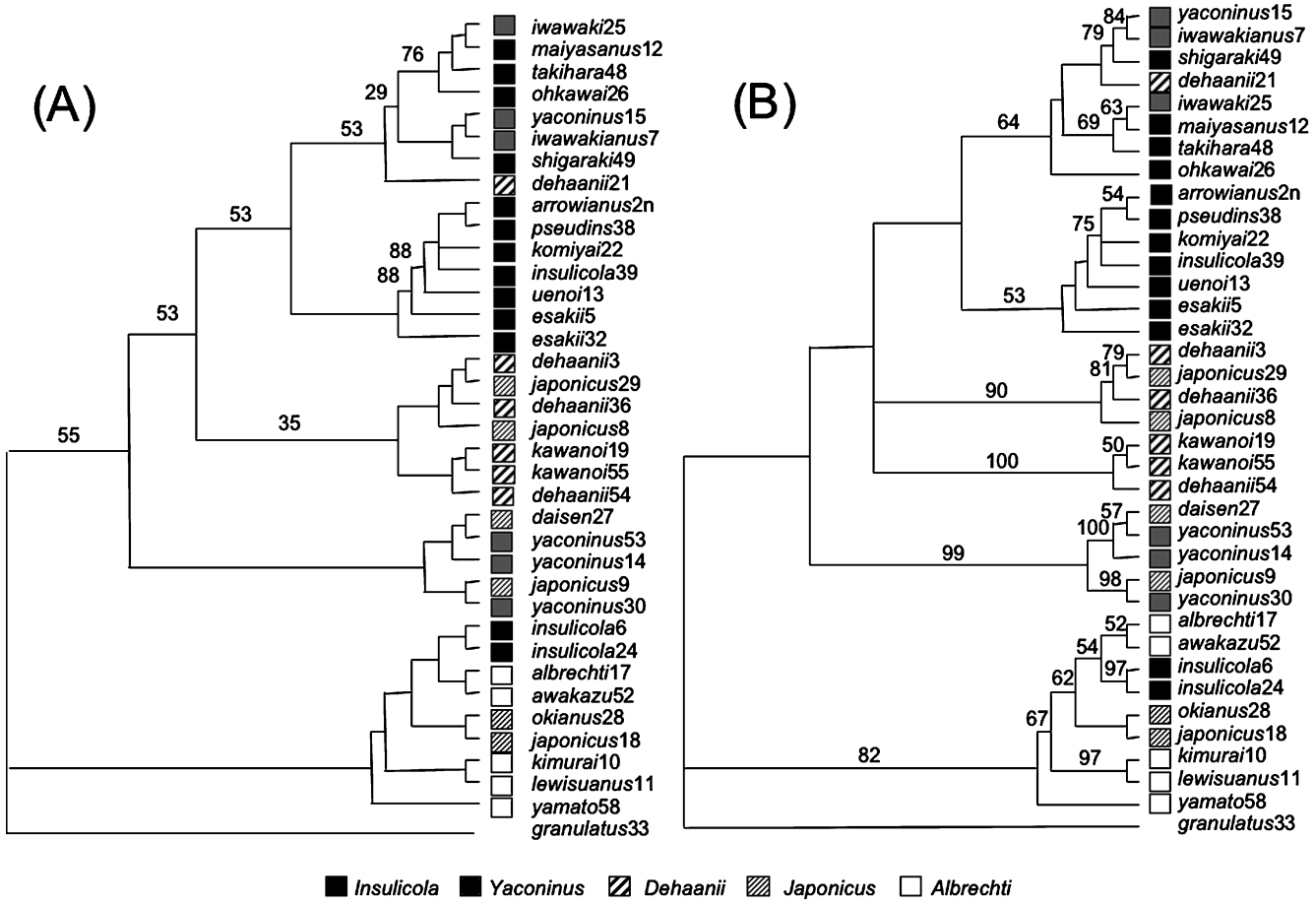


FIGURE 5. Multimodel phylogeny of *Ohomopterus* carabid beetles. (A) Consensus of trees estimated under 56 candidate models, and constructed using Akaike weights (with the $AIC_c$) as tree weights. The values above branches represent the weights for each branch. All branches without a number received a weight of 100%. (B) Consensus of the two maximum likelihood trees under the best $AIC_c$ model (TN93+I+Γ), one of which had a branch of zero length. Numbers above nodes are nonparametric bootstrap proportions. Nodes that received less than 50% are not indicated. The five species groups are indicated by shaded boxes.

topology. The greatest variability between the model averaged parameter and best-fit model parameter estimates is observed for the transversion rate parameters. This is not surprising given that relatively few transversions have occurred in these data and therefore there is not much information from which to gain stable estimates.

