Syst. Biol. 53(5):673–684, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490522232

Bayesian Estimation of Ancestral Character States on Phylogenies

MARK PAGEL, ANDREW MEADE, AND DANIEL BARKER

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

Abstract.—Biologists frequently attempt to infer the character states at ancestral nodes of a phylogeny from the distribution of traits observed in contemporary organisms. Because phylogenies are normally inferences from data, it is desirable to account for the uncertainty in estimates of the tree and its branch lengths when making inferences about ancestral states or other comparative parameters. Here we present a general Bayesian approach for testing comparative hypotheses across statistically justified samples of phylogenies, focusing on the specific issue of reconstructing ancestral states. The method uses Markov chain Monte Carlo techniques for sampling phylogenetic trees and for investigating the parameters of a statistical model of trait evolution. We describe how to combine information about the uncertainty of the phylogeny with uncertainty in the estimate of the ancestral state. Our approach does not constrain the sample of trees only to those that contain the ancestral node or nodes of interest, and we show how to reconstruct ancestral states of uncertain nodes using a most-recent-common-ancestor approach. We illustrate the methods with data on ribonuclease evolution in the Artiodactyla. Software implementing the methods (*BayesMultiState*) is available from the authors. [Ancestral states; comparative methods; maximum likelihood; MCMC; phylogeny.]

Given a collection of species, information on their attributes, and a phylogeny that describes their shared hierarchy of descent, the prospect is raised of inferring the characteristics of the ancestors to these species. This is an intriguing idea, holding out as it does the possibility of glimpsing the past, of discovering how traits evolve and understanding their function. Reconstructions of the probable ancestral states of organisms have been used to infer proteins and genes that existed millions of years ago (e.g., Jermann et al., 1995; Schluter, 1995) to investigate ancient features of life on Earth (Galtier et al., 1999), and to test ecological and evolutionary hypotheses. Reconstructed ancestral states complement traditional palaeontological approaches to studying the past, and are able to investigate traits such as behaviors or physiology that do not fossilize (Pagel, 1999a).

To infer ancestral characters requires some model of how the traits are presumed to evolve, and a representation of the phylogenetic relationships among species. The model of evolution is used to characterize the evolutionary processes responsible for trait evolution, and the phylogeny specifies the probable routes by which ancestral species gave rise to the contemporary species. Statistical models of trait evolution (e.g., Pagel, 1994, 1999a; Schluter, 1995; Schluter et al., 1997; Mooers and Schluter, 1999; Cunningham, 1999) are now widely used. They have a number of desirable features, including the ability to estimate evolutionary processes such as rates of evolution, correlations among traits, and ancestral states. They also highlight the statistical uncertainty of their estimates.

Investigators often conduct comparative analyses on a single phylogenetic tree. This practice relies upon that tree being a valid representation of the hierarchical relationships among species as well as their relative degrees of divergence. However, phylogenies are seldom known with certainty, normally being inferences from data. This poses a problem because different trees can and frequently do give different answers to comparative questions. These include questions about ancestral character states, but also about other evolutionary questions such as the tempo and mode of evolution, timings, and correlations among traits.

It is tempting to think that the 'best' tree, defined as the tree that is optimal under some tree building criterion, such as maximum likelihood, maximum parsimony, or minimum evolution, offers a haven from phylogenetic uncertainty. But the 'best' tree need not in any given case be the true tree—it may be, and it may even have a reasonably high probability of being the true tree, but there is no way in general to know this in advance. Consensus trees highlight areas of the phylogeny that receive stronger or weaker support, but they do not by themselves provide a way to account for the effects of uncertainty on estimates of other parameters. Moreover, one knows a priori that consensus trees are not the 'best' trees.

If consensus trees are not 'best' trees and best trees cannot be presumed to be true trees, investigators must turn to other solutions. One possibility is to enumerate all possible trees and evaluate the comparative hypothesis in each. However, the number of possible topologies increases factorially with increasing numbers of species, and even for small trees this approach is impractical if one allows that branch lengths can vary continuously.

