Syst. Biol. 48(3):612–622, 1999

The Maximum Likelihood Approach to Reconstructing Ancestral Character States of Discrete Characters on Phylogenies

MARK PAGEL

School of Animal and Microbial Sciences, University of Reading, Whiteknights, Reading RG6 6AJ, England; E-mail: m.pagel@reading.ac.uk

A phylogeny describes the hierarchical pattern of descent of some group of species from a common ancestor. If information is available on the character states of the contemporary species, the possibility is raised of using that information in combination with the phylogeny to reconstruct the historical events of evolution. These reconstructions can be used to retrieve a picture of the world as the species evolved along what would become the branches of the phylogeny. This, in turn, provides a way to test hypotheses about evolution and adaptation.

Methods based on the principle of parsimony reconstruct the ancestral character states to minimize the number of historical character changes required to produce the diversity observed among the contemporary species (see Maddison et al., 1984, for a general account). An alternative to parsimony approaches makes use of the principle of maximum likelihood. Maximum likelihood solutions make the observed data most likely given some model of the process under investigation (see Edwards, 1972). In a phylogenetic context this means reconstructing the ancestral character states to make the character states observed among the contemporary species most probable, given some statistical model of the way evolution proceeds. Maximum likelihood solutions may or may not be the most-parsimonious solution.

I restrict myself here to using maximum likelihood models to infer ancestral character states for binary discrete characters, that is, for characters that can adopt only two states, although the generalization to more than two states requires no new concepts. My approach to reconstructing ancestral states makes use of a Markov model of binary character evolution on phylogenies (Pagel, 1994). Sanderson (1993) describes a related model for investigating rates of gains and losses of characters for which the ancestral states are assumed to be known. Schluter (1995), Yang et al. (1995), and Koshi and Goldstein (1996) derive methods that are similar to the procedures I will describe here. However, Yang et al. (1995) and Koshi and Goldstein (1996) use what I shall term "global" methods for estimating ancestral characters, I argue for a "local" approach on grounds that the global method does not produce a maximum-likelihood estimate of the hypothesis of interest. Schluter (1995) reported global and local estimators in his investigation of artiodactyl ribonucleases, and Schluter et al. (1997) reported global estimators.

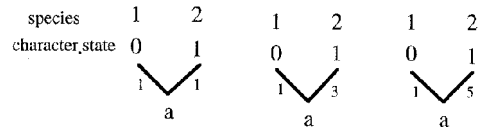
In several recent papers, Schluter (1995, Schluter et al., 1997) called attention to the usefulness of reconstructing ancestral character states for testing ideas about adaptation and evolution, and much of what I say here owes its inspiration to these investigations. Mooers and Schluter (1999) now provide important additional examples of how maximum likelihood methods can return both more information about ancestral character states than parsimony approaches, as well as information that is at odds with parsimony reconstructions.

I intend this article to act as a primer to those who are interested in using maximum likelihood methods but who may not be familiar with the mathematics of the approach. Accordingly, I begin with the simplest case of estimating the ancestral state of two species.

THE SIMPLEST CASE

Figure 1 shows two species, their character states, and the lengths of the branches of

the phylogenetic tree leading to them from their ancestor. Branch lengths can be measured in units of time or in units of the "opportunity for selection" or "operational time" (Pagel, 1994, 1997), including such things as genetic distance or possibly time scaled by generation time.



$\hat{L}(d)$	0.250	0.256	0.296
$\hat{L}(a = 0 \mid d)$	0.125	0.236	0.291
$\hat{L}(a = 1 \mid d)$	0.125	0.125	0.125

FIGURE 1. Phylogenetic tree of two species. Likelihood calculations are based on the character states as shown, given three different sets of branch lengths $L(m, d)$, $L(m, a = 0)$, and $L(m, a = 1)$ as defined in the text.

Denote the data observed at the tips of the tree in Figure 1 as $d = \{d1, d2\}$, where here $d = \{0, 1\}$, and denote the ancestral node as $n = \{a\}$. For larger trees, both d and n will contain more elements. We will presume that the tree and the lengths of its branches are known and fixed. The character states at the ancestral nodes will normally not be known and thus will constitute parameters to be estimated. Our interest will be to derive statements that can be used to suggest that one value at a node is more or less likely than some other.

Given a model of character evolution, it will be straightforward to estimate the probability of observing the data d on any tree. Let m denote a model of character evolution that describes a character that can adopt two states, and we shall suppose that transitions between character states can proceed in either direction. I have earlier (Pagel, 1994, 1997) described a continuous-time Markov model for this process that represents the probability of a character transition as a function of two transition-rate parameters and time. Let $P_{ij}(t) = m(\alpha, \beta, t)$ represent the probability of a transition from character state i to character state j along a branch of length t. The parameter α is the instantaneous rate of transition from state 0 to state 1, and β is the instantaneous transition from state 1 to state 0. These are sometimes referred to as "forward" and "backward" transitions, respectively. This formulation leads to four possible probabilities corresponding to the beginning and ending states of each branch of the phylogeny, as shown in Table 1.