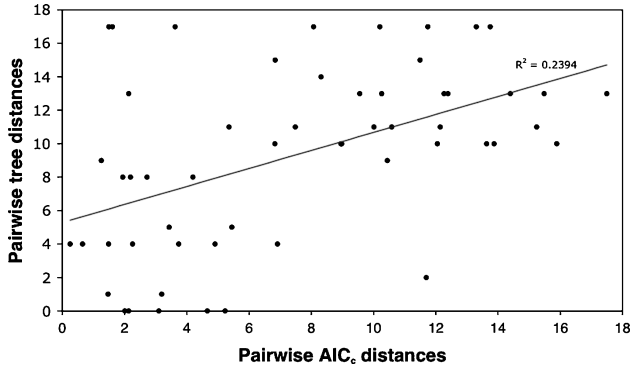


FIGURE 6. AIC differences and phylogeny estimation. For each pair of models out of the 11 models with noticeable $AIC_c$ support, we calculated the differences in AIC scores (Pairwise $AIC_c$ distances) and the Robinson and Foulds (1981) tree distances (Pairwise tree distances) using $AIC_c$ scores calculated on a NJ-JC tree.

Not all model parameters have the same *importance* for this data set (Table 3). The alpha shape parameter from the gamma distribution of among-site rate variation and the base frequency parameters have a relative importance of 1.0 because they appear in all of the supported models. The proportion of invariable sites is also a very important parameter although a few models with low weight without this parameter are supported. This observation suggests that these properties of the evolutionary process are very important for obtaining a good model fit. The $\varphi_{A-G}$ and $\varphi_{C-T}$ substitution rate parameters have higher relative importance values that the transversion parameters. This indicates that for these data it is important to allow the two transition types to have different rates, more so than the transversion types. The results shown in Table 2 make sense in light of our current knowledge of the dynamics of animal mitochondrial DNA evolution (e.g., Brown et al. 1982; Tamura and Nei 1993; Buckley et al. 2001a).

TABLE 2. Model-averaged estimates of nucleotide substitution parameters. These estimates were obtained from the carabid beetles *Ohomopterus* mitochondrial DNA data set using the Akaike weights ($w_i$) derived from the $AIC_c$ for models with $w_i > 0.0001$. Which estimates contributed from which models are indicated in Table 3. Included also are the estimates corresponding to the best $AIC_c$ model (TN93+I+Γ) and to the model selected by the hLRT procedure (HKY85+I+Γ). $\pi_A - \pi_T$: base frequencies; $\kappa$: transition/transversion parameter; $\varphi_{A-C} - \varphi_{A-T}$: substitution rates; $\alpha$: shape of the gamma distribution for rate variation among sites; $\alpha$ (I+Γ) shape of the gamma distribution for rate variation among sites under an I+Γ model; $p_{inv}$ (I+Γ) proportion of invariable sites under an I+Γ model.

| Parameter | Model-averaged estimate | $AIC_c$ model estimate | hLRT model estimate |
|---|---|---|---|
| $\pi_A$ | 0.3330 | 0.3342 | 0.3303 |
| $\pi_C$ | 0.0683 | 0.0667 | 0.0725 |
| $\pi_G$ | 0.1362 | 0.1369 | 0.1335 |
| $\pi_T$ | 0.4625 | 0.4622 | 0.4637 |
| $\kappa$ | 14.8483 | 14.8476 | 14.8476 |
| $\varphi_{A-C}$ | 0.6290 | 1.0 | — |
| $\varphi_{A-G}$ | 13.4111 | 13.1823 | — |
| $\varphi_{A-T}$ | 1.0536 | 1.0 | — |
| $\varphi_{C-G}$ | 0.4189 | 1.0 | — |
| $\varphi_{C-T}$ | 20.0553 | 19.7583 | — |
| $\alpha$ | 0.1011 | — | — |
| $\alpha$(I+Γ) | 0.7149 | 0.7658 | 0.5849 |
| $p_{inv}$(I+Γ) | 0.6874 | 0.7038 | 0.6644 |

Lastly, model averaging could also be applied to other problems in evolutionary biology in which inferences can be drawn from several models, for example as in the detection of positive selection from sequence alignments (Yang et al., 2000), and the estimation of divergence times using relaxed molecular clocks (Aris-Brosou and Yang, 2002), where different models can frequently yield different results.

## PHILOSOPHICAL CONSIDERATIONS ON MODEL SELECTION

There is still an important philosophical debate about model selection in general (Burnham and Anderson, 1998, 2003; Forster and Sober, 1994, 2004; Forster, 2000, 2001; Kass and Raftery, 1995; Kieseppä, 2002; Myrvold and Harper, 2002; Popper, 1959; Sober, 2002a; Wasserman, 2000), and here we do not attempt to address all the issues, but just those we think are most relevant. The information-theoretic and the Bayesian approaches represent different philosophical approaches to the problem of model selection (Forster and Sober, 1994; Kuha, 2003; Sober, 2002a). The AIC is designed to choose the model that best *approximates* reality. The conclusions of AIC are never about the truth or falsity of a hypothesis, but about its closeness to the truth (Forster and Sober, 2004). On the other hand, Bayesian approaches are designed to *identify* the true model, given the data. Both the AIC and Bayesian approaches have been criticized on different grounds.

