Syst. Biol. 50(6):913–925, 2001

A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data

PAUL O. LEWIS

Department of Ecology and Evolutionary Biology, The University of Connecticut, Storrs, Connecticut 06269-3043, USA; E-mail: paul.lewis@uconn.edu

Abstract.—Evolutionary biologists have adopted simple likelihood models for purposes of estimating ancestral states and evaluating character independence on specified phylogenies; however, for purposes of estimating phylogenies by using discrete morphological data, maximum parsimony remains the only option. This paper explores the possibility of using standard, well-behaved Markov models for estimating morphological phylogenies (including branch lengths) under the likelihood criterion. An important modification of standard Markov models involves making the likelihood conditional on characters being variable, because constant characters are absent in morphological data sets. Without this modification, branch lengths are often overestimated, resulting in potentially serious biases in tree topology selection. Several new avenues of research are opened by an explicitly model-based approach to phylogenetic analysis of discrete morphological data, including combined-data likelihood analyses (morphology + sequence data), likelihood ratio tests, and Bayesian analyses. [Discrete morphological character; Markov model; maximum likelihood; phylogeny.]

The increased availability of nucleotide and protein sequences from a diversity of both organisms and genes has stimulated the development of stochastic models describing evolutionary change in molecular sequences over time. Such models are not only useful for estimating molecular evolutionary parameters of interest but also important as the basis for phylogenetic inference using the method of maximum likelihood (ML) and Bayesian inference. ML provides a very general framework for estimation and has been extensively applied in diverse fields of science (Casella and Berger, 1990); however, the popularity of ML in phylogenetic inference has lagged behind that of other optimality criteria (such as maximum parsimony), primarily because of its much greater computational cost for evaluating any given candidate tree. Recent developments on the algorithmic aspects of ML inference as applied to phylogeny reconstruction (Olsen et al., 1994; Lewis, 1998; Salter and Pearl, 2001; Swofford, 2001) have succeeded in reducing this computational cost substantially, and ML phylogeny estimates involving hundreds of terminal taxa are now entering the realm of feasibility. Bayesian methods (based on a likelihood foundation) offer the prospect of obtaining meaningful nodal support measures without the unreasonable computational burden imposed by existing methods such as bootstrapping (Rannala and Yang, 1996; Yang and Rannala, 1997; Larget and Simon, 1999; Mau et al., 1999; Huelsenbeck, 2000a). Furthermore, the Bayesian approach makes it possible to test hypotheses involving phylogenies without depending on any particular hypothesized tree (e.g., Huelsenbeck, 2000b), so likelihood models are expected to play an ever-increasing role in systematics and related disciplines.

ML, least squares, and minimum evolution are all distinguished from maximum parsimony in being model-based optimality criteria. ML and maximum parsimony are similar in being discrete character methods, unlike minimum evolution and least squares, which are based on a matrix of pairwise evolutionary distances between terminal taxa. Despite the early availability of a likelihood model for continuous traits (Felsenstein, 1973), the use of model-based optimality criteria has heretofore been restricted primarily to molecular data, with maximum parsimony being the only criterion applied to both discrete morphological and molecular data. Models have been applied to discrete morphological traits, but the purpose of these models has been to infer ancestral states (e.g., Schluter et al., 1997; Mooers and Schluter, 1999; Pagel, 1999), to assess the magnitude of the evolutionary correlation between different traits (Pagel, 1994), or to investigate the properties of other optimality criteria (Felsenstein, 1981a), but not for phylogeny reconstruction per se.

Although no one has suggested using likelihood for estimating trees, two models have

been previously described for purposes of investigating properties of the parsimony method. Goldman (1990) described a simple likelihood model (hereafter referred to as the G90 model) that always chooses the exact same tree (or trees) as equal-weighted Fitch parsimony. Later, Penny et al. (1994) and Tuffley and Steel (1997) found that a very different model (hereafter, the TS97 model) also selects trees identical to those selected by parsimony. The G90 model has only one branch (i.e., edge) length parameter that governs the probability of observing a change across any branch of the tree; however, the model requires implicit estimation of the ancestral character states at each interior node of the tree. Goldman (1990) emphasized that a negative side effect of these nuisance parameters, the number of which grows with the number of characters, is likely to be statistical inconsistency. A method is statistically consistent if the estimates produced by the method come closer to the true value of the quantity being estimated as the sample size increases to infinity (Casella and Berger, 1990:323). Statistical consistency is thus a desirable asymptotic property of a statistical inference method, as has been pointed out numerous times with respect to the choice of likelihood versus parsimony methods (e.g., Felsenstein, 1978).

The TS97 model is also very parameter-rich. For a problem involving n taxa and m characters, the TS97 model has effectively $m(2n-3)$ separate parameters (a separate parameter for every branch/character combination). This model was called the "no common mechanism" model by Tuffley and Steel because it allowed the rate of evolution for one particular branch and one particular character to be independent of the rate for any other branch and every other character. Tuffley and Steel (1997:599) cautioned, however, that "...the number of parameters being estimated grows linearly with the number of characters, so the statistical consistency of these two methods is not guaranteed by standard results. Indeed, the former method can be provably statistically inconsistent...." Here, "former method" refers to the "no common mechanism" model. The G90 and TS97 models thus have very little in common except the fact that they are both parsimony models (i.e., the set of tree topologies chosen is identical to the set chosen by parsimony) and the number of parameters in both grows as a function of the number of characters.

Goldman (1990) emphasized the importance of using only *structural* parameters (parameters that appear in the likelihood function for all characters) and avoiding the use of *incidental* parameters (parameters that appear in the likelihood functions for only some characters) in models used for phylogenetic inference. In the classical models currently used for ML phylogeny reconstruction, all parameters are structural parameters. For example, the transition/transversion rate ratio parameter used in the HKY85 model (Hasegawa et al., 1985) is necessary for calculating the likelihood for every site, and the same can be said for any branch length parameter and any nucleotide frequency parameter in this model. In contrast, the ancestral states estimated in the G90 model are incidental parameters, since their value is only used in calculating the likelihood associated with a single character. Likewise, the branch probability parameters of the TS97 model are incidental parameters because each is used in computing the likelihood for only one character. Models incorporating incidental parameters are susceptible to problems with statistical inconsistency, and Goldman (1990) noted that the presence of incidental parameters can make estimates of the structural parameters in the model inconsistent as well. There is a growing tendency to discount the importance of statistical consistency in phylogeny inference (e.g., Farris, 1999); however, avoiding (where possible) models that may be statistically inconsistent even when their assumptions are not violated seems prudent. The G90 and TS97 parsimony models both have this property.

