Annu. Rev. Ecol. Evol. Syst. 2005. 36:445–66
doi: 10.1146/annurev.ecolsys.36.102003.152633

First published online as a Review in Advance on September 16, 2005

MODEL SELECTION IN PHYLOGENETICS

Jack Sullivan[1,2] and Paul Joyce[2,3]
[1]*Department of Biological Sciences, University Idaho, Moscow, Idaho 83844-3051; email: jacks@uidaho.edu*
[2]*Initiative in Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, Moscow, Idaho 83844*
[3]*Department of Mathematics, University of Idaho, Moscow, Idaho 83844-1103; email: joyce@uidaho.edu*

Key Words AIC, BIC, decision theory, likelihood ratio, statistical phylogenetics

■ **Abstract** Investigation into model selection has a long history in the statistical literature. As model-based approaches begin dominating systematic biology, increased attention has focused on how models should be selected for distance-based, likelihood, and Bayesian phylogenetics. Here, we review issues that render model-based approaches necessary, briefly review nucleotide-based models that attempt to capture relevant features of evolutionary processes, and review methods that have been applied to model selection in phylogenetics: likelihood-ratio tests, AIC, BIC, and performance-based approaches.

INTRODUCTION

In this review, we assume the well-known view first voiced by Box (1976) that all models are wrong, but some are useful. After a brief introduction, we discuss alternatives for evaluating the adequacy of the chosen model. Finally, we illustrate how each of the traditional approaches to model selection fit well within the framework of decision theory (DT) and that DT facilitates an understanding of the goals and assumptions of these approaches.

The Importance of Models

Phylogenetic analysis is entering the genomics era, and as tools for surveying genomes (e.g., expressed sequence tags, single-nucleotide polymorphisms, genome sequencing, etc.) become more widely available, phylogenetic studies at all levels, from intraspecific phylogeography to the tree of life, will increasingly use data from multiple-gene loci. Concurrent with the advent of phylogenomics is the application of phylogenies to an ever-widening array of disciplines. For example, statistical phylogenetics have been permitted as evidence in a criminal court recently in which a Louisiana physician was convicted of infecting his former girlfriend with HIV from one of his HIV-positive patients (Metzker et al. 2002),

1543-592X/05/1215-0445$20.00

and phylogenetic testing has been used recently to refute the hypothesis that contaminated polio vaccine was the origin of the AIDS epidemic (Worobey et al. 2004).

Applying the emerging wealth of data to such an array of issues, however, presents difficulties because multiple loci are likely to be evolving under very different constraints and, therefore, may be subject to diverse substitution processes. One must, therefore, decide how best to account for the diversity of substitution processes in model-based phylogeny estimation, even for potential partitions in a single-gene data set. In our review of model choice in phylogenetics, we begin by introducing first the importance of probabilistic models in science generally, and then in the particular case of phylogenetics.

Models in Science

Statistical models allow scientists to exceed a mere description of their data and extend to proposing and testing general principles that can explain the data. Thus, statistical models add precision to the formulation of a scientific hypothesis and provide a rigorous means by which to assess the evidence for or against a hypothesis by providing a context for making predictions. Statistical models and methods are therefore ubiquitous in science.

Interestingly, the founder of modern statistics, R.A. Fisher, discovered the likelihood principle and invented maximum likelihood (ML) (Fisher 1958) primarily to answer questions related to evolutionary genetics. However, he did most of his work before the discovery of DNA, and, thus, he focused on quantitative genetics. Fisher's paradigm has been the centerpiece of data analysis throughout much of science in general, and much of biology in particular (e.g., Johnson & Omland 2004), but application of the ML principle and its explicit modeling approach has been slow in coming to phylogenetics. This delay was caused partly by the computational complexity of the problem and partly by an antithetical attitude of some systematists toward statistical approaches (e.g., Siddall & Kluge 1997). Computational difficulties have been ameliorated by a number of advances in theory and implementation (e.g., Huelsenbeck & Ronquist 2001, Swofford 1998), and philosophical objections have not proved sufficiently compelling to the broader community of systematics to halt the advance of model-based approaches to phylogenetics. Thus, the fact that Fisher's methodology is now dominating the field of phylogenetic biology, particularly in the analysis of molecular data, seems particularly appropriate to us.

Models in Phylogenetics

The necessity of models in molecular phylogenetics and evolution was recognized in the first comparative analyses of DNA sequence data (e.g., Brown et al. 1982, Jukes & Cantor 1969). Sequence divergence is roughly linear with time only shortly after a divergence event. The cause of this deviation from linearity is multiple substitutions at the same site (i.e., multiple hits), and the earliest molecular

evolutionary studies attempted to accommodate multiple hits in estimating the number of substitutions that have occurred since two sequences diverged from a common ancestor by use of explicit models (Jukes & Cantor 1969).

Furthermore, the consequence of ignoring multiple substitutions was also recognized early: underestimation of the number of substitutions that have occurred since two sequences last shared a common ancestor. More importantly, however, this underestimation is not uniform. Long branches (and large genetic distances) will be underestimated disproportionately more than will short branches and genetic distances (e.g., Gillespie 1986). Some of the implications of this nonuniform underestimation are well studied [e.g., long-branch attraction (LBA) (Felsenstein 1978)], but the effect of model choice on data exploration seems to be less appreciated.

MODELS IN EXPLORING DATA VIA SATURATION PLOTS The recognition that multiple hits can occur led to the concept of substitutional saturation (e.g., Brown et al. 1982), which is still of concern to many molecular phylogeneticists. However, because most studies lack the fossil data that Brown et al. (1982) and others have used to establish the *x*-axis in early saturation plots, most assessments of saturation use some measure of pairwise genetic distance on the *x*-axis as a proxy for time. Some other aspect of molecular evolution, say, the absolute number of transitions, is then plotted on the *y*-axis to make inferences about the relationship between that variable and genetic distance. Such plots are frequently used as exploratory tools with which to understand the processes that have generated a data set of interest (e.g., López-Fernández et al. 2005) and are frequently used to justify decisions about data elimination (e.g., Han & Ro 2005).

However, for the *x*-axis to be at all meaningful, estimates of genetic distances for use as the *x*-axis must be based on a model of evolution that estimates multiple substitutions adequately. If an underparameterized model is used, genetic distances will be undercorrected and will underestimate the actual number of substitutions disproportionately more for large distances than for small distances (e.g., Golding 1983). The effect that this error will have on saturation plots is simple to predict; the *x*-axis will be compressed nonuniformly and use of overly simple models in saturation plots (or even worse, use of uncorrected *p*-distances) will obfuscate understanding of the processes of molecular evolution.