TABLE 1. The four possible transitions of a binary character between the beginning and end of a branch.

	State at end of branch	
State at beginning of branch	0	1
0	$P_{00}(t) = 1 - P_{01}(t)$	$P_{01}(t)$
1	$P_{10}(t) = 1 - P_{11}(t)$	$P_{11}(t)$

Let $P(d|m)$ represent the probability of the observed data given the model of evolution. Calculating $P(d|m)$ requires estimating the two transition-rate parameters, as the branch lengths are assumed known. Let $L(m; d)$ represent the likelihood of the observed data. It is a property of likelihood (e.g., Edwards, 1972) that

$$L(m; d) \propto P(d|m) \qquad (1)$$

where the constant of proportionality is arbitrary. The difference between the probability and likelihood approaches is that whereas probability approaches describe the data for a given and fixed hypothesis, the likelihood approach seeks the hypothesis that best describes the data. In the present context, that means choosing the values of the transition-rate parameters that maximize Equation 1. The likelihood associated with this state will not ordinarily be interpretable as a probability: Probabilities sum to 1.0 when all possible outcomes of the data are entertained; likelihoods, on the other hand, consider the data as fixed and instead vary with the values of m. For this reason, it is convenient to write $L(m; d)$ in a shorthand as simply $L(m)$.

The likelihood of the data in Figure 1 is found from

$$\begin{aligned} L(m) &\propto P(d|m) \\ &= \sum_{a=0}^{1} w(a)P(d|m, a) \\ &= \sum_{a=0}^{1} w(a)(P_{a0}(t) \cdot P_{a1}(t)) \qquad (2) \\ &= w(0)(P_{00}(t) \cdot P_{01}(t)) + w(1)(P_{10}(t) \cdot P_{11}(t)) \end{aligned}$$

This equation acknowledges that the probability of observing the characters at the tips is the sum of two terms. One corresponds to the product of the two probabilities that are implied if the node a is 0 and the other to the product if the node a is 1. These two alternatives for node a are weighted by their prior probability of occurrence. In the absence of any other information, the alternative states are equally probable and $w(a) = 0.5$.

The maximum likelihood solution is to find those values of α and β that make Equation 2 yield the largest value. Previously (Pagel, 1994), I have described the Markov model in more detail, and a computer program (Discrete) to calculate the transition-rate parameters and likelihoods is available upon request. I have used Discrete to perform all of the calculations reported here.

By setting the two branch lengths equal to 1.0, Equation 2 yields a value of 0.25 for these data (Figure 1) with $\hat{\alpha} = \hat{\beta} = 8.0$. These values of the transition-rate parameters make all of the probabilities (Table 1) equal to 0.50 and thus the overall probability is just $(0.5 \cdot 0.5) + (0.5 \cdot 0.5)$, which is then multiplied by the prior probability of 0.5.

The Most Likely Ancestral State

$L(m)$ was found above by maximizing over both states at the "root." This removes the root (or more generally any ancestral node) from the likelihood and makes the likelihood independent of any particular value. To estimate the more likely of the two possible ancestral states, two separate likelihoods are found, conditional on the state of node a.

Write

$$L(m, a = i) \propto w(a = i)P(d|m, a = i) \qquad (3)$$

where $i = 0, 1$ corresponds to each of the two alternative states at the node a. The two likelihoods are estimated separately, having fixed the root at either $a = 0$ or $a = 1$ (Figure 1) and re-estimating the parameters of m. The two likelihoods will not necessarily sum to $L(m)$ of Equation 2. The likelihoods for the ancestral states are estimated this way because the assignment of a value to the root (more generally to any node) implies a different set of maximum likelihood transition-rate parameters (the α and β's of m) from summing over both values at the root.

When the branch lengths are equal, the model finds that the likelihood of the ancestor being 0 is equal to the alternative of it being 1 , and thus we cannot distinguish between these two hypotheses on the data. This result agrees with that obtained from parsimony. Both approaches also agree with common sense: Knowing nothing other than that one species has state 0 and the other state 1, we must be indifferent, lacking any other information, as to the state of the ancestor.

Allowing the branches to be of different lengths, both intuition and the likelihood result suggest that we might now adopt a preference for one value at the root over the other. Letting t_2, the branch leading to species 2, be equal to 3 versus a length of 1 for the other branch, $\hat{L}(m, a = 0)$ is roughly 1.7 times greater than $\hat{L}(m, a = 1)$ (Fig. 1); letting $t_2 = 5$, $\hat{L}(m, a = 0)$ now exceeds $\hat{L}(m, a = 1)$ by a factor of 2.0. The logic of this is that if there has been greater opportunity for change along a branch, then the endpoint of that branch is less likely to reflect the starting point.