Haphazard samples, all or a sample of the equally parsimonious trees (Hibbett et al., 2000; Holden, 2002), a bootstrap sample of phylogenies, or a sample of trees thought somehow to reflect the range of uncertainty in the estimate of the true tree, all avoid the problems with enumeration. These sampling approaches are a step in the right direction but produce distributions of trees whose statistical properties are uncertain (Felsenstein and Kishino, 1993; Newton, 1996; but see Efron et al., 1996).

By comparison, Bayesian Markov chain Monte Carlo (MCMC) methods (e.g., Gilks et al., 1996) offer a formal statistical procedure for sampling from the probability distribution of phylogenetic trees, and have been for several years now a topic of interest in biology (Rannala and Yang, 1996; Wilson and Balding, 1998; Larget and Simon,

1999, Mau et al., 1999; Huelsenbeck et al., 2001, Lutzoni et al., 2001; Pagel and Lutzoni, 2002). Given a probability sample of trees, phylogenetic uncertainty is taken into account by estimating the parameter of interest in each tree, and combining the estimates across trees.

Our aim in this paper is to describe a general procedure for estimating comparative parameters across statistically justified samples of phylogenetic trees, with an emphasis on reconstructing ancestral character states. In particular, we shall describe how to use Bayesian methods to estimate the parameters of a model of discrete trait evolution and to infer the ancestral states of specified internal nodes. The methods we describe build on our earlier work using Bayesian inference to account for phylogenetic uncertainty when estimating trait evolution on phylogenies (Lutzoni et al., 2001; Pagel and Lutzoni, 2002).

An issue that arises when estimating ancestral states in a sample of trees is that not all of the trees necessarily contain the internal node or nodes of interest. One solution is to restrict analyses only to those trees that have the nodes (Huelsenbeck and Bollback, 2001; Lutzoni et al., 2001). This has the limitation that different trees might be used to estimate different nodes, and we show below that constraining the sample overestimates the confidence that can be placed in a reconstruction. The method we develop here avoids these difficulties by combining information on the uncertainty about the existence of a node with information on the uncertainty in the estimate of the ancestral state. It turns out that the uncertainty about a node's existence puts an upper limit on the confidence that can be placed in its ancestral state. We therefore show how, using a 'most recent common ancestor' approach, this limit can be overcome. We illustrate the methods with data on the evolution of the ribonuclease gene in the Artiodactyla.

A Primer of Bayesian Phylogenetic Inference

Accounting for phylogenetic uncertainty requires that the comparative hypothesis is evaluated in a statistically justified sample of trees, and so we begin with a brief discussion of how to characterize the probability distribution of phylogenies, and how to sample from it.

Bayesian methods offer a formal statistical procedure for calculating the posterior probability distribution of phylogenetic trees. Given an aligned set of sequence data, **S**, Bayes' rule as applied to phylogenetic inference states that the posterior probability of tree T_i is

$$p(T_i \mid \mathbf{S}) = \frac{p(\mathbf{S} \mid T_i)p(T_i)}{\sum_T p(\mathbf{S} \mid T_i)p(T_i)} \quad (1)$$

where $p(T_i \mid \mathbf{S})$ is the probability of tree T_i given the sequence data **S**, $p(\mathbf{S} \mid T_i)$ is the probability or likelihood of the data given tree T_i, and $p(T_i)$ is the prior probability of T_i. The denominator sums the probabilities over all possible trees.