That Bayesian approaches are designed to identify the true model can be surprising when surely we know that all models of evolution are false (i.e., their probability is zero). The standard interpretation of $P(M_i|D)$ is that it is the probability that $M_i$ is the true model given the data, even though we know that this statement is false a priori (Gelfand, 1996). A common response to this criticism is that we can hope that at least one of the models is approximately true, and that the posterior distributions allows us to compare the relative merits of the models (Wasserman 2000). On the other hand, it has been argued that the derivation of the BIC does not require that the true model is contained within the set of candidate models (Burnham and Anderson, 2003, pp. 293–295; Cavanaugh and Neath, 1999). Interestingly, it is possible to obtain the AIC as a Bayesian result if a particular prior (the so called K-L prior) is used with the BIC (Burnham and Anderson, 2003, pp. 302–305).

It has been alleged in the statistical literature that, under certain conditions, the BIC is statistically consistent (it does converge to truth as more data is added), whereas the AIC is not (but see Bozdogan, 1987; Findley, 1991; Keuzenkamp and McAleer, 1995; Nishii, 1984, 1988; Shibata, 1986; Woodroofe, 1982) but the relevance of statistical consistency in this context is not clear (Forster, 2002).

We can think of a model as a set or family of sharp hypotheses. For example, the K80 model contains all hypotheses representing different values of the transition/transversion parameter, $\kappa$. The JC69 model,

TABLE 3. Relative parameter importance. Included here are Akaike weights ($w_i$) and relative parameter importance values for the *Ohomopterus* carabid beetles mitochondrial DNA data set, for models with $w_i > 0.0001$. Where a model contains a free parameter it is indicated with a black dot (note that $\varphi_{G-T}$ is often set to equal 1).

| | $w_i$ | $\pi_A$ | $\pi_C$ | $\pi_G$ | $\pi_T$ | $\kappa$ | $\varphi_{A-C}$ | $\varphi_{A-G}$ | $\varphi_{A-T}$ | $\varphi_{C-G}$ | $\varphi_{C-T}$ | $\varphi_{G-T}$ | $\alpha$ | $p_{inv}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TN93+I+Γ | 0.5221 | • | • | • | • | | | • | | | • | | • | • |
| TIM+I+Γ | 0.1913 | • | • | • | • | | | • | | | • | | • | • |
| HKY85+I+Γ | 0.1692 | • | • | • | • | • | | | | | | | • | • |
| K81uf+I+Γ | 0.0642 | • | • | • | • | | | | | | | | • | • |
| GTR+I+Γ | 0.0344 | • | • | • | • | | • | • | • | • | • | • | • | • |
| TVM+I+Γ | 0.0165 | • | • | • | • | | • | | • | • | | • | • | • |
| TN93+Γ | 0.0011 | • | • | • | • | | | • | | | • | | • | |
| HKY85+Γ | 0.0005 | • | • | • | • | • | | | | | | | • | |
| TIM+Γ | 0.0004 | • | • | • | • | | | • | | | • | | • | |
| K81uf+Γ | 0.0002 | • | • | • | • | | | | | | | | • | |
| GTR+Γ | 0.0001 | • | • | • | • | | • | • | • | • | • | • | • | |
| *Relative parameter importance* | | 1.0 | 1.0 | 1.0 | 1.0 | 0.170 | 0.051 | 0.749 | 0.051 | 0.051 | 0.749 | 0.051 | 1.0 | 0.997 |

however, contains only one hypothesis, as all its parameters are fixed (equal base frequencies and equal rates for transitions or transversions). The AIC and the BIC work with maximized likelihoods, and therefore they are comparing the best point hypothesis within each model. However, it might be unwise to compare models based only on the merits of a single point, even if this point is optimal, and that is why Bayesians prefer models for which the sum of the likelihoods of all contained point hypotheses is largest (Holder and Lewis, 2003).

## WHICH MODEL SELECTION METHOD IS BEST FOR PHYLOGENETICS?

The use of different model selection strategies may lead to the selection of different models of evolution (Posada and Crandall, 2001a), and we know that model choice affects all aspects of phylogenetic analysis. Here we have attempted to compare different model selection strategies from a theoretical and practical point of view, in the context of phylogenetics. Previous Monte Carlo simulations on the performance of model selection in phylogenetics (Posada, 2001; Posada and Crandall, 2001b) showed that these methods work well when the aim is to *identify* the generating model. However, these simulations missed the point that the true model of evolution will never be one of the candidate models. It would be more useful to generate data from a model much more complex than any of the candidate models, and then study how well the selected models *approximate* this complex generating model (e.g., Minin et al., 2003). Clearly, we should seek models that are good approximations to the truth and from which therefore we can make valid inferences concerning the real process of molecular evolution. Too often we read expressions like "The best-fit model was selected with the program MODELTEST" without any reference to which model selection strategy was used (in this case, hLRT or AIC). When a method of model selection is used, this should be explicitly reported.