The General Case of Estimating Ancestral States

Consider the tree in Figure 2, with two ancestral nodes, such that $n = \{a \text{ and } b\}$. We may wish to estimate the ancestral states of chosen nodes of this tree, or we may wish to estimate the most likely single set of nodes, that is, the single best description of the past.

Finding the Most Likely Set of Ancestral States

We may wish to find the particular assignment of states to n that maximizes the likelihood of the observed data. To do this, find

$$L(m, n = a, b) \propto w(b)P(d|m, n = a, b) \quad (4)$$

where here $w(b) = 0.5$, as we have no information about the root.

There are four possible assignments for a binary character at two nodes, and Equation 4 says that we separately calculate each of them. The largest is the maximum likelihood estimate of the ancestral nodes. The likelihoods, shown in Figure 2, are always large when a 1 is placed at node a. Reassuringly, the largest corresponds to placing a 1 at both the root (node b) and at node a. As above, the four likelihoods will not necessarily sum to $P(d|m)$, which is 0.148 for these data.

It may seem curious that in Equation 4 the right-hand side is multiplied only by the prior probability of node b, rather than the joint prior probability of $\{a, b\}$. However, having fixed a prior probability for node b automatically sets the prior weights for a. The latter are functions of the relative probabilities of the character transitions in the branch leading to the node a.

Finding the Most Likely Value at a Single Node

The most likely state at a single node is found by maximizing Equation 4 while holding constant only the state of the node of interest. This is repeated separately for all possible states at that node (here two states, 0 or 1) and the largest likelihood corresponds to the maximum likelihood estimate for that node. If our interest is to estimate the state of node a, we find

$$L(m, a = i) \propto \sum_{b=0(a=i)}^{1} w(b)P(d|m, a = i) \quad (5)$$

Holding node a constant, we maximize the likelihood summed over the two states at node b. In words, to find the likelihood that node $a = i$, find the probability of the data on the tree having fixed node a at state i, and then multiply by the prior weight $w(b)$. For the data and tree of Figure 2, the likelihood of character state 1 is $0.5 \cdot 0.211$; for state 0, it is $0.5 \cdot 0.125$.

The General Case

Equations 2, 4, and 5 can be generalized to any tree to find the likelihood estimate of the data, of the state at any given node, or of the single set of states that maximize the likelihood. If the tree contained three nodes, such that $n = \{a, b, c\}$, Equation 2 would be written as

$$L(m) \propto P(d|m) = \sum_{c=0}^{1}\sum_{b=0}^{1}\sum_{a=0}^{1} w(c)P(d|m, n) \quad (6)$$

and, as before, a single model would be fit to the data after having summed over all possible states at each node. To find the best set of states, Equation 4 would be written as

$$L(m, n = a, b, c) \propto w(c)P(d|m, n = a, b, c) \quad (7)$$

and eight separate conditional probabilities would be calculated, estimating a different set of transition-rate parameters for each. Similarly, to study the state of a particular node, for example, node a, Equation 5 would be written as

$$L(m, a = i) \propto \sum_{c=0}^{1}\sum_{b=0}^{1} w(c)P(d|m, a = i) \quad (8)$$

and two separate conditional probabilities would be found. The eight probabilities derived from Equation 7 and the two derived from Equation 8 will not in general sum to $L(m)$.

Local Versus Global Estimators

The estimators of the likelihoods (Eqs. 4, 5, 7, and 8) I shall term 'local' estimators because the parameters of the model of evolution are found separately for each combination of ancestral states. The local procedure is used because the maximum likelihood approach is to compare the support for different hypotheses when they are assumed to be true. The support for the hypothesis that the ancestral state at a node is zero derives from how well the model can be shown to fit the data when the ancestor in question is fixed at zero. Support for the hypothesis that the state is 1 is likewise found by maximizing the likelihood but this time having