The purpose of this paper is to discuss the applicability of ML phylogeny inference to discrete morphological data. The TS97 model provides an excellent comparison because it gives results identical to parsimony, currently the only option for phylogenetic analyses involving discrete morphological characters. The G90 model is less attractive for comparison because its assumption of equal branch lengths and estimated ancestral states make it substantially different from the models currently used in phylogenetics for sequence data. In the terminology of Steel and Penny (2000), TS97 and the standard substitution models used

for sequence data are all "maximum average likelihood" methods, whereas G90 is in a different class, the "most-parsimonious likelihood" methods. In this paper, I strongly emphasize avoiding incidental parameters so that the model will be well-formulated and statistically well-behaved. I also show that standard Markov models, that is, generalizations of the Jukes and Cantor (1969:JC69) model, represent modified versions of the TS97 model and avoid the aforementioned problems with incidental parameters that lead to potential statistical inconsistency. Discussion will center around whether the modifications necessary to make the TS97 model statistically sound are biologically justified. I conclude with a discussion of interesting extensions to the basic model and touch on the wealth of opportunities that model-based approaches open up for systematic biologists.

A Basic Likelihood Model for Discrete Morphological Data

The model adopted here for ML phylogenetic analyses of discrete morphological data is not by any means novel. Some version of the model has been used by numerous authors (Jukes and Cantor, 1969; Neyman, 1971; Farris, 1973; Cavender, 1978; Felsenstein, 1981a; Pagel, 1994; Penny et al., 1994; Schultz et al., 1996; Schluter et al., 1997; Tuffley and Steel, 1997; Mooers and Schluter, 1999) for several different purposes, even for analysis of discrete morphological character evolution. By the time it was adopted by Jukes and Cantor, this model had already enjoyed a long history in biology, for example, forming the basis of Haldane's (1919) map distance function. Rather than propose a new model, my purpose here is to examine the consequences of applying this type of model to morphological character data and to describe modifications needed to accommodate the peculiarities of discrete morphological data sets. I will hereafter use the acronym Mk to refer to this family of models (where the "M" stands for "Markov" and "k" refers to the number of states observed). The Mk model is a generalized JC69 model, the latter representing the special case of $k = 4$ (the JC69 model could thus be referred to as the M4 model). The Mk model assumes that a lineage is always in one of k possible states ($k \geq 2$), with no state considered plesiomorphic or apomorphic a priori. Along a particular branch of the phylogeny, a character can change state at any instant in time, with the probability of such an event being equal for all such time intervals along the branch. An instant is defined to be an infinitesimal period of time, denoted dt, during which there can be at most one substitution (= change of state) event. Different instantaneous time periods are independent of one another with respect to the probability of a character state change, and the probability of change is symmetrical (i.e., the instantaneous probability of changing from state i to state j is the same as the instantaneous probability of changing from j to i). The length of a branch under the Mk model is defined to be the expected number of changes per character across the branch, which is equal to $(k-1)\alpha t$, where α is the instantaneous rate of any particular transition between states, and t is the amount of time represented by the branch. The $k \times k$ instantaneous rate matrix for the Mk model is

$$Q = \alpha \begin{bmatrix} 1-k & 1 & \cdots & 1 \\ 1 & 1-k & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1-k \end{bmatrix}$$

(Tuffley and Steel, 1997; their eq. 4), the transition probabilities are

$$P_{ii}(t) = \frac{1}{k} + \frac{k-1}{k} e^{-k\alpha t},$$

$$P_{ij}(t) = \frac{1}{k} - \frac{1}{k} e^{-k\alpha t},$$

and the stationary distribution (relative frequencies of the states at equilibrium) is the vector $[\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}]$ of length k. Once the model is thus specified, using it to infer phylogenies involves straightforward application of the methods outlined by Felsenstein (1981b; see also Swofford et al., 1996). The rationale for the use of ML inference in general is discussed at length in Edwards (1972).

Rationale for Using the Mk Model

The Mk model may strike many systematists as being highly unrealistic. One possible objection lies in the fact that the model predicts that the probability of observing a change along a branch in a phylogeny

increases with the amount of time associated with the branch. This appears on the surface to be explicitly gradualistic, excluding punctuated equilibria as a mode of morphological evolution. Another point of contention lies in the fact that the model allows characters to change freely back and forth between two states, which may strike many biologists as unrealistic for most discrete morphological features. Both of these concerns will be addressed at the end of this section.

My primary tool in the defense of the Mk model is the fact that Tuffley and Steel's (1997) parsimony model (the TS97 model), when used within the framework of ML inference, has the property of always choosing the same tree(s) as equally-weighted parsimony, even to the point of choosing multiple trees if there are multiple most-parsimonious solutions. That is, the likelihood under TS97 is a monotonically decreasing function of the parsimony score, meaning that a likelihood analysis using TS97 is identical to a parsimony analysis using equally-weighted parsimony (the tree that minimizes the number of steps also maximizes in L). Thus, the justification for using likelihood for morphological data instead of parsimony hinges on whether the differences between Mk and TS97 are acceptable from both a biological and a statistical standpoint.

The TS97 model collapses to the Mk model if each branch of the tree is assumed to have the same length for all characters. This reduces the number of parameters in the model by a factor of m, where m is the number of characters. More importantly, this restriction converts a model consisting entirely of incidental parameters (TS97) to one consisting entirely of structural parameters (Mk). The addition of more data in the form of characters to the Mk model thus provides more information relevant to estimating this fixed number of parameters (the branch lengths), whereas adding one more character to the TS97 model is used only in estimating the additional $2n - 3$ branch lengths specific to that character.