From the discussion above it should be clear that the Bayesian and AIC approaches present several important advantages over the hLRTs for model selection (see also Table 4). Namely, they are able to simultaneously compare multiple nested or nonnested models (see Chamberlain, 1890), account for model selection uncertainty, and allow for model-averaged inference. Although model selection uncertainty tools do not exist within the standard hLRTs framework, there are extensions of the LRT framework that allow for the specification of confidence sets of models. Evidence for a model can be also estimated by the "expected likelihood weights" (Strimmer, 2001; Strimmer and Rambaut, 2001). Criteria like the AIC or BIC are very simple to calculate from the maximum likelihood estimate, although they do rely on point estimates and do not take in account topological uncertainty (Bollback, 2002). The importance of the later effect has yet to be examined (but see Posada and Crandall, 2001b), as well as the potential impact of comparing models with parameters fixed at the boundary of their ranges (e.g., $\alpha = \infty$) in the AIC and BIC.

TABLE 4. Comparison of model selection strategies for phylogenetics. Indicated are what the authors think are good properties for a model section procedure. Exceptions to these may exist and the comments below are generalizations.

| Good properties for model selection methods | hLRT | Bayesian | AIC |
|---|---|---|---|
| *Applies easily to nonnested models* | No | Yes | Yes |
| *Allows for the simultaneous comparison of multiple models* | No | Yes | Yes |
| *Does not depend on a subjective significance level* | No | Yes[§] | Yes |
| *Incorporates topological uncertainty* | No | Yes[*] | No |
| *Easy to compute* | Yes | No[*] | Yes |
| *Assesses model selection uncertainty* | No | Yes | Yes |
| *Allows model averaging* | No | Yes | Yes |
| *Provides the possibility of specifying prior information for models* | No | Yes[*] | Yes |
| *Provides the possibility of specifying prior information for model parameters* | No | Yes[*] | No |
| *Designed to approximate, rather than to identify, truth* | No | No | Yes |

[*]Not the BIC.
[§]In a sense, the interpretation of Bayes factors could be considered as subjective.

The possibility of inferring model-averaging phylogenies will eliminate some of the criticisms that model-based methods are contingent on the single best-fit model selected. Obviously, the methods described above can facilitate model-averaged hypothesis testing, as one could test for the monophyly of a group by considering all models available. Sanderson and Kim (2000) already hinted at the possibility of model-averaging phylogenies, but claimed that such a composite solution would be computationally prohibitive. However, this computational burden will depend on the size of the data set (especially on the number of taxa) and the number of models considered (but one could work with the 95% confidence or credible set of models), and in some cases it will certainly be feasible.

Selecting a set of candidate models is not easy; there are 203 "standard" time-reversible models of nucleotide substitution, but model selection in phylogenetics is commonly limited to a subset of these (Huelsenbeck et al., 2004). Indeed, evaluating a large number of models is more problematic for the hLRT than for the AIC and Bayesian approaches for the reasons explained above. The implications of conditioning model selection on a subset of the possible set of models is currently unknown.

*Selection bias* (Zucchini, 2000) may occur when the number of candidate models is large. In such cases random fluctuations in the data will increase the score of some models more than others and therefore the chance that the best model won for spurious reasons increases. Indeed, the set of candidate models influences model

choice, and a careful a priori selection of candidate models is very important.

Both in the $AIC_c$ and the BIC descriptions above, the total number of characters was used as an estimate of sample size. However, effective sample sizes in phylogenetic studies are poorly understood, and depend on the quantity of interest (Churchill et al., 1992; Goldman, 1998; Morozov et al., 2000). Characters in an alignment will often not be independent, so using the total number of characters as a surrogate for sample size (Minin et al., 2003; Posada and Crandall, 2001b) could be an overestimate. Using only the number of variable sites as an estimate of sample size is a more conservative approach, but could be an underestimate (note that all sites are used when estimating base frequencies or the proportion of invariable sites). Indeed, sample size also depends on the number of taxa. Importantly, sample size can have an effect on the outcome of model selection with the $AIC_c$. In our example above, if we were to use the number of variable characters (301 sites) as the sample size, instead of the total number of characters (1927 sites), the best $AIC_c$ model would not change, but the second and third $AIC_c$ models would exchange their rankings. Furthermore, because the LRT, the AIC, and the BIC strategies rely on large sample asymptotics, it is also important to decide when a sample should be considered small. Although the $AIC_c$ was derived under Gaussian assumptions, Burnham et al. (1994) found that this second order expression performed well in product multinomial models for open population capture-recapture. Burnham and Anderson (2003, p. 66) suggest using this correction when the sample size is small compared to the number of adjustable parameters, $n/K < 40$. Alternatively, and because $AIC_c$ converges to the AIC with increasing $n/K$ ratios, one could always use the $AIC_c$ (D. Anderson, personal communications). Phylogenetic characters are mostly discrete, and the unconstrained model in phylogenetics is multinomial (Goldman, 1993). One may think of an alignment of nucleotide characters as a large and sparse contingency table with $4^T$ bins, where $T$ is the number of taxa. For large sample asymptotics to hold in a contingency table every cell should contain, in general, more than 5 observations (see Agresti, 1990, p. 49, 244–250), which gives a rule of thumb of $n/4^T > 5$. Clearly, more research is needed on sample size in phylogenetics.