Equation 1 usefully defines the posterior probability of any given tree but can be difficult to put into practice. The number of possible different unrooted topologies for n species is $(2n-5)!/(2^{n-3}(n-3)!)$ meaning that the summation in the denominator of Equation 1 is over a large number of topologies for all but the smallest datasets. In turn, for each of these possible topologies the quantity $p(\mathbf{S} \mid T_i)$ must be integrated over all possible values of the lengths of the branches of the tree and over the parameters of the model of evolution that describe the sequence data. Letting $\mathbf{t}$ be a vector of the branch lengths of the tree and $\mathbf{m}$ a vector of the parameters of the model of sequence evolution,

$$p(\mathbf{S} \mid T_i) = \int_{\mathbf{t}} \int_{\mathbf{m}} p(\mathbf{S} \mid T_i, \mathbf{t}, \mathbf{m})\, p(\mathbf{t})\, p(\mathbf{m}) d\mathbf{t} d\mathbf{m} \quad (2)$$

where $p(\mathbf{t})$ and $p(\mathbf{m})$ are the prior probabilities of the branch lengths and the parameters of the model.

Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods (e.g., Gilks et al., 1996) as applied to phylogenetic inference provide a computationally efficient way to estimate the posterior probability distribution of trees. A Markov chain is constructed, the states of which are different phylogenetic trees (Rannala and Yang, 1996; Wilson and Balding, 1998; Larget and Simon, 1999; Mau et al., 1999; Lutzoni et al., 2001; Huelsenbeck et al., 2001; Pagel and Lutzoni, 2002; Pagel and Meade, 2004). At each step in the chain a new tree is proposed by altering the topology, or by changing branch lengths or the parameters of the model of sequence evolution. The Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) is then used to accept or reject the new tree. A newly proposed tree that improves upon the previous tree in the chain is always accepted (sampled), otherwise it is accepted with probability proportional to the ratio of its likelihood to that of the previous tree in the chain.

If such a Markov chain is allowed to run long enough, it reaches a stationary distribution. At stationarity, the Metropolis-Hastings sampling algorithm ensures that the Markov chain 'wanders' through the universe of trees, sampling better and worse trees, rather than inexorably moving towards 'better' trees as an optimizing approach would do. A properly constructed chain samples trees from the posterior density of trees in proportion to their frequency of occurrence in the actual density. That is, the Markov chain draws a sample of trees that can be used to approximate the posterior distribution. In fact, the stationary distribution simultaneously samples the posterior density of trees, the posterior distributions of the branch lengths and parameters of the model of sequence evolution. By allowing the chain to run for a very long time—perhaps hundreds of thousands or millions of trees—the continuously varying posterior distribution defined in Equations 1 and 2 can be approximated to whatever degree of precision is desired.

The MCMC approach encourages investigators to move away from attempting to find best trees and towards estimating the statistical confidence they have in

particular aspects of the phylogeny. For example, the proportion of trees in the MCMC sample in which a given monophyletic group appears estimates the Bayesian posterior probability that that node defining that group actually exists, given the data and the model of evolution. This property is especially relevant to reconstructing ancestral states, as the posterior probability of a node directly measures the confidence that can be placed in the existence of that node. There is little to be gained in reconstructing ancestral states of nodes that have little probability of being true in the first place.

Modeling Trait Evolution

The model of trait evolution is used to characterize how a trait changes along the branches of a phylogeny. We restrict our discussion here to models for discrete traits, but all of the logic of our approach also applies to modeling continuously varying traits.

We use a continuous-time Markov model (Pagel, 1994, 1997; Lewis, 2001), mathematically equivalent to models of gene-sequence evolution, but not limited to investigating traits with four states. We shall describe the model for binary traits, or traits that adopt just two states, although our general approach is applicable to any number of states. Common binary traits might be the presence or absence of a morphological feature, two different forms of an animal mating system, or as here, the presence or absence of a particular amino acid at a specified site in a protein.