This problem is common and is illustrated in Figure 1. These plots were generated from the COI data of Cicero & Johnson (2001), who used them (along with data from Cyt b, ND2, and ND3) to estimate phylogenetic relationships among *Empidonax* flycatchers. They illustrated a linear relationship apparent between third-position transitions in the original saturation plots by use of *p*-distances (figure 3 in Cicero & Johnson 2001), and this apparent linearity was used to justify inclusion of those sites in an equally weighted parsimony analysis, whereas other data were eliminated (not shown). However, a plot based on the HKY + I + Γ distances (Figure 1*A*), which the authors chose for ML analysis by application of the hierarchical likelihood-ratio test (LRT, see below), leads to very different

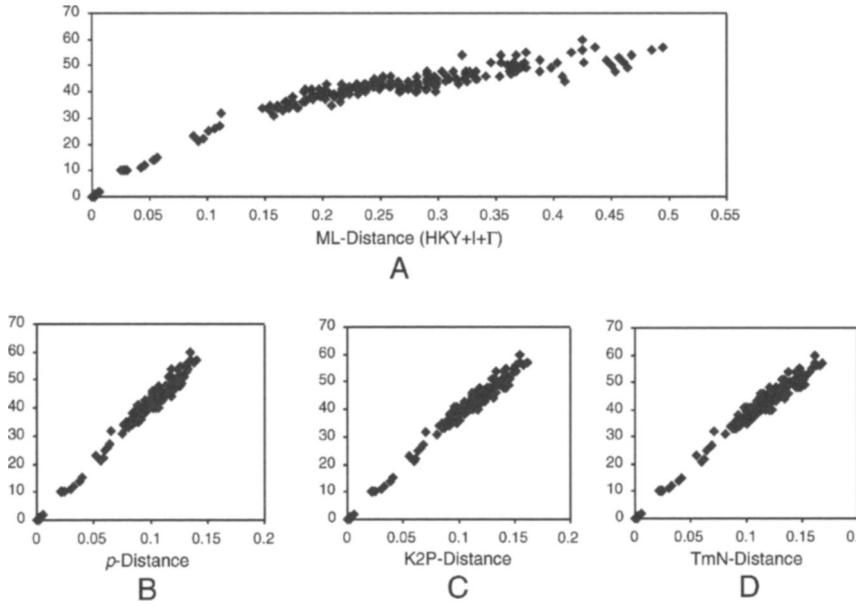

Figure 1 The effect of model choice on data exploration. Data are from Cicero & Johnson (2001); the *y*-axis is absolute number of third-position transitions, and the *x*-axis is genetic distance corrected by use of various models. Modeltest was used by the orignal authors to select HKY + I + Γ.

conclusions regarding the pervasiveness of multiple transitions at third-codon positions in their COI data than does their plot based on *p*-distances (Figure 1*B*). Their conclusions about the prevalence of multiple hits that involve third-position transitions in this data set are spurious and the result of use of a poorly chosen model in the saturation plots. Furthermore, many studies have used models such as the Kimura two-parameter (K2P) model (Kimura 1980) or Tamura-Nei distances (Tamura & Nei 1993) to calculate genetic distances for the *x*-axis in saturation plots. However, as is shown in Figure 1*C* and 1*D*, use of neither of these simple models as the *x*-axis in saturation plots results in detection of multiple substitutions that involve third-position transitions in the *Empidonax* COI data. Clearly, model choice has a dramatic effect on exploration of data.

THE EFFECT OF UNDERESTIMATION OF MULTIPLE SUBSTITUTIONS IN PHYLOGENY

Felsenstein (1978) was the first to point out that the underestimation of multiple hits can result in inconsistent estimation of phylogeny if the (unknown) true tree contains long branches separated by a short internal branch. This result is caused by the well-studied phenomenon of LBA and is the result of precisely the same underestimation of evolutionary change (number of substitutions) described above. Huelsenbeck & Hillis (1993) examined the performance of many methods across

a variety of tree shapes and demonstrated that accurate estimation of phylogenies is difficult, regardless of method, under the conditions that Felsenstein (1978) had described. This conclusion led them to dub that region of tree space (where two long branches are separated by a short internal branch) the Felsenstein zone. In subsequent studies, investigators have demonstrated via simulations that the underestimation of nucleotide substitutions associated with overly simplified models leads to LBA and inconsistent estimation in the Felsenstien zone, even when ML is used (e.g., Gaut & Lewis 1995, Sullivan & Swofford 2001). Furthermore, a few studies have demonstrated that use of inadequate likelihood models can lead to LBA in real data sets (e.g., Anderson & Swofford 2004, Sullivan & Swofford 1997).

The large body of simulation studies show that the shape of the underlying true tree has an enormous impact on the importance of model choice. In the ideal case (Figure 2), the underlying tree shape is such that all existing methods estimate phylogeny accurately; ML estimation is very robust to violations of model assumptions, and model choice is not critical (e.g., Sullivan & Swofford 2001). However, model choice is critical in the Felsenstein zone (Figure 2), and that observation is widely accepted.

Although perhaps not a widely appreciated, biases associated with violation of model assumption may favor the true tree. Specifically, if long terminal branches are adjacent to a short internal branch [termed the Farris zone by Siddall (1998) and the inverse Felsenstein zone by Swofford et al. (2001)] (Figure 2), the underestimation of long terminal branches will result in overestimation of the short internal branch and cause the most biased methods (such as parsimony and ML under an oversimplified model) to recover the true tree with high confidence and with very little data (Bruno & Halpern 1999, Siddall 1998, Sullivan & Swofford 2001, Swofford et al. 2001, Yang 1997). In fact, the most overly simplified method of phylogenetic estimation will be the most efficient (Sullivan & Swofford 2001). Some have suggested that this bias might be a useful attribute of methods such as parsimony and ML under simplistic models (Siddall 1998, Yang 1997). However, others have suggested that this bias is caused by misinterpretation of convergent substitutions as synapomorphies and should be avoided (Bruno & Halpern 1999, Sullivan & Swofford 2001, Swofford et al. 2001). Model choice is therefore critical here as well.

Figure 2 The effect of topology on robustness. At the center of the continuum, phylogenetics signal is strong and model choice is not critical (i.e., maximum likelihood is robust to violations of model assumptions). In the Felsenstein zone (*left*), model selection is critical, as is also the case for the inverse Felsenstein zone (*right*).

Although estimation of topology may not always be compromised by use of overly simple models, estimation of nodal support certainly is. This outcome has been demonstrated for nonparametric bootstrap values (Buckley & Cunningham 2002), parametric bootstrap tests of a priori hypotheses (Buckley 2002), and Bayesian posterior probabilities (Erixon et al. 2003, Huelsenbeck & Rannala 2004, Lemmon & Moriarty 2004). Furthermore, although most simulation studies have focused on the four-taxon cases shown in Figure 2, any large phylogeny can reasonably be expected to contain subtrees from across the continuum. That is, unless one has some assurance that no terribly short and no terribly long branches exist anywhere in the phylogeny that one is attempting to estimate, choice of an overly simple model is likely to impinge negatively on phylogeny estimation.