Other model selection methods exist, like cross-validation and the bootstrap (see Browne, 2000; Efron and Tibshirani, 1993; Linhart and Zucchini, 1986), but they seem too time-consuming—note that cross validation is asymptotically equivalent to the AIC (Stone, 1977)—for the selection of substitution models. There is an important role for more general tests of model fit and accuracy within the process of model selection. For example, tests of base frequency stationarity (Rzhetsky and Nei, 1995; Van Den Bussche et al., 1998) should be standard before a phylogenetic analysis. In addition, the global tests of Goldman (1993) and Bollback (2001) are useful for detecting model misspecification. When tests such as these indicate that the final model selected still does not fit the data well, our results must be interpreted with caution as the possibility remains that some vital evolutionary process has not been accounted for, which could potentially be misleading.

Model selection is a useful tool for research, but it is not a substitute for careful thinking and common sense reasoning (Browne, 2000). There are examples in the phylogenetic literature where the best-fit models have led to phylogenetic estimates that are clearly incorrect (Buckley and Cunningham, 2002; Posada and Crandall, 2001c). Consideration of model selection uncertainty and multimodel inference should lead to equal or better estimates of phylogenies and substitution parameters, and we should see more applications of these ideas in the future (see also Nylander, 2004). Computation of AIC differences, Akaike weights, model-averaged estimates, and relative parameter importance is currently implemented in the program MODELTEST (Posada and Crandall, 1998). Further developments will allow for the simultaneous use of different models for different partitions of the data (Nylander et al., 2004; Pupko et al., 2002; Suchard et al., 2003a; Yang, 1996b). It is now time to start thinking about how we will select those. Model selection in phylogenetics is indeed still an open area for research (Huelsenbeck et al., 2002).

## ACKNOWLEDGEMENTS

We are undoubtedly indebted to Kenneth Burnham and David Anderson for their enlightening book. David Anderson, Elliot Sober, and Carsten Wiuf provided very insightful comments on the manuscript. Robert Weiss, Janet Sinsheimer, Paul Lewis, Paul Joyce, Hidetoshi Shimodaira, and Rissa Ota helped clarify some ideas on Bayesian model selection. Nick Goldman and two anonymous referees provided useful comments on a first version. Jeff Thorne, Hirohisa Kishino, and two anonymous referees provide very valuable comments that considerably improved the manuscript. Thanks to David Swofford and Jack Sullivan for many valuable conversations on model selection throughout the years. DP was funded by the Spanish Ministry of Science and Technology, while funding for TRB was provided by the New Zealand Foundation for Research, Science, and Technology.

## REFERENCES

Adachi, J., and M. Hasegawa. 1996. MOLPHY version 2.3: Programs for molecular phylogenetics based in maximum likelihood. Comput. Sci. Monogr. 28:1–150.

Agresti, A. 1990. Categorical data analysis, 2nd edition. Wiley, New York.

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

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Aut. Control 19:716–723.

Akaike, H. 1981. Likelihood of a model and information criteria. J. Econometrics 16:3–14.

Akaike, H. 1983. Information measures and model selection. Int. Stat. Inst. 22:277–291.

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: Problems, prevalence, and an alternative. J. Wildl. Manage 64:912–923.

Aris-Brosou, S., and Z. Yang. 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. Syst. Biol. 51:703–714.

Bartlett, M. S. 1957. A comment on D. V. Lindley's statistical paradox. Biometrika 44:533–534.

Berger, J. O., and T. Sellke. 1987. Testing a point null hypothesis: The irreconcilability of *P* values and evidence. J. Am. Stat. Assoc. 82:112–122.

Bernardo, J. M., and A. F. M. Smith. 1994. Bayesian theory. Wiley and Sons, New York.

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

Box, G. E. P. 1976. Science and statistics. J. Am. Stat. Assoc. 71:791–799.

Bozdogan, H. 1987. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika 52:345–370.

Browne, M. 2000. Cross-validation methods. J. Math. Psychol. 44:108–132.

Bruno, W. J., and A. L. Halpern. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol. Biol. Evol. 16:564–566.

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.

Buckland, S. T., K. P. Burnham, and N. H. Augustin. 1997. Model selection uncertainty: An integral part of inference. Biometrics 53:603–618.

Buckley, T. R., and C. W. Cunningham. 2002. The effects of nucleotide substitution model assumptions on 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: The 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. 1998. Model selection and inference: A practical information-theoretic approach, 1st ed. Springer-Verlag, New York.

Burnham, K. P., and D. R. Anderson. 2003. Model selection and multimodel inference: A practical information-theoretic approach, 2nd ed. Springer-Verlag, New York.

Burnham, K. P., D. R. Anderson, and G. C. White. 1994. Evaluation of the Kullback-Leibler discrepancy for model selection in open population capture-recapture models. Biometrica J. 36:299–315.

Cavanaugh, J. E., and A. A. Neath. 1999. Generalizing the derivation of the Schwarz information criterion. Commun. Stat. Theory Methods 28:49–66.

Chamberlain, T. C. 1890. The method of multiple working hypotheses. Science 15:93.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. J. R. Stat. Soc. A 158:419–466.

Churchill, G. A., A. Von Haeseler, and W. C. Navidi. 1992. Sample size for a phylogenetic inference. Mol. Biol. Evol. 9:753–769.