The continuous-time Markov model presumes that the trait can evolve repeatedly between its two possible states in any branch of the phylogenetic tree. Letting the two states be called "0" and "1," the rate parameter q_{01} measures the rate at which the trait changes from state 0 to state 1 over short interval dt, and q_{10} measures the rate at which the trait changes back again. Formally, the product of the rate parameters and the interval dt define the probability of a change over interval dt:

$$P_{01}(dt) = q_{01}dt$$
$$P_{10}(dt) = q_{10}dt$$

Over longer interval t, such as the branch of a phylogeny, the model allows that more than one change can occur. For example, the trait could change from state 0 to state 1 and then back again, and thus the expressions for $P_{01}(dt)$ and P_{10} (dt) above do not apply for a longer interval t. To derive the probabilities of change for longer intervals, the model is written in the form of a matrix $\mathbf{Q}$:

$$Q = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} -q_{01} & q_{01} \\ q_{10} & -q_{10} \end{bmatrix} \end{array}$$

This matrix shows the two rate parameters corresponding to the trait changing state and the two implied rate parameters corresponding to no change in the state of the trait. The latter are given by minus the sum of the other rate coefficients in the row of the matrix, such that each row sums to zero. Given $\mathbf{Q}$, the probabilities over longer intervals t are found by exponentiating the $\mathbf{Q}$ matrix multiplied by the length of the interval:

$$\mathbf{P}(t) = e^{-\mathbf{Q}t} = \begin{bmatrix} P_{00}(t) & P_{01}(t) \\ P_{10}(t) & P_{11}(t) \end{bmatrix}$$

$\mathbf{P}(t)$ is a square matrix of order equal to the number of states. The quantities $P_{ij}(t)$ and $P_{ji}(t)$ record the probability of beginning a branch of length t in one state and ending in another. This is not to say that there has been just a single change in the trait although that may be the most probable scenario; so long as the branch begins and ends in a different state, any number of changes is allowed to intervene. The quantities $P_{ii}(t) = 1 - P_{ij}(t)$ and $P_{jj}(t)$ record the probability of beginning and ending and branch of length t in the same state. As with the other probabilities, this is not to say that the character has remained unchanged, although again that is often the most probable interpretation, just that it began and ended the interval in the same state.

Branch lengths can be in units of time or, as is true of most molecular phylogenies, in units of genetic divergence. Branch lengths in units of genetic divergence may usefully record the underlying opportunity for trait evolution (Pagel, 1994). Because the continuous time Markov model allows multiple (but unseen) changes per branch, the rate parameters can adopt large ($\gg$1) values, especially if branch lengths are short and many changes are implied.

Estimating the Rate Parameters

The rate parameters are of interest in themselves and because they are sufficient to calculate the character states at ancestral nodes of a phylogeny (Pagel, 1999b). Their numerical values will not normally be known but can be estimated given a phylogeny and observations on the value of the trait in each species. The maximum likelihood approach finds the values of the rate parameters that make the observed data most probable (Pagel, 1994). Alternatively, we can use Bayes' theorem to estimate not just the maximum likelihood value of the rates but their entire posterior probability distribution.

Presume we have a single phylogenetic tree, let $\mathbf{Q}_i$ denote a particular set of rate coefficients, and let $\mathbf{D}$ stand for the dataset of traits observed across the species in the tree. The joint posterior probability of $\mathbf{Q}_i$ given the data $\mathbf{D}$ is

$$p(\mathbf{Q}_i \mid D, T) = \frac{p(D \mid \mathbf{Q}_i)\, p(\mathbf{Q}_i)}{\int_{\mathbf{Q}} p(D \mid \mathbf{Q})\, p(\mathbf{Q}) d\mathbf{Q}} \qquad (3)$$

where $p(D \mid \mathbf{Q}_i)$ is the probability of the data given the rate parameters in $\mathbf{Q}_i$, $p(\mathbf{Q}_i)$ is the prior probability of $\mathbf{Q}$, and the denominator integrates this same probability over all values of $\mathbf{Q}$.

Equation 3 says that the probability of any particular set of rate coefficients is just their proportion of the total probability, as summed over all possible sets of rate coefficients. Because the rate parameters of **Q** vary continuously, it will normally be difficult to calculate the integral. However, one way to approximate the integral is to sample many times from a Markov chain that implements the model of trait evolution. Successive steps in the trait-model Markov chain propose new values of the rate parameters. The stationary distribution of the chain samples **Q**'s posterior density.