Placing restrictions on the number of branch lengths makes the Mk model more statistically reliable, but also results in an apparent loss of biological realism: A character is no longer (under Mk) allowed to change its rate to an arbitrary value from branch to branch (as it is allowed to do in the TS97 model). If the TS97 model is examined closely, however, determining which of the two models is actually more biologically realistic becomes less clear. ML estimators tend to adopt extreme values when the amount of data applicable to their estimation drops below a critical threshold. A simple example concerns estimating the proportion of heads for a coin. If we are not willing to assume that the coin is perfectly fair, and thus estimate the proportion of heads (p) rather than assume its value is 0.5, the formula for the maximum likelihood estimator is simply the number of heads observed divided by the total number of flips. If the coin is flipped just once, the ML estimate using this formula will either be 0.0 or 1.0, depending on whether a tail or a head was observed, respectively. These extreme values give way to much more reasonable estimates if the coin is flipped many times, of course. If p is treated as a structural parameter, adding more data is beneficial to the estimation process because the model is statistically consistent. If, instead, a new parameter (p_i) is added to the model for every flip i (i.e., we are not willing to assume that the probability of heads is the same from one flip to the next), the estimates for all of these parameters will be either 0.0 or 1.0, and adding data from flip j clearly will not help refine the estimates of parameters p_i ($i < j$).

A similar pathology affects the TS97 model, at least for the two-state case. The ML estimates of the branch lengths in the TS97 model are all either ∞ or 0.0, depending on whether a parsimony reconstruction would yield a change or no change across the branch, respectively. In a direct analogy to the coin-flipping example, there is simply too little information to reliably estimate $2n - 3$ parameters with only n observations. Thus, TS97 is also less than realistic from a biological standpoint: Few biologists would be comfortable with a morphological trait changing state an infinite number of times along a particular branch. Although Mk could (with a suitable rate heterogeneity model attached) be made to allow rates to vary across characters, it would allow only tree contraction or expansion in this case. The TS97 model allows each character to have a different rate on every branch (an "anything goes" approach), but this results in estimates of rates that are not realistic (the rate of change is either zero or infinity).

It should be pointed out (see the example in Tuffley and Steel, 1997:596–597) that these

extreme branch length estimates are characteristic of TS97 only in the two-state case. When the number of states exceeds two, some of the branch lengths are nonidentifiable, meaning the likelihood is the same for any possible value. Although perhaps more biologically acceptable, this behavior is nevertheless a symptom of model overparameterization.

Before leaving the topic of biological realism, I will briefly address the two issues raised at the outset. The first objection concerned the fact that the probability of change increases with time in the Mk model. This might seem to enforce a gradualist perspective on morphological evolution. In fact, it is only the average amount of change (functions involving the product αt) that appears in the transition equations for the Mk model. It does not matter whether this average amount was realized in one bout of change (at the speciation event, for example) or gradually over the entire period represented by the branch.

The second objection concerned the fact that characters are allowed to change state numerous times under the Mk model. However, a modified Mk model can be created with the stipulation that the character can change either once or not at all. Interestingly, if rate homogeneity across characters is assumed, the ML score of a tree under such a model is identical to the ML score under the Mk model. The rate of evolution required to achieve a specific probability of observing a difference across a branch differs between these two models, but both explain the data equally well and thus yield identical likelihood scores. This equivalence means there is no disadvantage to using the Mk model: It will never lead us to prefer a different tree from the modified version in which characters are allowed to change only at most one time on any given branch.

The Problem of Constant Characters

A potential obstacle to the use of the Mk model lies in the fact that systematists never record characters if they are constant (i.e., if every taxon in the analysis has the exact same state for the character). In fact, it is difficult to imagine a way in which a set of morphological characters could be circumscribed such that the "proper" number of constant characters is included in the data set. This *acquisition bias* never arises in the use of likelihood for molecular phylogenetics because the linear nature of nucleic acids and proteins allows easy circumscription of a range of characters, including constant as well as variable characters. Acquisition bias is problematic because mean rates of evolution embodied in the branch length parameters will be overestimated if only variable characters are present in the dataset. Because branch lengths play an important role in determining the overall likelihood for a particular tree topology, such overestimation, if not corrected, would lead to bias in tree topology inferences. Fortunately, one can correct for acquisition bias, and the remainder of this section is devoted to an explanation of how this is accomplished for the Mk model.

Characters can be divided into (1) parsimony-informative characters, which can potentially have different parsimony character lengths on different trees; (2) autapomorphic characters, which are variable but have the same length on all trees; and (3) constant characters, which have only one state. Because autapomorphic characters are considered uninformative by many systematists using parsimony, these too are often left out of data sets used in phylogenetic analyses (for an exception, see Funk and Wagner, 1995). One of the reasons likelihood methods resist long-branch attraction problems is that they can accept an explanation of similarity based on convergent or parallel evolution, whereas parsimony allows only historical explanations of similarity among the terminal taxa. Branch length estimates determine whether likelihood is willing to accept, so to speak, an explanation based on convergence or parallelism over an explanation based solely on history (i.e., simple inheritance). Likelihood methods may choose to keep separate two lineages having very long branches because the evidence for convergence/parallelism is (in this case) stronger than the evidence for shared history (Felsenstein, 1978). If branch lengths are incorrectly estimated (say, consistently overestimated), likelihood methods would be biased in their choice of tree topology. Thus, it is important to in some way correct for the systematic omission of constant characters. Autapomorphic and highly variable characters do not present the same problem as constant characters, being identifiable and at least enumerable.