Deleeuw, J. 1992. Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. Pages 599–609 *in* Breakthroughs in statistics (S. Kotz, and N. L. Johnson, eds.). Springer-Verlag, London.

Edwards, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge, UK.

Efron, B., and R. J. Tibshirani. 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

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

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

Felsenstein, J. 1981b. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol. J. Linnaean Soc. 16:183–196.

Findley, D. F. 1991. Counterexamples to parsimony and BIC. Ann. Inst. Stat. Math. 43:505–514.

Fisher, R. A. 1921. On the 'probable error' of a coefficient of correlation deduced from a small sample. Metron I, part 4:3–32.

Forster, M. R. 2000. Key Concepts in model selection: Performance and generalizability. J. Math. Psychol. 44:205–231.

Forster, M. R. 2001. The new science of simplicity. Pages 83–119 *in* Simplicity, inference and modeling (A. Zeller, H. A. Keuzenkamp, and M. McAleer, eds.). Cambridge University Press, Cambridge, UK.

Forster, M. R. 2002. Predictive accuracy as am achievable goal of science. Phil. Sci. 69:S124–S134.

Forster, M., and E. Sober. 1994. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. Br. J. Phil. Sci. 45:1–35.

Forster, M. R., and E. Sober. 2004. Why likelihood? *in* Likelihood and Evidence (M. Taper, and S. Lele, eds.). University of Chicago Press, Chicago.

Foulds, L. R., M. D. Hendy, and D. Penny. 1979. A graph theoretic approach to the development of minimal phylogenetic trees. J. Mol. Evol. 13:127–149.

Foutz, R. V., and R. C. Srivastava. 1977. The performance of the likelihood ratio test when the model is incorrect. Ann. Stat. 5:1183–1194.

Frati, F., C. Simon, J. Sullivan, and D. L. Swofford. 1997. Gene evolution and phylogeny of the mitochondrial cytochrome oxidase gene in Collembola. J. Mol. Evol. 44:145–158.

Gelfand, A. E. 1996. Model determination using sampling-based methods. Pages 145–161 *in* Markov chain Monte Carlo in practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). Chapman & Hall, London, New York.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (eds.) 1996. Markov chain Monte Carlo in practice. Chapman & Hall, London, New York.

Golden, R. M. 1995. Making correct statistical inferences using a wrong probability model. J. Math. Psychol. 38:3–20.

Goldman, N. 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. Syst. Zool. 39:345–361.

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

Goldman, N. 1998. Phylogenetic information and experimental design in molecular systematics. Proc. R. Soc. Lond. B Biol. Sci. 265:1779–1786.

Goldman, N., and S. Whelan. 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol. Biol. Evol. 17:975–978.

Green, P. J. 1995. Reversible jump MCMC computation and Bayesian model determination. Biometrika 92:711–732.

Hasegawa, M. 1990a. Mitochondrial DNA evolution in primates: Transition rate has been extremely low in the lemur. J. Mol. Evol. 31:113–121.

Hasegawa, M. 1990b. Phylogeny and molecular evolution in primates. Jpn. J. Genet. 65:243–266.

Hasegawa, M., K. Kishino, and T. Yano. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75:800–802.

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

Holder, M., and P. O. Lewis. 2003. Phylogeny estimation: Traditional and Bayesian approaches. Nat. Rev. Genet. 4:275–284.

Hsiao, C. K. 1997. Approximate Bayes factors when a mode occurs on the boundary. J. Am. Stat. Assoc. 92:656–663.

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., and N. S. Imennov. 2002. Geographic origin of human mitochondrial DNA: Accommodating phylogenetic uncertainty and model comparison. Syst. Biol. 51:155–165.

Huelsenbeck, J. P., B. Larget, and M. E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 21:1123–1133.

Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51:673–688.