Accounting for Phylogenetic Uncertainty

Rate parameters.—Equation 3 describes how to find the posterior distribution of the rate parameters on a single tree. To account for the influence of phylogenetic uncertainty on **Q**'s posterior distribution, it is necessary to integrate Equation 3 over all possible trees as well as over rate coefficients. Write

$$p(\mathbf{Q}_i \mid D) = \frac{\int_T p(D \mid \mathbf{Q}_i, T)\, p(\mathbf{Q}_i)\, p(T) dT}{\int_T \int_{\mathbf{Q}} p(D \mid \mathbf{Q})\, p(\mathbf{Q})\, p(T) d\mathbf{Q} dT} \quad (4)$$

to denote the posterior probability of $\mathbf{Q}_i$ integrated over trees. The integral for trees is over all possible branch lengths and parameters of the model of sequence evolution (Equation 2). Equation 4 says that the probability of a given set of rate parameters in **Q** is their proportion of the total likelihood, now calculated over all possible trees and sets of rates. By sampling repeatedly from a Markov chain that moves among phylogenetic trees and values of **Q**, an estimate of the joint distribution of the rate parameters across trees can be obtained.

Equations 3 and 4 show that **Q**'s posterior density has components attributable to *within-tree* and *between-tree* influences: the integral over **Q** on a single tree accounts for the within-tree uncertainty and the integral over trees for a given **Q** accounts for the between tree uncertainty. If **Q**'s posterior density is identical in every tree, then there is no between-tree variance and Equations 3 and 4 will produce the same posterior distribution. If, on the other hand, **Q**'s posterior density differs from tree to tree, then the posteriors derived from Equation 4 will have a larger variance than those from Equation 3. The joint density distribution of Equation 4 therefore accounts for phylogenetic uncertainty in the sense that **Q**'s posterior density contains all of the variability that is attributable to trees.

The ancestral states.—Any given set of rate coefficients in **Q** implies a set of probabilities of the ancestral character states at the internal nodes of the tree. Let s_{ij} denote that j is the character state of ancestral node i, where here $j = 0$ or 1. Let $p(s_{ij} \mid D)_{i \in T}$ be the posterior probability of observing state j at node i, where $i \in T$ indicates that node i appears in the tree. The posterior probability is found by integrating over trees the probability of observing the data given that node i adopts state j, then dividing by the probability of observing either state at the node, also integrated over trees:

$$p(s_{ij} \mid D)_{i \in T} = \frac{\int_T \int_{\mathbf{Q}} p(D \mid s_{ij}, \mathbf{Q}, T)\, p(s_{ij})\, p(\mathbf{Q})\, p(T) d\mathbf{Q} dT}{\int_T \int_{\mathbf{Q}} \sum_j p(D \mid s_{ij}, \mathbf{Q}, T)\, p(s_{ij})\, p(\mathbf{Q})\, p(T) d\mathbf{Q} dT} \quad (5)$$

Here, $p(s_{ij})$ is the prior probability that j is the state at node i, and $p(T_i)$ is the prior probability of the tree. For some trees and combinations of the rate parameters, $p(s_{ij} \mid D)$ will be relatively large and for other combinations it will be small. Its distribution is estimated by sampling from a Markov chain that considers a large number of sets of rate coefficients and visits many trees.

The summation in the denominator of Equation 5 means that $p(s_{ij} \mid D)_{i \in T}$ can only be calculated for trees that have node i. If, say, 3/4 of the trees have the node, the remaining 1/4 will be ignored in deriving the posterior distribution. This is a general problem that arises whenever a node has a posterior probability of less than 1.0: how do we treat the trees in the MCMC sample that lack the node of interest? Huelsenbeck and Bollback (2001) constrain their estimates of ancestral states only to those trees that have the nodes of interest, and Lutzoni et al. (2001) choose only to examine nodes present in 95% or more of the trees. We show below that both of these approaches estimate the ancestral state in a biased subsample of the posterior density of trees, and overestimate the value of $p(s_{ij} \mid D)$.