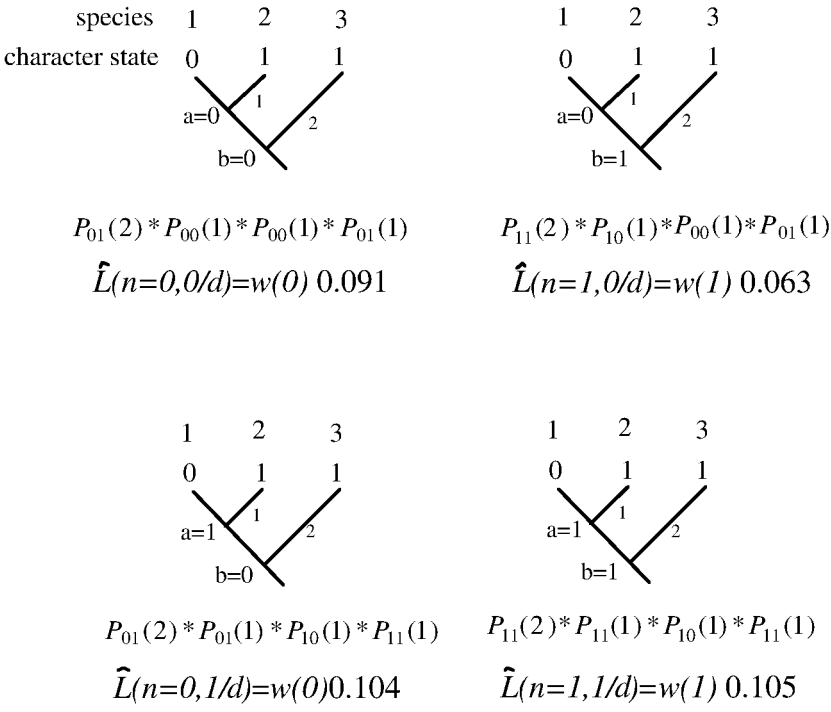


FIGURE 2. Four possible reconstructions of the ancestral character states on a three-species tree, and the likelihoods associated with each reconstruction (see text for definition of likelihood).

fixed the ancestral value at 1. The different possible states at a node become hypotheses for which we seek to estimate the support, given the data and our model of evolution.

The local method of estimation differs in calculation and logic from the estimators of ancestral states that Yang et al. (1995) and Koshi and Goldstein (1996) suggest for nucleotide and amino acid data. These authors first estimate $L(m) \propto P(d|m)$, as defined above. They then partition the overall likelihood into its additive components to find the proportion of the total likelihood attributable to the different character states at a given node—the parameters of the model of evolution having been estimated only once and then not for any particular state, but as the single best set maximized over all possible states. I call these estimates 'global' because unlike the local estimators, the global likelihoods are not estimated as being conditional on any particular state at a node or set of nodes, and consequently they do not reflect the best possible fit of the model, given the hypothesis.

Schluter (1995) independently reported both local and global estimators, as defined above, in his reconstructions of artiodactyl ribonucleases. Later, Schluter et al. (1997) reported only the global estimators in their investigations of behavioral and morphological characters.

Global estimators are maximum likelihood estimates of the likelihoods of ancestral character states, and they cannot be used to compare the support for alternative hypotheses. This is because the global likelihoods have not been maximized under the assumption that the hypothesis is true. This can only be done by fixing a node in a given state and then estimating the parameters of the model of evolution.

The key point is that the parameters of the model of evolution assumed to describe the data may vary, and sometimes greatly, depending on the assignment of a state to a

given node. We do not have a way of estimating the support for a state at a node independently of these background parameters. To calculate conditional probabilities from a single model fitted to all possible background states misses this point and as a result fails then to yield the maximum likelihood estimator.

I have calculated local and global likelihoods for successive subsets of the tree of Figure 3, each subset including an additional outgroup. For each subset I calculated the likelihood by using Equation 8 and the local procedure. I calculated global estimates from first finding the single set of transition-rate parameters (α, β) that maximized $P(d|m)$ over all possible ancestral character states at each node and then merely partitioned it into its two additive parts.

Table 2 shows the two likelihoods and the estimated values of α and β for the subset tree of the first three species of Figure 3. Not only do the transition-rate parameters differ substantially for different assignments to node *a*, but also they produce quite different likelihoods in comparison with the global estimators.

TABLE 2. Likelihoods derived from local and global procedures for the three-species phylogeny of Figure 2. α's and β's are the forward and backward transition rates, (see text). Subscript refers to character state at node *a*.

Probability	Local $\hat{\alpha}_0 = 1.52, \hat{\beta}_0 = 1.39, \hat{\alpha}_1 = 24, \hat{\beta}_1 = 8$	Global $\hat{\alpha} = 11, \hat{\beta} = 5.5$
$\hat{L}(m, a = 0)$	0.063	0.050
$\hat{L}(m, a = 1)$	0.106	0.099
$\hat{L}(m)$		0.148
$\hat{L}(m, a = 1)/\hat{L}(m, a = 0)$	1.69	1.99

The ratio of the likelihoods record the relative amount of support for different hypotheses on the data. Figure 4 plots the natural logarithm of the ratio of the two local likelihoods and the two global likelihoods against the number of outgroups used in the calculation. The value the ratio takes is a measure of the relative preference for character state 1 at node *a* over character state 0. A value of 1.0 means that the ratio of the conditional probabilities is ~ 2.7 and a value of 2.0 corresponds to a ratio of ~ 7.4.

As outgroups are added, preference for a 1 at node *a* grows rapidly at first and then appears to plateau for these data. The qualitative pattern of differences between the local conditional probabilities being smaller than the global probabilities is expected. The global solution favors a 1 at node *a*, and consequently the transition rate parameters (the α's and β's) are biased towards a 1 at node *a*. The "global" model accordingly does not fit the data very well when node *a* takes the value 0. In contrast, the local procedure shows how well the model can fit the data when it is optimized to the state at node *a*.