Overparameterization

Given the potential problems associated with overly simplistic models, an obvious reaction would be to always use the most complex model available. Indeed, use of the most complex model available has been advocated at times, at least for Bayesian estimation (e.g., Huelsenbeck & Rannala 2004). However, in general, this approach seems like a poor strategy. Although an increase in the number of parameters will always increase the fit between model and data (i.e., increase the likelihood), if that increase is simply the result of parameterizing stochastic variation, nothing is gained. With increased use of mutlilocus data for phylogeny estimation, the temptation will inevitably arise to partition data excessively. Such overparameterization can result in nonidentifiability of parameters because of a loss of degrees of freedom (Rannala 2002). Furthermore, Buckley et al. (2001) examined the performance of several models with regard to branch-length estimation from a data set containing 25 sequences of three mtDNA genes (COI, A6, and $tRNA_{Asp}$) from *Maoricicada* and two outgroups. They found that both GTR + I + Γ and GTR + Γ models (applied to all sites) provided better estimates of branch lengths than did a 10-class, site-specific rates (SSR) model (GTR + SSR_{10}), despite the fact that the SSR model is more parameter rich and has a better likelihood. Models with the best likelihood score are not guaranteed to produce the best estimates of branch lengths from finite data and, by extension, should not necessarily be expected to perform best in phylogeny estimation.

This suggestion by Huelsenbeck & Rannala (2004) was generated by the fact that, when they simulated data under a simple Jukes-Cantor (JC) model, they were able to estimate nodal probabilities accurately by estimating with an overparameterized GTR + Γ, even with sequences as short as 100 nt. This result is encouraging, but the recommendation based on that conclusion should be tempered somewhat for two reasons. First, the simulation conditions are very artificial. When the true model (JC) is a special case of the estimating model (GTR + Γ), the overparameterized estimating model will converge on the special-case true model (i.e., base frequencies will be estimated to be equal). This situation will never occur in real data, for which all models are almost certainly wrong. Similarly, some nonnested

models may be simpler than the most complex model available but may account for some important feature not addressed by the more highly parameterized model. In these cases, the simpler model may have better predictive ability.

Uniqueness of Phylogeny Estimation

The statistical nature of phylogeny estimation is very unusual. Standard statistical software packages, even ones as powerful as SAS or R, are unlikely to be of much use in phylogenetic analysis. The reason is that the fundamental parameter in phylogenetics is usually the tree topology, which is inherently discrete, whereas the wealth of statistical methodology and theory centers on continuously varying parametric models. Therefore, standard χ^2 goodness-of-fit tests are untrustworthy in the phylogenetic context, and methods such as parametric bootstrap or Bayesian posterior analysis (that do not rely on asymptotic theory) represent better statistical procedures for phylogeny estimation.

REVIEW OF MODELS

Reviews of models of nucleotide substitution have been provided by Swofford et al. (1996) and, more recently, by Felsenstein (2004). However, potentially important models are not presented in either of those publications and a brief review of models is therefore appropriate here.

GTR Family

Widely used models of nucleotide substitution are usually time reversible; an A→T transversion is treated as equivalent to a T→A transversion [i.e., $r_{(AT)} = r_{(TA)}$]. Thus, six possible substitution types exist among the four nucleotides. Each of these transformation types may be treated as equivalent (Jukes & Cantor 1969), transitions may be treated separately from transversions (e.g., Hasegawa et al. 1985, Kimura 1980), all six may be treated as unique (Tavaré 1986, Yang 1994), or any combination of the six types may be grouped. Thus, 203 transformation matrices are possible, each of which represents a special case of the GTR model. Furthermore, base frequencies may be assumed to be equal (i.e., Jukes & Cantor 1969, Kimura 1980) or allowed to vary.

Early models assumed that all sites in a collection of sequences evolve at a uniform rate. However, several methods have been developed to account for the observation that sites usually evolve at different rates (e.g., Uzzell & Corbin 1971). One may assume that some portion of the sites are invariable (e.g., Hasegawa et al. 1985), that rates across sites conform to a Γ-distribution (e.g., Uzzell & Corbin 1971, Yang 1993), or that rate heterogeneity is better described by a mixture of invariable sites and Γ-distributed rates, the $I + \Gamma$ model, in which some sites are invariable (p_{inv}) and rates at variable sites conform to a Γ-distribution (Gu et al. 1995, Waddell & Penny 1996).

Swofford et al. (1996) reviewed the development and conceptual relationships among some of the commonly used equal-rates models; these relationships can be expanded to accommodate the heterogeneous-rates models mentioned above. Under this framework, the most general and parameter-rich model (GTR + I + Γ) has the following substitution parameters:

- Rate matrix parameters: $r_{(AC)}$, $r_{(AG)}$, $r_{(AT)}$, $r_{(CG)}$, and $r_{(CT)}$, with $r_{(GT)} = 1$
- Base frequencies: π_A, π_C, π_G, with $\pi_T = 1 - (\pi_A + \pi_C + \pi_G)$
- Rate heterogeneity parameters: gamma shape (α), proportion of sites that are invariable (p_{inv})

All other submodels within this family are special cases of GTR + I + Γ, with one or more of the parameters constrained.

Nonreversible Models

In many data sets, base frequencies change in different parts of the tree, and a few models have been proposed that accommodate this change. Base frequencies may be allowed to change on every branch, for $3(2n - 2)$ compositional parameters (because trees must now be rooted), or only on terminal branches, for $3n$ compositional parameters (Yang & Roberts 1995). Alternatively, nucleotide frequencies may be pooled, so that only GC content varies across a tree (Galtier & Guoy 1998). Foster (2004) has made important advances in modeling nonuniform base frequencies. In particular, he has made the number of base-frequency vectors a parameter that can be estimated and, for several real data sets, has demonstrated that even a single change in base frequencies on the tree is sufficient to provide an adequate improvement in model fit.

Other nonreversible models are based on the covarion hypothesis of Fitch & Markowitz (1970), in which rates of sites can change across the tree. Tuffley & Steele (1998) were the first to model this situation explicitly, and it has been incorporated into corrections for evolutionary distances and likelihood frameworks (Galtier 2001, Hueslenbeck 2002). These advances are likely to be important in phylogeny estimation across the tree of life and will almost certainly require application of Markov chain Monte Carlo approaches (Felsenstein 2001).

Nonindependence Across Sites

CODON-BASED MODELS Because of the nature of the genetic code, one can expect nonindependence across sites within a codon. Codon-based models are particularly appealing for protein-coding genes because they account for the genetic code explicitly. Instead of a 4 × 4 rate matrix for transformations among nucleotides at a site, these models approximate a 61 × 61 matrix (with 3660 implied relative rates for the nonreversible version) to account for transformations among all possible (non–stop) codons for each triplet. The rate matrix is filled by use of the relevant genetic code, and rates of synonymous versus nonsynonymous codon substitution are

optimized. Underlying nucleotide substitution models assume uniform rates [i.e., a single-nucleotide substitution type but with nonequal base frequencies (Muse & Gaut 1994)], a difference between transitions and transversions [i.e., two nucleotide substitution types (Goldman & Yang 1994)], or allow all six substitution types (Halpern & Bruno 1998).