To account for phylogenetic uncertainty when reconstructing an ancestral character state, we want to know the joint probability of a tree having a node and that the node adopts a particular state. This is given by Equation 6, which removes the summation in the denominator of Equation 5, replacing it with the integral over all trees of the probability of the data, independently of whether the tree contains node i:

$$p(s_{ij} \mid D) = \frac{\int_T \int_{\mathbf{Q}} p(D \mid s_{ij}, \mathbf{Q}, T)\, p(s_{ij})\, p(\mathbf{Q})\, p(T) d\mathbf{Q} dT}{\int_T \int_{\mathbf{Q}} p(D \mid \mathbf{Q}, T)\, p(\mathbf{Q})\, p(T) d\mathbf{Q} dT} \quad (6)$$

To see why Equation 5 overestimates $p(s_{ij} \mid D)$ consider that the term $p(s_{ij})$ now records the prior probability that the tree contains node i and that the node adopts state j. Because $p(s_{ij})$ is zero in trees that do not have node i, the numerator of Equation 6 is equal to that of Equation 5. However, because the denominator of Equation 6 now integrates over all trees, and not just those with node i, it will always exceed the value of the denominator of Equation 5 unless the node is present in all trees.

An intuitive understanding of Equation 6 can be derived from considering that

$$p(s_{ij} \mid D) \approx p(s_{ij} \mid D)_{i \in T} * p(i)$$

where $p(i)$ is the probability that node i exists, as estimated from the MCMC sample of trees. Thus the probability that j is the character state of ancestral node i is approximately equal to the product of two probabilities: that node i exists and the probability of state j at the node when it does exist. This returns the intuitively pleasing result that the probability that the node exists puts an upper limit on how certain one can ever be about the ancestral state of that node.

Application to Ribonuclease Evolution

Pancreatic ribonuclease is an enzyme involved in digestion. Jermann et al. (1995) studied the evolutionary history of the ribonuclease gene in Artiodactyls, using parsimony methods applied to a phylogeny derived from multiple sources. They suggested that ancestral ribonuclease genes had a glycine (G) at position 38 of the protein, and that aspartic acid (D) replaced it in the common ancestor to the ruminants. In vitro studies demonstrated that the G and D versions of ribonuclease differed five-fold in catalytic activity against double-stranded RNA. Schluter (1995) reanalyzed Jermann et al.'s result using a maximum likelihood (ML) model of trait evolution (Discrete; Pagel, 1994). The ML method reconstructed a slight and non-significant preference for D as the ancestral state with several replacements by G throughout the Artiodactyl clade. However, the dominant feature of the analysis was that ancestral states were uncertain such that either state could plausibly be assigned to the ancestral nodes.

Here we revisit this question to illustrate how to account for phylogenetic uncertainty in estimates of ancestral states.

MCMC Sampling of the Phylogeny

We found pancreatic ribonuclease gene-sequences in *GenBank* for 16 Artiodactyl species, plus a long-tailed Chinese hamster (*Cricetulus longicaudatus*) as an outgroup (see Fig. 2 caption for species names and accession numbers). Jermann et al.'s analysis included a number of nonpancreatic ribonucleases, which we have excluded. We used the 17 sequences to reconstruct the phylogeny and to provide information on the probable ancestral states of the amino acid (G or D) at position 38 of the enzyme. We make no claim that ribonucleases provide the best data for reconstructing the Artiodactyl phylogeny, as our aim is to show how to account for phylogenetic uncertainty rather than to advance a phylogenetic hypothesis (but see Results).