Global weights are appropriate for finding the maximum likelihood description of the species data, $L(m)$, when no hypothesis is made about the value of an internal node of the tree. The relative support for alternative hypotheses of evolution can then be compared to test which best describes the species data. This is, for example, the approach that I (Pagel, 1994) used to test hypotheses about the independent versus correlated evolution of two characters on a phylogeny.

Interpretation and Hypothesis Testing

To compare two likelihoods, define LR = $-2 \log_e[H_{a=i}/H_{a=j}]$ where $H_{a=i}$ is the smaller of the two likelihoods, and *a* refers to any node. When the two hypotheses are "nested," the LR statistic is asymptotically distributed as a χ^2 variate with degrees of freedom equal to the number of parameters that differ between the models that define the two hypotheses. Two hypotheses are considered nested if one can be defined as a special case of the other.

Here we wish to compare two hypotheses on the data that are not nested, as they differ by having assigned a 0 or a 1 to node *a*. When the hypotheses are not nested, the logarithm of the ratio of the likelihoods does not necessarily follow a χ^2 distribution and

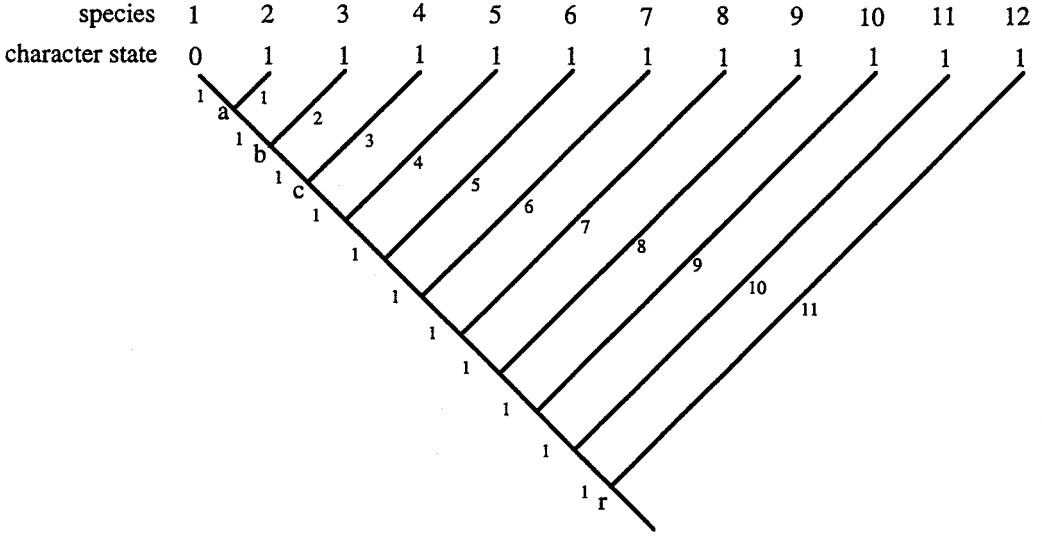


FIGURE 3. Phylogeny of 12 species (10 outgroups to node *a*), with branch lengths shown. Nodes are identified by lower-case letters. All species are equidistant from the root.

therefore no *p*-value can be assigned (see Goldman [1993] and Pagel [1994] for discussions of how to obtain *p*-values by simulation for other sorts of nonnested hypotheses). In such cases the LR statistic is usually interpreted as a measure of "support" (Edwards, 1972), and following Edwards (1972), a support of ~ 2 log units is frequently taken as rule-of-thumb evidence that the likelihoods are significantly different.

This is a stringent criterion as, on this view, not even with 10 outgroups all taking state 1 is there "significantly" more support for $L(m, a = 1)$. It is tempting to compare this outcome unfavorably with parsimony, which would unambiguously favor a 1 at node *a*. Because forward and backward transitions are nearly equally likely in this example (estimated transition-rate parameters are such as to make all probabilities of change = 0.25), support for the two states at node *a* does not differ greatly. Other trees could give more or less support. But it should be borne in mind that the likelihood approach also favors a 1 at node *a* and does so beginning with the first outgroup. What the likelihood procedure reminds us is that, if we are willing to take our model of evolution seriously, a zero at node *a* is not out of the question, even if less likely.

Inferring the State at the Root of the Tree: An Iterative Bayesian-like Approach

The likelihood procedure of Equation 8 can also be applied to estimate the root of the tree. By definition, at the root there are no further outgroups, and the prior probabilities of the two values are 0.50 in the absence of any other information. The likelihoods as defined above can, however, be used to gather information to shift values away from these priors.