Another approach to deal with nonindependence of sites is use of hidden Markov models (Felsenstein 2001) to permit the autocorrelation of rates regionally. For some reason, the hidden Markov models have not been utilized extensively.

rRNA MODELS For ribosomal RNA (rRNA) genes, the primary transcript is the functional product. These rRNAs fold into a secondary structure in which some regions form pair-bonded stems and others form single-stranded loops. Substitutions in stem regions are constrained by the complementary nucleotide and compensatory changes (substitutions that maintain pair bonding) are well known. Models specific to rRNA have been developed (e.g., Smith et al. 2004, Tillier & Collins 1995) in which loop regions are treated as distinct from stem regions and the latter treated as hydrogen-bonded pairs, although these models are yet to be implemented in many phylogeny estimation packages [with the exception of MrBayes (Huelsenbeck & Ronquist 2001)]. Kjer (2004) used this model, coupled with a mixed-distribution model of among-site rate variation (the Doublet + I + Γ model) in analysis of 18S rRNA among insects. The parameters of the doublet model include 16 doublet frequencies (which sum to 1 for 15 free parameters), 3 free base frequencies, 5 free transformation rates for loops (from the reversible 4×4 nucleotide matrix), 119 free transformation rates for stems (from the reversible 16×16 doublet matrix), a separate p_{inv} for stems and loops (2 parameters), and a gamma across all variable sites. Clearly, this model is extremely parameter-rich.

Partitioned Models

If one has natural partitions in ones data sets (e.g., codon positions, multiple genes, etc.), an intuitively appealing option is to apply different models to the various partitions. The simplest of these approaches are the site-specific rate (SSR) models (although they really should be called partition-specific rate models), and these models apply a separate, equal-rates, GTR to each partition. Because partitions often have very different nucleotide frequencies, the simple SSR models often improve the likelihood score considerably. However, this improvement in fit may not equate to improved phylogeny estimates, because other simpler models (nonnested) may better account for rate variation among sites (Buckley et al. 2001).

Alternatively, one may apply a full GTR + I + Γ model to each partition (e.g., Castoe et al. 2004), and any of the parameters may be linked (apply across partitions) or unlinked (be partition specific). If one had, for example a 10-gene data set, from two genomes (nuclear and organellar), one could imagine an enormous array of potential, plausible partitioning schemes. Some way of evaluating the partitioned models is necessary to guide the choice.

MODEL SELECTION CRITERIA IN PHYLOGENTICS

Given that model choice is critical in phylogeny estimation and the vast array of potential models from which to choose, one is faced with the decision of how to select from among these. Obviously, the requirement is to select a model or models from the set available that account for processes that impinge on phylogeny estimation sufficiently well to avoid the biases discussed above without sacrificing the predictive power of the chosen model. Posada & Buckley (2004) recently published an excellent overview of model choice in systematics and focus on a justification for model averaging by use of AIC weights (see below).

Likelihood-Ratio Tests

By far, the most widely used method of choosing a model objectively is through use of LRTs. This approach takes advantage of two issues. First, the likelihood score can be interpreted as measuring the fit between model and data that is comparable across models. Second, the commonly used models in phylogeny estimation from DNA sequences are members of the GTR + I + Γ family (i.e., are special cases or submodels). Thus, one may evaluate the effect of including one or more parameters by calculating the likelihood of a model in which the parameter of interest is optimized versus a model in which it is fixed and comparing the likelihoods of the two models by use of the classical test statistic

$$\delta = 2(\ln L_1 - \ln L_0),$$

where $\ln L_1$ is the likelihood score of the more complex model. The test statistic is then typically evaluated under the assumption of asymptotic convergence to a χ^2 distribution; the degrees of freedom are the difference in number of free parameters in the two models.

This approach was first used in a hierarchical fashion (the hLRT) by Frati et al. (1997) and Sullivan et al. (1997), who selected a model for phylogeny estimation from among a set of 16 models. It was suggested independently by Huelsenbeck & Crandall (1997). Posada & Crandall (1998) hard-coded this approach in the production of their program Modeltest and expanded the set of candidate models examined to include 56 members of the family. The release of Modeltest had an enormously important impact on phylogenetics because it permitted many systematists to select good models in a nonarbitrary fashion.

A potential weakness of LRTs (Sanderson & Kim 2000) is that an initial estimate of topology, usually from either a parsimony search or a neighbor-joining tree, is required to conduct hLRTs. However, although model parameters are not as invariant across tree topologies as initially postulated, analyses of real data have shown that extremely poor estimates of model parameters are typically only derived from very poor trees (e.g., Sullivan et al. 1996). Similarly, Posada & Crandall (2001) demonstrated that use of initial trees has little effect on the model chosen by hLRTs.

Nevertheless, serious weaknesses remain in the use of hLRTs for model selection. One of these weaknesses is the requirement to traverse model space via a series of pairwise comparisons without relevant theory to guide the traversal. Model space can be represented by a decision tree (Posada & Crandall 1998), and the first choice one must make in applying hLRTs is where to start on this tree. One may start with the most general and parameter-rich model (typically GTR + I + Γ) and simplify by fixing the values of certain parameters (e.g., setting the proportion of invariable sites equal to zero). Conversely, one may start with the simplest model (JC) and add parameters (e.g., base frequencies) that are then optimized. Once the decision has been made as to which direction to follow (top down or bottom up) in traversing model space, one must decide the order in which to subtract or add parameters. This traversal may either be hard-coded, as is the case with Modeltest, or be done interactively. Swofford & Sullivan (2003) and Sullivan (2005) demonstrate the interactive approach to hLRTs, by starting with the most general model and subtracting parameters that appear closest to their fixed values in the simpler model. Not surprisingly, this approach often leads to selection of models that would never be examined in current hard-coded approaches.

Similarly, several authors have demonstrated that the manner in which the model space is traversed influences model choice (Cunningham et al. 1998, Felsenstein 2004, Pol 2004). In the most extensive examination, Pol (2004) examined 32 different traversals for 18 data sets and found that mode of traversal influenced model selection in 15 of the 18 data sets and that the selected models differed by as many a 6 parameters (for one data set). He further demonstrated for two data sets that the ML tree was different under models selected by use of different traversal schemes (however, in both cases, trees only differed very slightly, by one or two nearest-neighbor interchanges). These problems in how best to implement hLRTs arise because no relevant theory exists to guide traversal of model space.

In addition to these issues of implementation (as well as others; for example, multiple testing), several authors have pointed out that LRTs were not intended to be used to select from a series of models (e.g., Posada & Buckley 2004). Similarly, the hypothesis-testing approach inherent in hLRTs is poorly suited to model selection, and LRTs typically favor the complex model (e.g., Burnham & Anderson 2002). Thus, despite the extremely widespread use of hLRTs to select models for phylogenetics, and the enormous improvement that this approach has made to model-based phylogenetics, time has probably come to move to other alternatives, including some that have been developed recently.