The technique for freeing the Mk model of this problem is borrowed from Felsenstein (1992), who encountered a similar problem in the analysis of restriction site data. The solution involves computing a conditional likelihood instead of the normal likelihood, the condition being that only variable characters are present in the data. The likelihood for character c (L_c) is proportional to the probability of the data (D_c) for character c, given the parameters of the model, which for the Mk model comprises the tree topology (T) and branch lengths (Λ):

$$L_c(T, \Lambda \mid D_c) \propto \Pr(D_c \mid T, \Lambda) \qquad (1)$$

If event E corresponds to the case in which character c is variable, the likelihood (as it is normally computed) can be written:

$$L_c(T, \Lambda \mid D_c) \propto \Pr(D_c, E \mid T, \Lambda) \qquad (2)$$

The likelihood for character c conditional on E is thus proportional to:

$$\Pr(D_c \mid T, \Lambda, E) = \frac{\Pr(D_c, E \mid T, \Lambda)}{\Pr(E)} \qquad (3)$$

where $\Pr(E)$ refers to the probability that evolution would have created a character that is variable. This quantity is just $1 - \Pr(\text{not } E)$, where $\Pr(\text{not } E)$ is the probability that evolution would create a constant character. $\Pr(\text{not } E)$ can be obtained by using a dummy character having the same state at all terminal nodes (see Felsenstein, 1992). The numerator on the right side of Eq. 3 is simply the likelihood as it would normally be computed, with the quotient being the conditional likelihood used to account for the acquisition bias in the data.

To illustrate the importance of making this correction, a computer simulation was performed in which data were generated according to the Mk model and the tree in Figure 1; constant characters were thrown away; and branch length estimates and the preferred tree topology were recorded under both the uncorrected Mk model and the version corrected for acquisition bias. The results (Table 1) indicate that one can easily obtain overestimated branch lengths unless the conditional likelihood is utilized, and (for this specific example) the probability of reconstructing the correct tree dropped from 0.998 (conditional likelihood approach) to 0.740 (uncorrected likelihood approach).

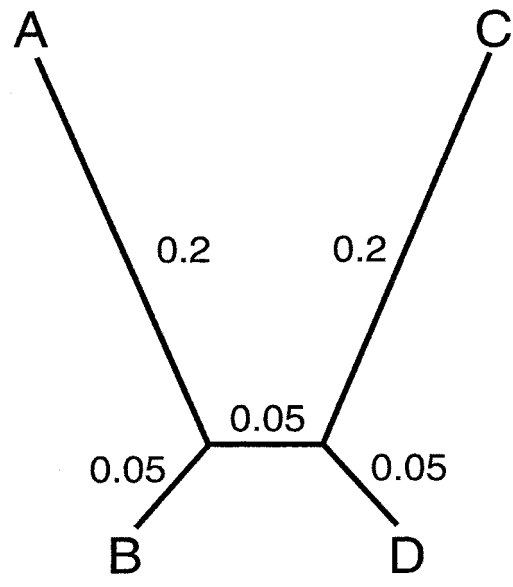

FIGURE 1. Model tree used for simulations to illustrate the importance of conditioning the likelihood on the fact that all characters are variable. The numbers beside the branches represent the expected number of changes per character along the branch.

To avoid possible confusion between the use of the Mk model and the version that conditions on variable characters, I will use the term Mkv to refer specifically to the conditional version of the Mk model.

EXTENDING THE MK MODEL

The Mk model may be easily extended to accommodate other factors deemed important in the evolution of morphological characters. A recent emphasis in models for nucleotide sequence data involves allowing rates of evolution to vary across sites (rate heterogeneity) instead of assuming a constant rate across all sites (rate homogeneity) (Churchill et al., 1992; Reeves, 1992; Sidow and Speed, 1992; Yang, 1993, 1994; Felsenstein and Churchill, 1995). Rate heterogeneity is possibly as much of a factor in morphological character data as in sequence data. Allowing rate heterogeneity among characters for morphological data is therefore reasonable, and is easily accomplished by using the same techniques currently available for sequence data under a likelihood framework. Of the three basic methods in use, assuming a discrete gamma distribution

TABLE 1. Results of simulations designed to illustrate the importance of correcting for acquisition bias. The model tree used for the simulations was (A:0.2, B:0.05, (C:0.2, D:0.05)X:0.05), for each simulated data set, characters were generated until 200 variable characters were obtained (thus, the overall number of simulated characters varied among simulation replicates). Each simulated dataset was analyzed with both the uncorrected model and the Mk model corrected by conditioning on the fact that only variable characters were observed (i.e., the Mkv model). The means (±SD) over 1,000 replicates are reported. Branch length estimates from the uncorrected model were edited such that values exceeding 1×10^6 were set to exactly 1×10^6. Thus, the actual means (and SDs) for the uncorrected model were greater than the numbers shown here.

	True branch length	Mk (uncorrected)	Mkv (corrected)
Percent correct	—	74.0	99.8
Branch A	0.2	241,750 (±349,100)	0.206 (±0.060)
Branch B	0.05	0.43210 (±0.13756)	0.050 (±0.018)
Branch X	0.05	54.646 (±1,725.3)	0.052 (±0.023)
Branch C	0.2	143,950 (±228,910)	0.206 (±0.059)
Branch D	0.05	0.022 (±0.054)	0.051 (±0.019)

for the relative rates (Yang, 1994) is probably the most appropriate for discrete morphological characters. The hidden Markov model of Felsenstein and Churchill (1995) requires estimation of more parameters than does the discrete gamma approach (which requires estimation of only a single extra structural parameter, the gamma shape parameter α), but has the advantage of not requiring the assumption that the relative rates are gamma-distributed. The hidden Markov model allows for correlation between neighboring characters, but this is not useful in the context of morphological characters because the position of characters within the data matrix is arbitrary. The invariant sites models (Churchill et al., 1992; Reeves, 1992; Sidow and Speed, 1992; Steel et al., 2000) assume that a certain fraction of characters do not evolve at all (i.e., they have a change rate of 0), which does not apply if all characters scored are variable.