Let

$$P_q(r = i) = \frac{L(m, r = i)}{\sum_{r=0}^{1} L(m, r = i)} \qquad (9)$$

be the proportion of the two likelihoods attributable to the root node, *r*, taking state *i*. This is not a true conditional probability (hence the subscript *q* to denote quasi-probability) because the summation in the denominator is over two probabilities calculated from the "local" procedure and hence

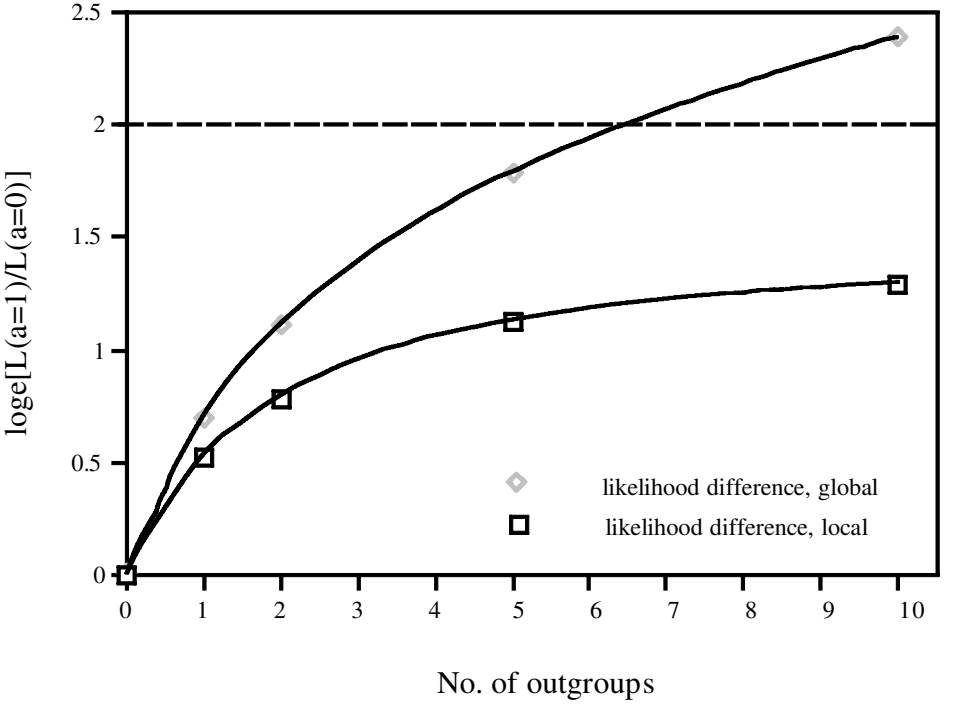


FIGURE 4. The "support" for placing a 1 at node *a* for subsets of the 12-species tree of Figure 3. *x*-axis, the number of outgroups to node *a* that were included in the calculation of support. Thus, two outgroups refer to a subset tree of Figure 3 based on the first four species, and 10 outgroups includes the entire tree. Support is defined as $\log_e[L(m, a = 1)/L(m, a = 0)]$. Large values indicate greater confidence in 1 relative to 0. The upper curve is the support derived from using global transition-rate parameters, and the lower curve is that from using local transition-rate parameters (see text for explanation). Dotted line at $y = 2.0$ is the conventional line of significance for this measure of support.

based on different transition-rate parameters. Equation 9 may, however, serve as a rough posterior weight to give a rule-of-thumb sense of how much the data can be used to shift away from the prior weights.

The availability of quasi-posterior weights suggests the possibility of using them in a new likelihood calculation as prior weights. That is, calculate $L(m)$ according to Equation 2 and using equal prior weights. From this, calculate the quasi-posterior weights, then use them as the prior weights in a new calculation of $L(m)$. Repeat this until no further improvement in the overall likelihood is obtained.

This may seem like cheating or even folly, but if real uncertainty exists as to the value of the root, the procedure is unlikely to "run away" and assign either a 0 or a 1 posterior weight to the alternative states at the root. It may also be preferable to direct maximum likelihood estimation of the relative weights to apply to the root, the quasi-posterior solution being constrained not to jump all the way towards a 0 or a 1 posterior weight in any single calculation. Nevertheless, it is a nonstandard Bayesian-like procedure, and much simulation work is required to assess its behavior. To the extent that the procedure generates weights that do not go to the extremes, it may suggest that the evidence for unambiguously assigning the root to one of the other states, as do parsimony solutions, may not always be clear-cut.

I have implemented the procedure in the program Discrete and applied it to the data of Figure 3, based on all 12 species. The two likelihoods for the root are equal and

TABLE 3. Applying the Bayesian-like procedure to the trees of Figures. 1 and 3.