Huelsenbeck, J. P., B. Rannala, and B. Larget. 2000. A Bayesian framework for the analysis of cospeciation. Evol. Int. J. Org. Evol. 54:352–364.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, 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.
Jeffreys, H. 1939. Theory of probability. Oxford University Press, Oxford.
Jermiin, L. S., G. J. Olsen, K. L. Mengersen, and S. Easteal. 1997. Majority-rule consensus of phylogenetic trees obtained by maximum-likelihood analysis. Mol. Biol. Evol. 14:1296–1302.
Johnson, J. B., and K. S. Omland. 2003. Model selection in ecology and evolution. Trends Ecol. Evol. 19:101–108.
Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. M. Munro, ed.) Academic Press, New York.
Kadane, J. B., and L. J. Wolfson. 1998. Experiencies in elicitation. J. R. Stat. Soc. D 47 Part 1:3–19.
Kass, R. E., and A. E. Raftery. 1995. Bayes factors. J. Am. Stat. Assoc. 90:773–795.
Kass, R. E., and L. Wasserman. 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J. Am. Stat. Assoc. 90:928–934.
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.
Kendall, M., and A. Stuart. 1979. The advanced theory of statistics, 4th edition. Charles Griffin, London.
Kent, J. T. 1982. Robust properties of likelihood ratio tests. Biometrika 69:19–27.
Keuzenkamp, H., and M. McAleer. 1995. Simplicity, scientific inference and economic modeling. Econ. J. 105:1–21.
Kieseppä, I. A. 2002. Statistical model selection and Bayesianism. Phil. Sci. 68:S141–S152.
Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.
Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Nat. Acad. Sci. USA 78:454–458.
Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170–179.
Kuha, J. 2003. AIC and BIC: Comparisons of assumptions and performance. Sociol. Methods Res. Submitted.
Kullback, S., and R. A. Leibler. 1951. On information and sufficiency. Ann. Math. Stat. 22:79–86.
Larget, B., and D. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.
Lindley, D. V. 1957. A statistical paradox. Biometrika 44:187–192.
Linhart, H. 1988. A test whether two AIC's differ significantly. S. Afr. Stat. J. 22:153–161.
Linhart, H., and W. Zucchini. 1986. Model selection. Wiley, New York.
Madigan, D., J. Gavrin, and A. E. Raftery. 1995. Eliciting prior information to enhance the predictive performance of Bayesian graphical models. Commun. Stat. Theory Methods 24:2271–2292.
Madigan, D. M., and A. E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's Window. J. Am. Stat. Assoc. 89:1335–1346.
Mau, B., and M. A. Newton. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. J. Comp. Grap. Stat.
Mau, B., M. A. Newton, and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12.
Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092.
Miller, A. J. 2002. Subset Selection in Regression, 2nd edition edition. Chapman & Hall/CRC, New York.
Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:674–683.
Morozov, P., T. Sitnikova, G. Churchill, F. J. Ayala, and A. Rzhetsky. 2000. A new method for characterizing replacement rate variation in molecular sequences: Application of the Fourier and Wavelet models to Drosophila and mammalian proteins. Genetics 154:381–395.
Myrvold, W. C., and W. L. Harper. 2002. Model Selection, Simplicity, and Scientific Inference. Philos. Sci. 69:S135–S149.
Nishii, R. 1984. Asymptotic properties of criteria for selection of variables in multiple regression. Ann. Stat. 12:758–765.
Nishii, R. 1988. Maximum likelihood principle and model selection when the true model is unspecified. J. Multivar. Ana. 27.
Nylander, J. A. 2004. Bayesian Phylogenetics and the Evolution of Gall Wasps. Pages 43 *in* Acta Universitatis Upsaliensis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 937 Uppsala University, Uppsala, Sweden.
Nylander, J. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47–67.
Occam, W. ca.1320. Scriptum in Librum Primum Sententiarum, Opera Theologica, I.
Ogishima, S., F. Ren, and H. Tanaka. 2000. Efficiencies of information criteria for topology selection in reconstructing molecular phylogenetic tree *in* Proceedings of International Symposium on Artificial Life and Robotics 2000:745–748.
Ota, R., P. J. Waddell, M. Hasegawa, H. Shimodaira, and H. Kishino. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Mol. Biol. Evol. 17:798–803.
Penny, D., P. J. Lockhart, M. A. Steel, and M. D. Hendy. 1994. The role of models in reconstructing evolutionary trees. Pages 211–230 *in* Models in Phylogenetic Reconstruction (R. W. Scotland, D. J. Siebert, and D. M. Williams, eds.). Clarendon Press, Oxford.
Pol, D. in press. Empirical problems of the hierarchical likelihood ratio test for model selection. Syst. Biol.
Popper, K. R. 1959. Logic of scientific discovery. Hutchinson, London.
Posada, D. 2001. The effect of branch length variation on the selection of models of molecular evolution. J. Mol. Evol. 52:434–444.
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, D. B. Davison, R. D. M. Page, G. A. Petsko, L. D. Stein, and G. D. Stormo, eds.). John Wiley & Sons, Inc.
Posada, D., and K. A. Crandall. 1998. Modeltest: Testing the model of DNA substitution. Bioinformatics 14:817–818.
Posada, D., and K. A. Crandall. 2001a. Selecting models of nucleotide substitution: An application to human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. 18:897–906.
Posada, D., and K. A. Crandall. 2001b. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50:580–601.
Posada, D., and K. A. Crandall. 2001c. Simple (wrong) models for complex trees: Empirical Bias. Mol. Biol. Evol. 18:271–275.
Pupko, T., D. Huchon, Y. Cao, N. Okada, and M. Hasegawa. 2002. Combining multiple data sets in a likelihood analysis: Which models are the best? Mol. Biol. Evol. 19:2294–2307.
Raftery, A. E. 1996. Hypothesis testing and model selection. Pages 163–187 *in* Markov chain Monte Carlo in practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). Chapman & Hall, London, New York.
Raftery, A. E. 1999. Bayes factors and BIC: Comment on "A critique of the Bayesian information criterion for model selection." Sociol. Methods Res. 27:411–427.
Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
Rzhetsky, A., and M. Nei. 1995. Tests of applicability of several substitution models for DNA sequence data. Mol. Biol. Evol. 12:131–151.
Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. Akaike information criterion statistics. Springer, New York.
Sanderson, M. J., and J. Kim. 2000. Parametric phylogenetics? Syst. Biol. 49:817–829.
Schwarz, G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464.
Shafer, G. 1982. Lindley's paradox (with discussion). J. Am. Stat. Assoc. 77:325–351.
Shibata, R. 1986. Consistency of model selection and parameter estimation. J. Appl. Prob. 23A:127–141.