Another extension to the Mk model of interest to systematists would allow the rate of change from state i to state j to differ from the rate governing the opposite change (from j to i). This "unequal rate" Mk model is uncomplicated for two-state characters, and the transition probabilities for this generalization are available in Taylor and Karlin (1984:256) and Schultz et al. (1996), among other places. Because such models imply a specific stationary (equilibrium) distribution for the character states, it is customary (to save time) in nucleotide sequence models to assume stationarity and use the empirical frequencies of the different nucleotides in lieu of estimating the base frequencies by maximum likelihood. If there are no missing data or gaps, the empirical frequency of the base A is simply the number of A's in the data matrix divided by the total number of nucleotides (number of taxa times the number of sites). Using the same approach for morphological characters would be meaningless, because state 0 for one character is not at all the same thing as state 0 for any other character in the data matrix. Thus, application of the unequal-rate Mk model makes sense for evaluating single characters, but application of the model for multiple characters involves the unrealistic assumption that all characters share the same ratio of forward to reverse rates of change (despite the fact that the states for different characters receive their 0 or 1 designation arbitrarily!).

The most statistically sound way around this complication would involve assuming a distribution for the equilibrium frequency of state 0 across characters. For example, the frequency of state 0 (π_0) could be assumed to have a Beta distribution. The Beta distribution is determined by two parameters, a and b, which would be the only additional parameters estimated from the data, and because data from all characters would participate in estimating a and b, these are structural parameters and hence in keeping with the ban on the use of incidental parameters. This is similar to the assumption of a gamma distribution for relative rates across characters, in which a single additional shape parameter is estimated from the data.

AN EXAMPLE

A recent paper by Quicke and Belshaw (1999) examined incongruence within a dataset based on the morphology of parasitic

wasps in the family Braconidae (Hymenoptera). Much of the incongruence within this dataset appears to be the result of convergence in morphology associated with the endoparasitic lifestyle of many of the wasps in this family. Excluding all characters except those related to the female reproductive system and larval development (their FEMALE + LARVAL character set), Quicke and Belshaw's parsimony analyses suggested a single origin of endoparasitism in the Braconidae, and even placed an endoparasitoid outgroup taxon *Alomya* among the endoparasitoid braconids. The parsimony analysis was repeated by using PAUP* 4b3 (Swofford, 2001) to obtain the 24,620 most-parsimonious trees at 128 steps. The strict consensus tree from this analysis is shown in Figure 2A. The log-likelihoods of these trees under the basic Mkv model (i.e., conditioning on characters being variable but without the other possible extensions such as rate heterogeneity or frequency heterogeneity) ranged from −478.68557 to −484.39209. None of the most-parsimonious trees was equivalent to the ML tree under the Mkv model (Fig. 2B), for which the log-likelihood was −472.98025. The ML tree is thus 5.7 log-likelihood units better than any of the most-parsimonious trees, and its 135 steps are

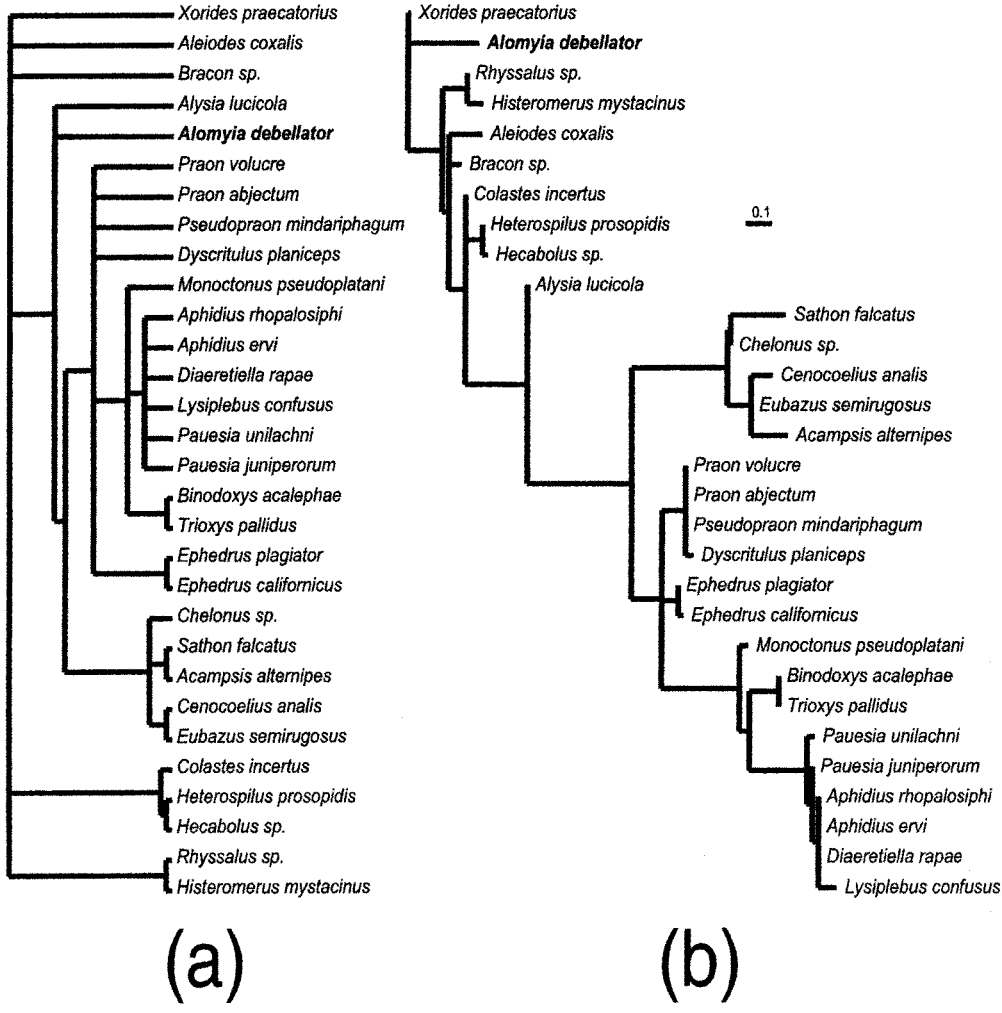

FIGURE 2. Parsimony and likelihood results for the parasitic wasp data (FL/EM characters only) of Quicke and Belshaw (1999). (a) Strict consensus of 24,620 most-parsimonious trees (128 steps). (b) The single maximum likelihood tree under the Mkv model ($\ln L = -472.98025$).