	Unweighted solution		Weighted solution			
Tree	$\hat{L}(m, a = 0)$	$\hat{L}(m, a = 1)$	$\hat{w}(r = 0)$	$\hat{w}(r = 1)$	$\hat{L}(m, a = 0)$	$\hat{L}(m, a = 1)$
12 species, Figure 3	0.016	0.016	0.50	0.50	0.016	0.016
12 species, short branch to root	0.015	0.016	0.00	1.0	0.0	0.018
2 species, $t_2 = 3$	0.24	0.13	1.0	0.0	0.24	0.0
2 species, $t_2 = 1.15$	0.14	0.13	0.83	0.17	0.13	0.12

this makes the quasi-posterior weights equal (Eq. 9) and equal to the priors (Table 3). As a consequence, the weighted posterior probabilities do not change. This reflects the fact that there is no information to bias the result towards one or the other value. Shortening to 0.05 the branch that leads from species 12 to the root leads to a slight increase in confidence that the root is 1 under the unweighted model and leads to a large increase in the posterior weights model. Applying the procedure to the simple tree of Figure 1 with $t_2 = 3$, the posterior weights bias the likelihoods to a clear preference for a 0 at node *a*. Setting $t_2 = 1.15$ for the simple two-species phylogeny returns an intermediate result.

Some Peculiar Features of Maximum Likelihood as Applied to Discrete Characters

Maximum likelihood solutions to character evolution can sometimes return puzzling results when applied to phylogenetic data. I will briefly discuss two issues in this section that can help to explain likelihood solutions for phylogenetic data: ambiguous or counterintuitive values at the root of the tree, and testing whether forward and backward transitions are equally likely.

The likelihoods for the root of the full ($n = 12$ species) phylogeny of Figure 3 suggest that 0 and 1 are equally likely (Table 4). Given that all but one species take the value of 1, this seems counterintuitive. This result arises because the maximum likelihood transition-rate parameters are estimated to be large, and thus all kinds of change are equally likely. The model allows change to happen anywhere on the tree and more than once in a branch. Placing a 0 at the root of Figure 2 implies that somewhere along the branch leading to the outgroup there must be at least one more transition to 1 than transitions back to 0 (hence, α is large). A similar scenario must hold true to yield 1's for each of the tips except for species 1. However, to yield a 0 in species 1, β must also be reasonably large relative to α. This, however, places further pressure on α to be large—to ensure that any transitions to 0 in other branches on the tree quickly return to 1. Placing a 1 at the

TABLE 4. Selected conditional probabilities. The transition-rate parameters are derived from a global solution but do not qualitatively differ from the separate local rate parameters.

Tree	$\hat{L}(m)$	$\hat{L}(m, a = 0)$	$\hat{L}(m, a = 1)$	$\hat{\alpha}$	$\hat{\beta}$
$n = 12$, Figure 2	0.032	0.016	0.016	13	1
$n = 12$, all species = 1's					
$\alpha \neq \beta$	0.98	0.49	0.49	49	5×10^{-5}
$\alpha = \beta$	0.49	8.9×10^{-4}	0.49	1.7×10^{-4}	1.7×10^{-4}

root implies no net changes to 0, except in the short branch leading to species 1. To ensure that this transition happens, the model makes β moderately large as before, but this must now be balanced in the rest of the tree by a large α to ensure that any transitions to 0 quickly go back to 1.

Thus, both values at the root potentially imply frequent character transitions throughout the tree. The parsimony solution in this instance implicitly rules out changes from 1 to 0 everywhere on the tree except in the branch leading to species 1, and so does not return an ambiguous assignment to the root.

A more puzzling result arises when all species of Figure 2 are assigned a 1. The maximum likelihood result for this situation also suggests that 0 and 1 are equally likely (Table 4). This somewhat disturbing result arises because the model is forced to consider both possibilities at the root. It turns out that the two possibilities make the same demands on the transition-rate parameters: Placing a 0 at the root requires a large α to ensure that all tips eventually have a 1. Placing a 1 at the root also requires a large α, this time to ensure that any stray $1 \rightarrow 0$ transitions must immediately change back to 1. The prediction is that α will be large and β small for both values at the root, and this is what the model returns. Surprisingly, the root is ambiguous even when there is no variance at the tips.

Here, however, the problem can be dealt with by asking whether a two-parameter model allowing unequal forward and backward transition rates is justified for these data. Restricting $\alpha = \beta$ and fitting what is now a one-parameter model to the data will yield a likelihood that can be compared with the two-parameter likelihood directly via a likelihood ratio test: $\mathrm{LR} = -2\log_e[L(m\{\alpha = \beta\})/L(m\{\alpha, \beta\})]$, and the test will have 1 df. For the data of Table 4, the LR statistic is $\mathrm{LR} = -2\log_e[0.49/0.98] = 1.38$, which is not significant and suggests that the two-parameter model does not lead to a significant improvement in the fit of the model to the data.

Adopting the simpler one-parameter $m\{\alpha = \beta\}$ model, the results now overwhelmingly prefer a 1 at the root. The conclusion that arises from this exercise is that we should always ask whether the data justify fitting a two-parameter model. Maximum likelihood will optimize the model to the data, independently of whether there is sufficient information in the data to justify unequal rates of forward and backward transitions. If a one-parameter model (e.g., setting $\alpha = \beta$) does not lead to a significant reduction in the likelihood, then this model should be used in preference to the two-parameter model. Mooers and Schluter (1999) discuss this idea in more detail.