Shimodaira, H. 1997. Assessing the error probability of the model selection test. Ann. Inst. Stat. Math. 49:395–410.
Shimodaira, H. 1998. An application of multiple comparison techniques to model selection. Ann. Inst. Stat. Math. 1:1–13.
Shimodaira, H. 2001. Multiple comparisons of log-likelihoods and combining nonnested models with applications to phylogenetic tree selection. Commun. Stat. Theory Methods 30:1751–1772.
Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1234.
Sober, E. 2002a. Bayesianism—its scope and limits. Pages 21–38 *in* Bayes's Theorem (R. Swinburne, ed.). Oxford University Press, Oxford.
Sober, E. 2002b. Instrumentalism, parsimony, and the Akaike framework. Phil. Sci. 69:S112–S123.
Sober, E., and M. Steel. 2002. Testing the hypothesis of common ancestry. J. Theoret. Biol. 218:395–408.
Sota, T., and A. P. Vogler. 2001. Incongruence of mitochondrial and nuclear gene trees in the Carabid beetles Ohomopterus. Syst. Biol. 50:39–59.
Steel, M., and D. Penny. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17:839–850.
Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J. R. Stat. Soc. 39:44–47.
Strimmer, K. 2001. Model selection using expected likelihood weights: A Bayes-frequentist compromise. Technical Report available at http://www.stat.uni-muenchen.de/∼ strimmer/cv.html.
Strimmer, K., and A. Rambaut. 2001. Inferring confidence sets of possibly misspecified gene trees. Proc. R. Soc. Lond. B Biol. Sci. 269:137–142.
Suchard, M. A., C. M. Kitchen, J. S. Sinsheimer, and R. E. Weiss. 2003a. Hierarchical phylogenetic models for analyzing multipartite sequence data. Syst. Biol. 52:649–664.
Suchard, M. A., R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer. 2002. Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. Syst. Biol. 51:715–728.
Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 18:1001–1013.
Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2003b. Testing a molecular clock without an outgroup: Derivations of induced priors on branch-Length restrictions in a Bayesian framework. Syst. Biol. 52:48–54.
Sugiura, N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. Commun. Stat. Theory Methods A7:13–26.
Sullivan, J., and D. L. Swofford. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenies. 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.
Suzuki, Y., G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA 99:16138–16143.
Swofford, D. L. 1998. PAUP* Phylogenetic analysis using parsimony and other methods, version 4.0 beta. Sinauer Associates, Sunderland, Massachusetts.
Swofford, D. L. 2000. PAUP*. Phylogenetic analysis using parsimony (*and other methods). version 4. Sinauer Associates, Sunderland, Massachusetts.
Tamura, K. 1994. Model selection in the estimation of the number of nucleotide substitutions. Mol. Biol. Evol. 11:154–157.
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.
Tanaka, H., F. Ren, T. Okayama, and T. Gojobori. 1999. Topology selection in unrooted molecular phylogenetic tree by minimum model-based complexity method. Pac. Symp. Biocomput. 4:326–337.
Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Pages 57–86 *in* Some mathematical questions in biology—DNA sequence analysis (R. M. Miura, ed.) American Mathematical Society, Providence, Rhode Island.
Van Den Bussche, R. A., R. J. Baker, J. P. Huelsenbeck, and D. M. Hillis. 1998. Base compositional bias and phylogenetic analyses: A test of the "flying DNA" hypothesis. Mol. Phylogenet. Evol. 10:408–416.
Verdinelli, I., and L. Wasserman. 1995. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. J. Am. Stat. Assoc. 90:614–618.
Vuong, Q. H. 1989. Likelihood ratio tests for model selection and nonnested hypotheses. Econometrica 57:307–333.
Wasserman, L. 2000. Bayesian model selection and model averaging. J. Math. Psychol. 44:92–107.
Weakliem, D. L. 1999. A critique of the Bayesian information criterion for model selection. Sociol. Methods Res. 27:359–397.
Whelan, S., and N. Goldman. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Mol. Biol. Evol. 16:1292–1299.
Woodroofe, M. 1982. On the model selection and the arc sine laws. Ann. Stat. 10:1182–1194.
Yang, Z. 1996a. Among-site rate variation and its impact on phylogenetic analysis. Trends Ecol. Evol. 11:367–372.
Yang, Z. 1996b. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.
Yang, Z., N. Goldman, and A. Friday. 1995. Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. Syst. Biol. 44:384–399.
Yang, Z., R. Nielsen, N. Goldman, and A.-M. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.
Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.
Zhang, J. 1999. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. Mol. Biol. Evol. 16:868–875.
Zharkikh, A. 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39:315–329.
Zucchini, W. 2000. An introduction to model selection. J. Math. Psychol. 44:41–46.

*First submitted 25 November 2003; reviews returned 29 January 2004; final acceptance 10 June 2004*
*Associate Editor: Jeffrey Thorne*