7 steps longer than the most-parsimonious trees. This points out that the Mkv model, unlike either the Goldman (1990) model or the Tuffley and Steel (1997) model, is not equivalent to parsimony in its choice of trees. Interestingly, the ML tree places the outgroup *Alomya* outside all of the ingroup taxa, which is the position it occupies when molecular data or all of the morphological data, or both, are used. Although the ML tree for just the FEMALE + LARVAL characters does not coincide in all respects to the molecular or full morphological trees, it does appear to correct the most egregious problem involving the placement of *Alomya*.

DISCUSSION

Adaptive Convergence and the Autapomorphy Trail

The likelihood criterion differs from parsimony in that all characters are used to estimate branch lengths (another way of saying that branch lengths are structural, not incidental, parameters), which in turn are used in calculating the overall score by which different tree topologies are compared. Autapomorphies *are* phylogenetically informative under the Mk model because they provide information about the amount of evolution that has occurred along terminal lineages; this, in turn, influences the estimated lengths of terminal branches and, ultimately, the overall likelihood used to compare a tree to other trees. In parsimony, each character contributes a certain number of steps (the character length) to the overall tree length, and the number of steps contributed is totally independent of all other characters. Under the likelihood criterion, each character contributes (as in parsimony) a value that is added to the contributions of other characters to form the overall log-likelihood score, which is the likelihood equivalent of the tree length in parsimony. In likelihood, however, the actual value contributed by a character depends to some extent on information contributed by all characters. One might argue, however, that there is no logical reason (at least within the context of morphological evolution) for allowing all characters to influence, through their effect on branch length estimation, the interpretation of all other characters and hence the tree topology preferred.

A simple example can be used to illustrate a situation in which this connection between characters is reasonable. Suppose two lineages (those leading to taxon W and taxon Y) have independently adapted to the same set of environmental conditions, resulting in one instance of convergence (or mistaken homology; character 2) and numerous autapomorphies (the changes involved in the separate adaptations that were not similar enough to be confused as synapomorphies linking W and Y). On the tree (Fig. 3) are shown two parsimony-informative characters (1 and 2) and five autapomorphous characters (3–7). Assuming that the tree shown, (W,X,(Y,Z)), is indeed the true topology, character 1 is the only true synapomorphic character and character 2 is homoplasious. Both likelihood and parsimony would agree on this outcome if this tree is assumed; because parsimony only allows the use of the two "informative" characters, however, it has no basis for recommending this tree over (W,Y,(X,Z)), which makes character 2 the synapomorphy. Both of these trees require three steps automorphies ignored with the third possible tree

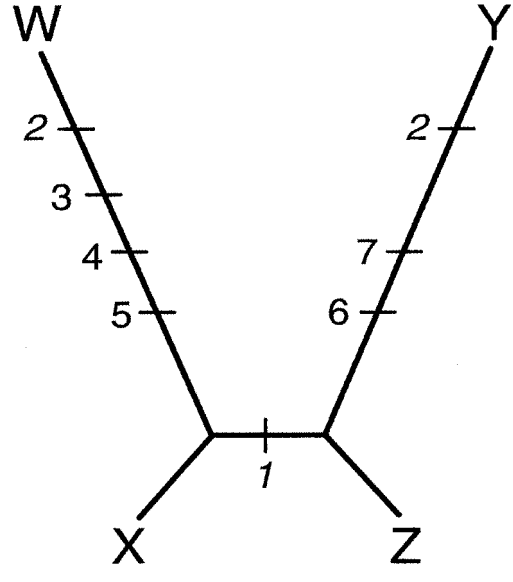

FIGURE 3. Unrooted phylogenetic tree showing the true evolutionary relationships among four taxa: W, X, Y, and Z. Two parsimony-informative characters (1 and 2) are mapped onto the tree and in italics. The character 2 is homoplasious on this tree, being the result of convergent evolution in both taxon W and taxon Y. Five autapomorphies are also mapped onto the tree. See the text for details of the example.

(W,Z,(X,Y)) requiring four steps, and thus parsimony is equivocal. From the perspective of likelihood, however, the autapomorphies are also informative, indicating that the lineages leading to taxon W and taxon Y both represent long branches (branches on which evolutionary change was more common than on other, shorter branches). Accordingly, likelihood would give more weight to explanations (of the similarity between taxon W and Y) based on convergence than to purely historical explanations and thus would prefer the tree that forces character 2 to evolve twice independently, whereas character 1 evolves only once.

In this case, likelihood is able to correctly choose which of two conflicting parsimony-informative characters is the homoplasious one, on the basis of evidence coming entirely from autapomorphies. This example is relevant because separate adaptation events leading to convergent similarities might be expected to leave behind just such a trail of autapomorphies pointing out the lineages involved. Parsimony does not take advantage of this evidence because autapomorphies are phylogenetically uninformative under the parsimony criterion. Likelihood using the Mk model has the ability to correctly diagnose this situation, given that the autapomorphies were included in the data matrix (unfortunately, autapomorphies are routinely omitted from morphological data sets, making comparisons of likelihood and parsimony difficult at present). Homoplasy, of course, has many different causes (see discussions by many authors in Sanderson and Hufford, 1996), of which adaptive convergence is only one. This suggests that careful scrutiny of all the causal factors leading to homoplasy is in order, paying attention to whether involvement of branch length estimates in the evaluation of individual characters aids or hinders the identification of such homoplasies. This would seem to be an area ripe for future study.

Advantages of a Likelihood Approach

Combining morphological data with molecular data, using the Mk model for the morphological data and a different model for the molecular data, involves straightforward addition of the log-likelihoods resulting from the separate data partitions. Other advantages involve ancestral state reconstruction, the ability to utilize likelihood ratio tests, and the ability to obtain Bayesian posterior probabilities for hypotheses involving morphology.