Akaike Information Criterion

The Akaike information criterion (AIC) (Akaike 1973) is a simple measure with a complex derivation. The AIC for model i (AIC_i) is calculated as follows:

$$\text{AIC}_i = -2 \ln L_i + 2k_i,$$

where ln L_i is the maximum log-likelihood of the model (i.e., with joint ML estimates across parameters) and k_i is the number of parameters in model i. In addition, a modification to correct for small sample sizes (where small is defined as $n/k_i \leq 40$, and n is typically the number of sites), the AIC_c (Burnham & Anderson 2002, 2004) is given by the following:

$$AIC_{Ci} = -2 \ln L_i + 2k_i + \frac{2k_i(k_i + 1)}{n - k_i - 1}.$$

The simple interpretation of the AIC is that it provides a measure of fit between model and data ($-2 \ln L_i$) and includes a penalty for overparameterization. Its first application to phylogenetics was by Hasegawa (1990), and the model favored is that model with the lowest AIC (or AIC_c). Ideally, one would find the ML topology and parameters for each model, but usually, some initial tree is used across all models. In other words, as typically applied, the AIC shares the reliance on an initial tree with hLRTs. However, Posada & Crandall (2001) demonstrated that this reliance has virtually no effect on the model chosen by comparing the AIC rankings based on the true tree (in simulated data) with the rankings based on an initial (NJ) tree. A similar conclusion was reached by Abdo et al. (2005), who compared the models selected by AIC calculated on an initial tree with those chosen by optimizing the tree under each model examined (i.e., on the ML tree for each model).

An obvious advantage of AIC over LRTs in model selection is that the AIC is calculated for each model in isolation, which eliminates the need to traverse model space by a series of pairwise comparisons. The AIC can, therefore, be used to compare nonnested models. Another advantage of the AIC is that it can allow for generation of a plausible set of models by computation of the Δ_i for each model as follows:

$$\Delta_i = AIC_i - AIC_{min},$$

where AIC_{min} is the score of the preferred model. These Δ_i values provide for evaluating the support in the data for each of the models that is examined (i.e., quantifying uncertainty in model selection). Burnham & Anderson (2002, 2004) provide the following benchmarks for discerning the relative support for alternative models: $\Delta_i \leq 2$ indicates substantial support, $4 \leq \Delta_i \leq 10$ indicates weak support, and $\Delta_i \geq 10$ indicates no support. Furthermore, these Δ_i values can be used to erect AIC weights for multimodel inferences (see below).

Although the interpretation of the AIC given above is sufficient to understand the properties of the AIC, the approach has a formal derivation from information theory. Suppose we have a distribution that has been generated by some true but unknown process. The AIC represents the Kullback-Leibler (K-L) distance between that model and the model being examined. The K-L distance can be thought of as quantifying the information lost by approximation to the true model. More details are provided in the online Supplemental Material of this review;

follow the Supplemental Material link from the Annual Reviews home page at http://www.annualreviews.org.

Bayesian Model Selection

BAYES FACTORS In Bayesian comparison of two models, the Bayes factor permits direct evaluation of the support in the data for one model versus another (Kass & Raftery 1995). This support is calculated as by $B_{12} = \text{pr}(D|M_1)/\text{pr}(D|M_2)$, and it can be multiplied by the ratio of the prior probabilities of each model to give the posterior odds that favor one model. Thus, if the priors are uniform (i.e., the ratio of priors equals 1), the posterior odds take a similar form as the LRT, with the important difference that $\text{pr}(D|M_i)$ is calculated by integrating across the parameters of M_i rather than by fixing parameter values at the ML point estimates. Bayes factors, therefore, account for uncertainty in parameter estimation, unlike hLRTs. As with the Δ_i under the AIC, benchmarks are provided by Raftery (1996) to interpret relative support on the basis of the magnitude of the Bayes factor. When $B_{12} > 20$, support for M_1 is strong; when $3 \leq B_{12} \leq 20$, M_1 is slightly favored; and when $1 \leq B_{ij} < 3$, the two models are supported roughly equally by the data. Suchard et al. (2002) used Bayes factors to examine a nested subset of the GTR + I + Γ family and rejected the K2P and HKY models in favor of the Tamura-Nei model (Tamura & Nei 1993). However, unlike in the case of LRTs, Bayes factors are not restricted to comparisons of nested models. For example, Nylander et al. (2004) used Bayes factors to select from an array of partitioned models that included nonnested variants. Interestingly, simpler models were preferred over more complex models only in comparisons of nonnested models. In this example, because no penalty was imposed for overparameterization, Bayes factors always favored the more general of two nested modes. They also noted symptoms of nonidentifiability (diffuse and highly skewed marginal posterior distributions) of p_{inv} and the Γ-shape parameter in the smallest partitions. Sullivan et al. (1999) have demonstrated the correlation of error in these two parameters, and this error impedes their estimation with limited data and likely explains the issues of nonidentifiability seen by Nylander et al. (2004).

BAYESIAN INFORMATION CRITERION An approximation of full Bayesian model evaluation was devised by Schwarz (1978): the Bayesian information criterion (BIC). In calculation, this quantity is similar to the AIC,

$$\text{BIC}_i = -2 \ln Li + k_i \ln n,$$

where k_i is the number of parameters in model i, $\ln L_i$ is the ML score (i.e., with all parameters fixed to their ML point estimates), and n is the sample size. As above, sample size is typically taken to be the number of nucleotide sites, but its appropriate interpretation in phylogenetics is not entirely clear. Again, a superficial characterization of the BIC is that it assesses fit via the ML score and penalizes overparameterization (more heavily than is the case for the AIC, especially with

large n). Moreover, the BIC resists the tendency for model selection to favor more complex models as n increases.

Again, as typically employed, BIC_i values are calculated on an initial tree, rather than the ML tree, under the model M_i. Just as for the AIC, Posada & Crandall (2001) demonstrated by use of simulations that this approximation is quite good, and Abdo et al. (2005) demonstrated the same by actually calculating the BIC_i on the ML tree for all M_i.

Just as the AIC has a more rigorous statistical justification than simply assessing fit plus a penalty for overparameterization (i.e., minimizing the K-L distance), the model with the minimum BIC will be the same as the model with the highest posterior probability, $\text{pr}(M_i|D)$, at least if one assumes uniform priors across models and certain approximations are valid. This derivation is discussed in more detail in the Supplemental Material available online.

Performance-Based Model Selection

Minin et al. (2003) developed a model-selection approach that ranks models on the basis of the weighted expected error in branch-length estimates, with the weights are derived from the BIC. This method focuses on the fact that both the tree topology and the branch lengths (the rate of evolution $\times$ the time between each node or speciation event in the tree) are critical. If we assume momentarily that topology is known, we can focus attention on accurate branch-length estimates; rather than worry about whether a model is correct, the accuracy of the branch lengths estimated under various models can be used to assess model quality.