These kinds of puzzling results can arise even when the data are not as extreme as those used here. Usually a puzzling result can be understood either by applying the logic of counting implicit transitions or by testing for the fit of one- versus two-parameter models to the data. In some cases, there may be evidence favoring a specific value at the root over another, and fixing the root to that value can cause the estimates of the forward and backward transition rates to settle down.

The continuous-time Markov model takes into account the lengths of the branches of the phylogeny in a way that accords with intuition. This also means that investigators must be aware of the assumptions implicitly built into any set of branch lengths they may use. If it is believed that the probability of a character changing state is a function of time or genetic distance, then branch lengths reflecting these sorts of distances should be employed. If, on the other hand, character transitions are thought likely to occur independently of branch length, then setting all branches to the same length may be appropriate. Pagel (1994) described a branch-length–scaling feature that can find by maximum likelihood the optimal scaling of a set of branches. This included a test for assuming all branches to be of the same length, an implicitly punctuational view of evolution.

Future Prospects

When methods of character reconstruction are applied to DNA or RNA, or to amino acid sequences (e.g., Schluter, 1995;

Golding and Dean, 1998), investigators can use the inferred ancestral states to reconstruct and examine both in vitro and in vivo the putative ancestral protein. Technological advances in developmental biology further raise the fascinating spectre of some day inferring the genotype of, and then producing, entire ancestral organisms. What Schluter (1995), Schluter et al. (1997), and Mooers and Schluter (1999) have gone some way in showing is that real uncertainty exists over what state the past took, and this uncertainty must be reflected in the hypothesis tests and the conclusions drawn from them.

Einstein once remarked that "all of our science measured against reality is primitive and childlike." The two-state Markov transition model is likely to be a useful starting point for modeling character transitions on an evolutionary scale, but several issues for further research already present themselves. The most pressing issue is, do we really believe, as the model expects us to, that the transition-rate parameters apply equally everywhere on the tree? Mooers and Schluter (1999) speculate that we might wish to treat outgroups as having different transition rates, by virtue of being outgroups. If rates are not uniform, how might we go about identifying regions of the tree among which the transition rates differ? Parsimony reconstructions can be seen as, in effect, presuming that transition rates are zero in some parts of the tree and greater than zero elsewhere. Is this realistic?

What is urgently needed are data sets for which the ancestral states are known, such as the bacteriophage data of Hillis et al. (1992). I have analyzed (Pagel, unpublished) some of the over 200 restriction enzyme sites that Hillis et al. report, using the two-state Markov model described here, and typically get accurate reconstructions. For these molecular data, it would appear that the Markov assumption of transition rates applying equally everywhere on the tree is reasonable.

Another area requiring development is in the generalization of the two-state Markov model to three- and four-state characters. Four-state models for gene sequence data are available but will not in general be appropriate for other kinds of characters. Incorporating information from second and third characters may also prove helpful in reconstructing ancestral states. If two characters are known to be correlated, and the state of one is known with some certainty, that knowledge may be useful in developing hunches about the state of the other character. This is an area for much future investigation.

Acknowledgments

This work was supported by the Natural Environment Research Council and the Biotechnology and Biological Sciences Research Council of the United Kingdom.

References

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

Golding, G. B., and A. M. Dean. 1998. The structural basis of molecular adaptation. Mol. Biol. Evol., 15:355–369.

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

Hillis, D., J. J. Bull, M. E. White, M. R. Badgett, and I. Molineux. 1992. Experimental phylogenetics: Generation of a known phylogeny. Science 255:589–592.

Koshi, J. M., and R. A. Goldstein. 1996. Probabilistic reconstruction of ancestral protein sequences. J. Mol. Evol., 42:313–320.

Maddison, W. P., M. J. Donoghue, and D. R. Maddison. 1984. Outgroup analysis and parsimony. Syst. Zool. 33:83–103.

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.

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

Pagel, M. 1997. Inferring evolutionary processes from phylogenies. Zool. Scr. (25th Anniversary Special Issue on Phylogenetics and Systematics) 26:331–348.

Sanderson, M. J. 1993. Reversibility in evolution: A maximum likelihood approach to character gain–loss bias in phylogenies. Evolution 47:236–252.

Schluter, D. 1995. Uncertainty in ancient phylogenies. Nature 377:108–109.

Schluter, D., T. Price, A. Ø. Mooers, and D. Ludwig. 1997. Likelihood of ancestor states in adaptive radiation. Evolution 51:1699–1711.

Yang, Z., S. Kumar, and M. Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–1650.

Received 1 September 1998; accepted 15 February 1999
Associate Editor: C. Cunningham