Ancestral state reconstruction.—There has been interest for some time in modeling discrete morphological data, but heretofore models have been applied to inferences concerning character correlation (Pagel, 1994) or ancestral state reconstruction (Maddison, 1995; Schlutz et al., 1996; Schluter et al., 1997; Mooers and Schluter, 1999; Pagel, 1999) and not to phylogeny reconstruction per se. Pagel (1994) used a model identical to the unequal rates version of M2 in his likelihood ratio test for correlated evolution between two binary morphological characters. Pagel's method does not involve using this model for phylogeny reconstruction; rather, his method assumes a given tree topology and branch lengths, with the model being applied only to the two characters of interest. Pagel's method does not compute likelihoods conditional on variable characters, but that was not necessary because branch lengths were specified, not estimated. Maddison (1995) considered stochastic models in the context of assessing the reliability of parsimony reconstructions. The model he used was essentially Goldman's parsimony model (G90) because homogeneity of branch lengths was assumed. Maddison's purpose was the calculation not of a likelihood but of the probability that the ancestral states inferred by standard parsimony were the true ancestral states. Schultz et al. (1996) reexamined the issue of reliability (raised by Frumhoff and Reeve, 1994) of the state inferred to be possessed by the common ancestor of a clade of a specified size. In this study, monomorphic (all tip nodes have the same observed state) terminal polytomies of size N were considered and the question addressed was, "How large does N need to be for it to be safe to assume that the ancestral state (at the root node of the polytomy) is identical to the states possessed by all the N tips?" The model used by Schultz et al. was identical to the model of Pagel (1994). Schluter et al. (1997) used likelihood methods to estimate ancestral states, comparing these with parsimony reconstructions for several examples. The model for discrete characters used was again identical to that used by Pagel (1994) and, as in Pagel's method, Schulter et al. assumed both topology and branch lengths a priori. Recently,

Mooers and Schulter (1999) and Pagel (1999) reexamined some of these issues as part of a symposium on ancestral state reconstruction.

All of these model-based approaches to morphological evolution allow branch lengths to play a significant role in the inferences made; however, none of them estimates branch lengths from the data collected. This paper thus presents a different, but complementary, use for models in studies of morphological character evolution. Often, branch lengths used in conjunction with Pagel's (1994) test or for making inferences concerning ancestral states are obtained by using molecular data, which might be quite inappropriate for morphological data. For example, adaptive radiations are characterized by a considerable amount of morphological evolution during a short (on a molecular scale) amount of time, leading to short branch length estimates from molecular data when morphological branch lengths are actually long. Using the Mk model for phylogeny reconstruction provides not only the topology needed for studies of character evolution and correlation but also the branch lengths that are appropriate for the suite of characters being investigated.

Likelihood ratio tests.—Model-based morphological phylogenetics opens up tremendous potential for testing explicit evolutionary hypotheses of morphological character evolution. Likelihood ratio tests have proven extremely useful in molecular phylogenetics (Huelsenbeck and Rannala, 1997), and similar successes can presumably also be achieved for morphological phylogenetics. A likelihood ratio test has already been formulated for the question, "Are these two characters correlated in their evolution such that the second tends to evolve from state 0 to state 1 following a similar change in the first character?" (Pagel, 1994). Likelihood ratio tests could also be applied to such questions as, "Are the forward and reverse rates of change significantly different for this character?" or "Is the species tree obtained using morphological data significantly different from the gene tree obtained using DNA sequences?" (see Huelsenbeck and Bull, 1996). This latter question could be addressed by comparing the log-likelihoods under two models: a constrained model in which the same tree topology is assumed for both morphological and sequence data, and an unconstrained model in which potentially different tree topologies are allowed for the two different data types. This provides an explicit means of testing whether the species tree differs from the gene tree.

Bayesian inference.—Recent advances in the application of Bayesian inference methods to phylogenetic analyses (Rannala and Yang, 1996; Yang and Rannala, 1997; Larget and Simon, 1999) are equally applicable to discrete morphological data because the likelihood function forms the foundation of Bayesian inference. Whereas likelihood methods seek to find the tree (and branch lengths) maximizing the probability of the observed data, Bayesian methods return the posterior probability, that is, the probability of the tree conditional on the observed data and the prior probability (the existing approaches specify equal prior probabilities for all possible trees). As illustrated by Larget and Simon (1999), Bayesian posterior probabilities can be obtained for individual branches in a tree in much less time than it would take to obtain bootstrap proportions by using a standard likelihood approach. Interesting applications of the Bayesian approach to questions involving morphology have already been published (Huelsenbeck, 2000b), and this new approach, because of its explicit incorporation of prior beliefs, will make possible novel ways of assessing the degree to which the data are in opposition to an investigator's convictions concerning the evolution of particular traits.

CONCLUSIONS

This paper has examined the feasibility of using models for discrete morphological character data for the purpose of inferring phylogenies. The necessary models are already available (and have been in use for other purposes for many years); bringing discrete morphological data into the likelihood framework has many advantages, including but not limited to the following:

- Providing alternatives to parsimony with respect to ancestral states, branching patterns, and the degree of homoplasy present in the data
- Allowing information on morphology to be combined in a meaningful way with sequence data for a combined analysis in which each distinct data type is modeled separately and appropriately

- Allowing testing of explicit evolutionary hypotheses of morphological character evolution by way of likelihood ratio tests
- Providing a basis for using Bayesian methods for inferring nodal support and making inferences about model parameters

The availability of likelihood ratio tests and the ability to apply Bayesian approaches increase the value of discrete morphological data in addressing phylogenetic questions. Besides being useful to systematists themselves, a likelihood model for morphology has the potential to greatly increase the usefulness of systematic work in other, related subfields of biology.

PROGRAM AVAILABILITY

The Mkv model has been incorporated into the widely used computer program PAUP* 4.0 (beta version 9; Swofford, 2001).