Because the method of Minin et al. (2003) (available in the program DT-ModSel) relies on decision theory (DT), we defer explanation of the details of the method to the Supplemental Material available online; that material focuses on the decision-theoretic foundations of all the model-selection criteria. However, a few points are worth noting here. First, accuracy in branch-length estimation is justified as a performance measure by the observation that the reason ML estimation can be inconsistent under some topological conditions under strongly violated models is because of the underestimation of long branches discussed above. Thus, models that are expected to estimate branch lengths similarly are expected to perform similarly in phylogeny estimation. Abdo et al. (2005) validated the assumption by using data simulated under very complex conditions. Second, because the approach uses BIC weights, it typically selects simpler models than does either hLRTs (Minin et al. 2003) or AIC (Abdo et al. 2005). Nevertheless, these simpler models produce estimates of branch length with less error (both absolute error and relative error) and produce phylogeny estimates at least as accurate as the complex models selected by hLRTs, AIC, and BIC (Abdo et al. 2005). Third, inclusion of several poor models in the set examined has no effect on model choice, because the poor models receive extremely low BIC weights (Abdo et al. 2005). Fourth, although the method uses an initial estimate of topology (as do the other methods), this approximation does not compromise model choice (Abdo et al. 2005).

Finally, the loss function need not focus on branch-length estimates; any feature of the analysis can be used to erect a loss function.

Tests of Model Adequacy

Given the increasing uses of both Bayesian and frequentist tests of evolutionary hypotheses on model-based phylogenies, the adequacy of models should be assessed in an absolute sense. That is, all the methods described above permit us to choose objectively one or more models from a preselected set, but, although we certainly do not anticipate that the selected model or models will be true, any statistical tests conducted by use of the selected model or models may be compromised if the best available alternative is nevertheless insufficient. Thus, an absolute test of model adequacy is critical (Sanderson & Kim 2000), and two have been used in phylogenetics.

PARAMETRIC BOOTSTRAP The first test of the absolute goodness-of-fit between model and data in phylogeoentics was proposed by Goldman (1993) and is described in detail in Whelan et al. (2001). This test is a simulation-based test, and it uses as a test statistic the difference between the multinomial likelihood, which sets an upper bound on the likelihood for the data set under examination, and the ML achievable under that model. This difference measures the deterioration in fit associated with forcing all the data to conform to a single (albeit potentially heterogeneous-rates) model and a single tree. Replicate data sets are then simulated on the ML tree under the model being examined, with parameters fixed to their ML estimates, and the difference between multinomial likelihood and ML under the model is examined for each data set. This difference represents the expected difference under the null hypothesis of a perfect fit between model and data (simply due to stochasticity) because the model was used to generate the data. The distribution of this difference across replicates then becomes the null distribution to which the observed difference is compared.

In the first application of this test, Goldman (1993) evaluated the absolute fit of the simple equal-rates models available at the time and could reject them for real data sets. Similarly, Whelan et al. (2001) rejected the GTR model (without rate variation) for primate mtDNA by use of this test, and these results have led to the perception that current modes are inadequate (e.g., Sanderson & Kim 2000). However, a number of studies have applied the multinomial test of model adequacy to heterogeneous rates models, and in many of these studies (e.g., Carstens et al. 2004, Demboski & Sullivan 2003, Sullivan et al. 2000), the model selected by one of the selection methods could not be rejected in terms of absolute goodness-of-fit. Thus, despite the early conclusions, many conditions exists in which models chosen from among a pool of candidates appear to be adequate, at least as judged by these tests.

However, one limitation of this test is that it relies on point estimates of topology, branch lengths, and model parameters to simulate null distributions. This limitation has the effect of underrepresenting uncertainty in the simulations and may compromise the power of those tests. An analysis of error rates by use of this approach is currently lacking, and the effect of its reliance on point estimates is not known.

POSTERIOR PREDICTIVE SIMULATIONS Huelsenbeck et al. (2001) and Bollback (2002) have circumvented the weakness of Goldman's test by making use of posterior predictive simulations. This approach uses Bayesian estimation under the model being examined to provide posterior probability distributions of topologies, branch lengths, and substitution-model parameters. Simulations are then conducted under the model under examination, and each replicate samples the tree, branch lengths, and parameter values from the marginal posterior distributions. The idea is that future data should be predictable under a good model, but future data do not exist. Therefore, future data are simulated under conditions selected from the marginal posterior distributions derived from Bayesian analysis of real data, and replicates, therefore, account for uncertainty in parameter estimation. The multinomial likelihood from the real data is used as the test statistic and it is compared with the distributions of multinomial likelihoods derived from the posterior predictive simulations.

Interestingly, Bollback (2002) examined one of the same data sets that Goldman (1993) examined: the primate $\psi\eta$-globin data set. Whereas Goldman (1993) rejected the JC model for this data set by use of the parametric bootstrap test of absolute goodness-of-fit, Bollback could not (the P value was 0.123). Bollback attributes this outcome to the uncertainty in model parameters, topology, and branch lengths and the fact that the posterior predictive simulations account for this uncertainty explicitly. Comparison of the two methods on a diversity of real data sets would be extremely useful (e.g., Foster 2004). A second interesting result from Bollback's analysis of that data set is that the PPS test suggested that the HKY model (four parameters) is a better fit than the more general GTR model (eight parameters).

INCORPORATING UNCERTAINTY IN MODEL SELECTION

Classical parameter estimation involves choosing the appropriate statistical model and then estimating the parameter in the context of that model. Typically one only accounts for error in the estimate assuming the particular model chosen but does not account for the error associated with the model choice. This approach produces bias in the estimates, and the standard error of estimates calculated with a single model underrepresents the true error in the estimates. Model averaging is a way to overcome these problems (Burnham & Anderson 2002); this technique involves assigning each model a certain weight, estimating the parameter of interest under each model, and then producing an average estimate that is weighted across models. In the phylogeny context, Posada & Buckley (2004) have advocated AIC weights (w_i). These weights are a function of Δ_i as defined above (Burnham & Anderson 2002, 2004), and a few examples of model-averaged phylogenies that use AIC weights are in the literature (e.g., Posada & Buckley 2004).

However, model averaging requires that one accepts that models can be viewed as random variables, and one assigns a probability distribution to each of the models given the data. From the perspective of statistical philosophy, this approach

requires the Bayesian view of statistical inference. Under the Bayesian view, the only logically coherent way to weight each model is to assign each model a weight according to the posterior probability of the model given the data. Thus, one could make the argument that if one is willing to use model averaging as a legitimate statistical procedure, only Bayesian approaches make sense, although Burnham & Anderson (2004) provide a Bayesian interpretation of AIC weights. In particular, the posterior probability of a model is equivalent to the AIC weight [$\text{pr}(M_i \mid D) = w_i$], when the prior probabilities across models assume a particular form (for the derivation, see Burnham & Anderson 2004). Therefore, model averaging by use of AIC weights can be viewed as ad hoc; that is, to be consistent with Bayesian statistics, one is required to assume particular priors across models.

An alternative approach to model averaging by use of AIC weights in phylogenetics is reversible-jump Markov chain Monte Carlo (Huelsenbeck et al. 2004, Nylander et al. 2004, Suchard et al. 2002). This approach includes proposals to change models randomly in the Markov chain Monte Carlo proposal mechanism. Because this approach does not require any particular form of the priors across models, it seems to us to be a theoretically more justifiable approach to model averaging than is the use of AIC weights. Alternatively, many researchers seem to take a pragmatic approach to statistics and use methods that can be shown to work well under a variety of relevant conditions. AIC weights may prove to work sufficiently well in model averaging in phylogenetics.

CONCLUSIONS

Phylogenetics is beginning to grapple with model-selection issues, just as have other disciplines. Although what will ultimately be viewed as optimal model selection may depend on whether one is willing to adopt a Bayesian statistical philosophy, the fact that all current approaches to model selection can be formalized as a loss function within a DT framework facilitates direct comparison of the various approaches (Table 1). Minimizing loss in the DT interpretation of LRTs is equivalent to minimizing type II error (for a fixed type I error). The loss function for the AIC is the K-L distance, that is, the information lost by use of an assumed model rather than the true model. In Bayesian model selection, if we assume uniform priors across models, a binary loss function is proportional to the inverse of the posterior probability of a model, given the data. In performance-based methods, a nonbinary loss function can be erected on the basis of any feature of an analysis that one deems important to method performance (such as expected branch-length error). The derivations of these methods in the decision-theory framework is provided in the Supplementary Material available online at http://www.annualreviews.org/. Of the methods examined here, all but LRTs can easily be incorporated into model averaging, either manually (e.g., Posada & Buckley 2004) or through incorporation into reversible-jump Markov chain Monte Carlo (e.g., Huelsenbeck et al. 2004). Given the increasing numbers of taxa in phylogenetics data sets and the advantages of using partitioned models (e.g., Castoe et al. 2004, Nylander et al.

TABLE 1 Model-selection approaches interpretable from the perspective of decision theory

Approach[a]	Loss	Decision rule	Philosophy	Comments
hLRT	Binary	Minimize type II error rate	Non-Bayesian	Assume a fixed type I error rate
AIC	Nonbinary	Miminize Kullback-Leibler distance	Non-Bayesian	Assume candidate models are close to true model; Taylor expansion approximation[b]
BIC	Binary	Maximize posterior probability	Bayesian	Assume uniform priors across models; Taylor expansion approximation[b]
Performance based	Nonbinary	Minimize risk based on any feature of analysis (e.g., branch-length error)	Bayesian	Performance-measure dependence; Taylor expansion approximation[b]

[a]The derivations for interpreting these approaches in this framework are presented in the Supplemental Material online at http://www.annualreviews.org/.

[b]The Taylor expansion approximation permits priors across model parameters to be ignored and evaluation of a model at its joint maximum-likelihood estimates (Raftery 1995).

2004), simply choosing the most complex model available may result in loss of predictive ability and nonidentifiability of model parameters, both a function of too few degrees of freedom. Simulation studies with extremely complex models of sequence evolution to generate data (e.g., Minin et al. 2003) are likely to be very fruitful in evaluating alternative model-selection and model-averaging strategies.

ACKNOWLEDGMENTS

We thank Z. Abdo, D. Althoff, K. Segraves, B. Shaffer, and D. Vanderpool for critiquing the manuscript and for their many helpful comments. This work is part of the University of Idaho Initiative in Bioinformatics and Evolutionary Studies (IBEST). Funding was provided by NSF EPS-0080935 (IBEST), NSF Systematic Biology DEB-9974124 (J.S.), NSF Probability and Statistics DMS-0072198 (P.J.), NSF EPS-0132626 (P.J.), NSF Population Biology DEB-0089756 (P.J.), and NIH NCCR 1P20PR016448-01 (IBEST: PI, L.J. Forney). Long-term interactions with several excellent scientists outside of IBEST have contributed to our thinking about model selection. They include T. Buckley, K. Crandall, V. Minin, D. Posada, C. Simon, and D. Swofford.

LITERATURE CITED

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

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, ed. PN Petrov, F Csaki. pp. 267–81. Budapest: Akad. Kiado

Anderson FE, Swofford DL. 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. *Mol. Phylogenet. Evol.* 33:440–51

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.* 19:1171–80

Box GEP. 1976. Science and statistics. *J. Am. Stat. Assoc.* 71:791–99

Brown W, Prager EM, Wang A, Wilson AC. 1982. Mitochondrial DNA sequences of primates. *J. Mol. Evol.* 18:225–39

Bruno WJ, Halpern AL. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.* 16:564–66

Buckley TR. 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst. Biol.* 51:509–23

Buckley TR, Cunningham CW. 2002. The effects of nucleotide substitution model assumptions on estimates of non-parametric bootstrap support. *Mol. Biol. Evol.* 19:394–405

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

Burnham KP, Anderson DA. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* New York: Springer-Verlag. 488 pp. 2nd ed.

Burnham KP, Anderson DA. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Method Res.* 33:261–304

Carstens BC, Stevenson AL, Degenhardt JD, Sullivan J. 2004. Testing nested phylogenetic and phylogeographic hypotheses in the *Plethodon vandykei* species group. *Syst. Biol.* 53:781–92

Castoe TA, Doan TM, Parkinson CL. 2004. Data partitions and complex models in Bayesian analysis: the phylogeny of gymnophthalmid lizards. *Syst. Biol.* 53:448–59

Cicero C, Johnson N. 2001. Phylogeny and character evolution in the Empidonax group of tyrant flycatchers (Aves: Tyrannidae): a test of W.E. Lanyon's hypothesis using mtDNA sequences. *Mol. Phylogenet. Evol.* 22:289–302

Cunningham CW, Zhu H, Hillis DM. 1998. Best-fit maximum likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution* 52:978–87

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

Erixon P, Svennblad B, Britton T, Oxelman B. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst. Biol.* 52:665–73

Felsenstein J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.* 27:401–10

Felsenstein J. 2001. Taking variation of evolutionary rates between sites into account in

inferring phylogenies. *J. Mol. Evol.* 53:447–55

Felsenstein J. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer. 664 pp.

Fisher RA. 1958. *Statistical Methods for Research Workers*. New York: Hafner. 239 pp. 13th ed.