ACKNOWLEDGMENTS

I greatly appreciate the comments and suggestions provided by the following people on previous versions of this manuscript; however, I have chosen (probably unwisely) to not heed some of their suggestions, so I am squarely to blame for any mistakes, misunderstandings, and misrepresentations that remain: Kent Holsinger, John Huelsenbeck, Louise Lewis, Jim McGuire, Mike Steel, David Swofford, David Wagner, and the members of the PhyloBrew discussion group of the Smithsonian Museum of Natural History and the Systematics Seminar of the Department of Ecology and Evolutionary Biology at the University of Connecticut. Special thanks is due David Swofford for many productive discussions about using the Mk model for morphology, and for providing much assistance in performing calculations confirming the results presented here by incorporating the model into PAUP*. John Huelsenbeck provided excellent feedback and gentle prodding for several years and demonstrated a remarkable level of patience with me during the gestation of this paper. Special thanks goes to Jeff Thorne, who saved the day for me by pointing out that conditional likelihoods could be used to correct for acquisition bias, and to Mike Steel and Mark Holder for considerable enlightenment with respect to the TS97 model. I also extend my sincere appreciation for the funding provided by Alfred P. Sloan Foundation/National Science Foundation Young Investigator Award 98-4-5 ME.

REFERENCES

CASELLA, G., AND R. L. BERGER. 1990. Statistical inference. Duxbury Press, Belmont, California.

CAVENDER, J. A. 1978. Taxonomy with confidence. Math. Biosci. 40:271–280.

CHURCHILL, G. A., A. VON HAESELER, AND W. C. NAVIDI. 1992. Sample size for a phylogenetic inference. Mol. Biol. Evol. 9:753–769.

EDWARDS, A. W. F. 1972. Likelihood. Oxford Univ. Press, Oxford.

FARRIS, J. S. 1973. A probability model for inferring evolutionary trees. Systematic Zoology 22:250–256.

FARRIS, J. S. 1999. Likelihood and inconsistency. Cladistics 15:199–204.

FELSENSTEIN, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Genet. 25:471–492.

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

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

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

FELSENSTEIN, J. 1992. Phylogenies from restriction sites: A maximum-likelihood approach. Evolution 46:159–173.

FELSENSTEIN, J., AND G. A. CHURCHILL . 1995. A hidden Markov chain approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13:93–104.

FRUMHOFF, P. C., AND H. K. REEVE. 1994. Using phylogenies to test hypotheses of adaptation: A critique of some current proposals. Evolution 48:172–180.

FUNK, V. A., AND W. L. WAGNER. 1995. Biogeography of seven ancient Hawaiian plant lineages. Pages 160–194 *in* Hawaiian biogeography: Evolution on a hot spot archipelago (W. L. Wagner and V. A. Funk, eds.). Smithsonian Institution Press, Washington, DC.

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.

HALDANE, J. B. S. 1919. The combination of linkage values and the calculation of distances between the loci of linked factors. J. Genet. 8:299–309.

HASEGAWA, M. H., H. KISHINO, AND T. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 21:160–174.

HUELSENBECK, J. P. 2000a. MrBayes: Bayesian inference of phylogeny, version 1.1. Distributed by the author, Department of Biology, Univ. of Rochester, New York.

HUELSENBECK, J. P. 2000b. Accommodating phylogenetic uncertainty in evolutionary studies. Science.

HUELSENBECK, J. P., AND J. J. BULL. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. Syst. Biology 45:92–98.

HUELSENBECK, J. P., AND B. RANNALA. 1997. Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. Science 276:227–232.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

LARGET, B., AND D. L. SIMON 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.

LEWIS, P. O. 1998. A genetic algorithm for maximum likelihood phylogeny inference using nucleotide sequence data. Mol. Biol. Evol. 15:277–283.

MADDISON, W. P. 1995. Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. Syst. Biol. 44:474–481.

MAU, B., M. A. NEWTON, and B. LARGET. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12.

MOOERS, A. Ø., AND D. SCHLUTER. 1999. Reconstructing ancestor states with maximum likelihood: Support for one- and two-rate models. Syst. Biol. 48:623–633.

NEYMAN, J. 1971. Molecular studies of evolution: A source of novel statistical problems. Pages 1–27 *in* Statistical decision theory and related topics (S. S. Gupta and J. Yackel, eds.). Academic Press, New York.

OLSEN, G. J., H. MATSUDA, R. HAGSTROM, AND R. OVERBEEK. 1994. fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. 10:41–48.

PAGEL, M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. Proc. R. Soc. London B 255:37–45.

PAGEL, M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Syst. Biol. 48:612–622.

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 phylogeny reconstruction (R. W. Scotland, D. J. Siebert, and D. M. Williams, eds.). Clarendon Press, Oxford.

QUICKE, D. L., AND R. BELSHAW. 1999. Incongruence between morphological data sets: An example from the evolution of endoparasitism among parasitic wasps (Hymenoptera: Braconidae). Syst. Biol. 48:436–454.

RANNALA, B., AND Z. YANG. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

REEVES, J. H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J. Mol. Evol. 35:17–31.

SALTER, L. A., AND D. K. PEARL. 2001. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. Syst. Biol. 50:7–17.

SANDERSON, M. J., AND L. HUFFORD (eds.). 1996. Homoplasy: The recurrence of similarity in evolution. Academic Press, New York.

SCHLUTER, D., T. PRICE, A. Ø. MOOERS, AND D. LUDWIG. 1997. Likelihood of ancestor states in adaptive radiation. Evolution 51:1699–1711.

SCHULTZ, T. R., R. B. COCROFT, AND G. A. CHURCHILL. 1996. The reconstruction of ancestral character states. Evolution 50:504–511.

SIDOW, A., AND T. P. SPEED. 1992. Estimating the fraction of invariable codons with a capture-recapture method. J. Mol. Evol. 35:253–260.

STEEL, M., D. HUSON, AND P. J. LOCKHART. 2000. Invariable sites models and their use in phylogeny reconstruction. Syst. Biol. 49:225–232.

STEEL, M., AND D. PENNY. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17:839–850.

SWOFFORD, D. L. 2001. PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods), version 4.0b6. Sinauer Associates, Sunderland, Massachusetts.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

TAYLOR, H. M., AND S. KARLIN. 1984. An introduction to stochastic modeling. Academic Press, New York.

TUFFLEY, C., AND M. STEEL. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59:581–607.

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. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314.

YANG, Z. H., AND B. RANNALA. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method: Mol. Biol. Evol. 14:717–724.

Received 21 March 2000; accepted 30 October 2000
Associate Editor: R. Olmstead