Fitch WM, Markowitz E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* 4:579–93

Foster PG. 2004. Modeling compositional heterogeneity. *Syst. Biol.* 53:485–95

Frati F, Simon C, Sullivan J, Swofford DL. 1997. Evolution of the mitochondrial COII gene in Collembola. *J. Mol. Evol.* 44:145–58

Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.* 18:866–73

Galtier N, Guoy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* 15:871–79

Gaut BS, Lewis PO. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.* 12:152–62

Gillespie JH. 1986. Rates of molecular evolution. *Annu. Rev. Ecol. Syst.* 17:636–65

Golding GB. 1983. Estimates of DNA and protein sequence divergence: an examination of some assumptions. *Mol. Biol. Evol.* 1:125–42

Goldman N. 1993. Statistical tests of models of DNA substitution. *J. Mol. Evol.* 36:182–98

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11: 511–23

Gu X, Fu YX, Li WH. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.* 12:546–57

Halpern A, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.* 15:910–17

Han HY, Ro KE. 2005. Molecular phylogeny of the superfamily Tephritoidea (Insecta: Diptera): new evidence from the mitochondrial 12S, 16S, and COII genes. *Mol. Phylogenet. Evol.* 34:416–30

Hasegawa M. 1990. Phylogeny and molecular evolution of primates. *Jpn. J. Genet.* 65:243–65

Hasegawa M, Kishino H, Yano T. 1985. Dating the human-ape split by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22:160–74

Hueslenbeck JP. 2002. Testing a covariotide model of DNA substitution. *Mol. Biol. Evol.* 19:698–707

Huelsenbeck JP, Crandall KA. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* 28:437–66

Huelsenbeck JP, Hillis DM. 1993. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* 42:247–64

Huelsenbeck JP, Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* 904–13

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17:754–55

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310–14

Huelsenbeck JP, Larget B, Alfaro M. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol. Biol. Evol.* 21:1123–33

Johnson JP, Omland KS. 2004. Model selection in ecology and evolution. *Trends Ecol. Evol.* 19:101–08

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*, ed. N Munro, pp. 21–132. New York: Academic

Kass RE, Raftery AE. 1995. Bayes factors. *J. Am. Stat. Assoc.* 90:773–95

Kimura M. 1980. A simple model for estimating evolutionary rates of base substitutions between homologous nucleotide sequences. *J. Mol. Evol.* 16:111–20

Kjer K. 2004. Aligned 18S and insect phylogeny. *Syst. Biol.* 53:506–14

Lemmon AR, Moriarty EC. 2004. The importance of proper model assumption in Bayesian phylogenetics. *Syst. Biol.* 53:265–77

López-Fernández H, Honeycutt RL, Winemiller KO. 2005. Molecular phylogeny and evidence for an adaptive radiation of geophagine cichlids from South America (Perciformes: Labroidei). *Mol. Phylogenet. Evol.* 34:227–44

Metzker ML, Mindell DP, Liu X, Ptak RG, Gibbs RA, Hillis DM. 2002. Molecular evidence of HIV-1 transmission in a criminal case. *Proc. Natl. Acad. Sci. USA* 99:14293–97

Minin V, Abdo Z, Joyce P, Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. *Syst. Biol.* 52:674–83

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitutions rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11:1139–51

Nylander JAA, Ronquist F, Huelsenbeck JPP, Nieves-Aldrey JL. 2004. Bayesian phylogenetic analysis of combined data. *Syst. Biol.* 53:47–67

Pol D. 2004. Empirical problems of the hierarchical likelihood ratio test for model selection. *Syst. Biol.* 53:949–62

Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* 53:793–808

Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14:817–18

Posada D, Crandall KA. 2001. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* 50:580–601

Raftery AE. 1995. Bayesian model selection in social research (with discussion by A Gelman, DB Rubin, and RM Hauser). In *Sociological Methodology*, ed. PV Marsden, pp. 111–96. Oxford, UK: Blackwell Sci.

Raftery AE. 1996. Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice*, ed. WR Gilks, S Richardson, DJ Speigelhalter, pp. 163–87. New York: Chapman & Hall

Rannala B. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Syst. Biol.* 51:754–60

Sanderson MJ, Kim J. 2000. Parametric phylogenetics? *Syst. Biol.* 49:817–29

Schwarz G. 1978. Estimating the dimensions of a model. *Ann. Stat.* 6:461–64

Siddall ME. 1998. Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris zone. *Cladistics* 14:209–20

Siddall ME, Kluge AG. 1997. Probabilism and phylogenetic inference. *Cladistics* 13:313–36

Smith AD, Lui TWH, Tillier ERM. 2004. Empirical models for substitution in ribosomal RNA. *Mol. Biol. Evol.* 21:419–27

Suchard MA, Weiss RE, Sinsheimer JS. 2002. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* 18:1001–13

Sullivan J. 2005. Maximum-likelihood estimation of phylogeny from DNA sequence data. In *Molecular Evolution, Producing the Biochemical Data. Part B. Methods in Enzymology*, ed. E Zimmer, E Roalson. In press

Sullivan J, Swofford DL. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J. Mammal. Evol.* 4:77–86

Sullivan J, Swofford DL. 2001. Should we use model-based methods for phylogenetic inference when we know assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Syst. Biol.* 50:723–29

Sullivan J, Arellano EA, Rogers DS. 2000.

Comparative phylogeography of Mesoamerican highland rodents: concerted versus independent responses to past climatic fluctuations. *Am. Nat.* 155:755–68

Sullivan J, Holsinger KE, Simon C. 1996. The effect of topology on estimates of among-site rate variation. *J. Mol. Evol.* 42:308–12

Sullivan J, Markert JA, Kilpatrick CW. 1997. Phylogeography and molecular systematics of the *Peromyscus aztecus* species group (Rodentia: Muridae) inferred using parsimony and likelihood. *Syst. Biol.* 46:426–40

Sullivan J, Swofford DL, Naylor GJP. 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Mol. Biol. Evol.* 16:1347–56

Swofford DL. 1998. *PAUP*: phylogenetic analysis using parsimony (*and other methods)*. Version 4.0b3a. Sunderland, MA: Sinauer Assoc. CD-ROM

Swofford DL, Sullivan J. 2003. Phylogenetic inference using parsimony and maximum likelihood using PAUP*. In *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny*, ed. M Salemi, AM Vandamme, pp. 160–96. Cambridge, UK: Cambridge Univ. Press

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In *Molecular Systematics*, ed. DM Hillis, C Moritz, BK Mable, pp. 407–514. Sunderland, MA: Sinauer Assoc. 2nd ed

Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* 50:525–39

Tamura K, Nei M. 1993. Estimation of the number of nucleotides substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10:512–26

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* 17:57–86

Tillier ERM, Collins RA. 1995. Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. *Mol. Biol. Evol.* 12:7–15

Tuffley C, Steele M. 1998. Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* 147:63–91

Uzzell T, Corbin KW. 1971. Fitting discrete probability distributions to evolutionary events. *Science* 172:1089–96

Waddell P, Penny D. 1996. Evolutionary trees of apes and humans from DNA sequences. In *Handbook of Symbolic Evolution*, ed. AJ Lock, CR Peters, pp. 53–73. Oxford: Clarendon

Whelan S, Lio P, Goldman N. 2001. Molecular phylogenetics: state of the art methods for looking into the past. *Trends Genet.* 17:262–72

Worobey M, Santiago ML, Keele BF, Ndjango JBN, Joy JB, Labamall BL, et al. 2004. Origin of AIDS: contaminated polio vaccine theory refuted. *Nature* 428:820

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10:1396–1401

Yang Z. 1994. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39:105–11

Yang Z. 1997. How often do wrong models produce better phylogenies? *Mol. Biol. Evol.* 14:105–08

Yang Z, Roberts D. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.* 12:451–58