We aligned our sequences with Clustal X (Thompson et al., 1997) using default settings and excluding position 38. We then estimated the posterior distribution of trees from a Markov Chain implementing a single HKY + Γ (Yang, 1994, Swofford et al., 1996) model of sequence evolution. Our chain used uniform prior probabilities on trees and the parameters of the model of sequence evolution, and an exponential prior on branch lengths. We allowed the chain to reach convergence then sampled 500 trees at intervals of 20,000 trees to ensure that successive trees in our sample were statistically independent (see Results). This sample was then used in all of our analyses of trait evolution. We defined initial convergence as no change in the mean log-likelihood of the data on the tree over 200,000 iterations of the Markov chain. If at any point in the $500 \times 20{,}000 = 10{,}000{,}000$ tree sampling period that followed, the log-likelihood showed any long-term directional change, we abandoned the run. However, this never happened, and we ran at least five independent analyses of the data beginning from random starting points to check that each converged to the same point in tree space as judged by average log-likelihoods and posterior probabilities of nodes on the consensus tree. The chains all converged to the same region.

Five hundred trees is an arbitrary sample size, and it is difficult to provide rules of thumb for how many trees the sample should include. The sample should be large enough that all of the major different tree topologies are included. Fewer trees are required when the posterior probabilities of all nodes are high, and more when the posteriors are small. In the limiting case of all posteriors being 100, it is still necessary to sample enough trees to ensure that the chain has not simply missed an alternative topology, and to sample variation in branch lengths. If the posteriors seem not to change after a large number of trees has been sampled, this may be an indication that most or all of the frequently occurring topologies have been visited.

We might have employed Metropolis-coupled MCMC, or MCMCMC (Gilks et al., 1996), as an alternative to running a series of independent Markov chains. MCMCMC methods run some number of chains in parallel, and allow those chains to swap states. All but one of the chains is heated, allowing these chains to explore the tree space more readily. Swapping between chains promotes mixing by reducing the probability of any one chain getting stuck in a nonoptimal region of the universe of trees. All inferences are drawn from the unheated chain, and the results of the heated chains are discarded.

Our (unpublished) experience is that MCMCMC is of limited value in a phylogenetic context. Swapping of states is rare before convergence when the chains might be in different regions of the universe; and yet it is in these parts of the runs when exchanging information could be most valuable. Swapping becomes increasingly likely and even common once the chains have all converged, but little new information is gained. Moreover, a complication arises owing to some of the chains being heated. The heated chains will always have different stationary distributions to the unheated chain. This means that following every swap between a heated and the converged unheated chain, the unheated chain must be allowed to regain its stationary distribution.

By comparison, some number of independent MCMC chains begun from random starting points, requires the same computing power as one MCMCMC run with the same number of chains, but each chain can be used for inference. This makes the MCMC procedure, other things equal, more efficient. If all of the independent runs

converge to the same region of the tree space, this provides evidence that the chains have explored the tree space effectively.

MCMC Modeling of Trait Evolution

We constructed a Markov chain to implement the model of trait evolution, and used the chain to estimate the posterior probability distributions of the rate coefficients and of ancestral states. The model is defined by two rates, q_{GD} and q_{DG}, the likelihood function to evaluate the data on a tree given the rates, prior probability distributions for the rate parameters, and the distribution of trees from the MCMC sample.

The data consist of assigning a "D" to each species with aspartic acid at position 38 of the ribonuclease, and a "G" to species with a glycine. At each iteration the chain proposes a new combination of rate parameters and randomly selects a new tree from the sample of 500 trees. Random sampling of trees operates as a tree proposal mechanism within the Markov chain. The likelihood of the new combination is calculated and this new state of the chain is accepted or rejected following evaluation by the Metropolis-Hastings term.

We allowed this Markov chain to run for 1,000,000 generations after convergence (checked as for trees), ensuring that each tree in the MCMC sample was visited repeatedly. Rate coefficients were sampled from the Markov chain every 20 generations (autocorrelation = 0.05), yielding 50,000 sets of rate coefficients from which we estimated their joint probability distribution. Each sampled set of rate coefficients and its associated tree was also used to reconstruct ancestral states at specified nodes. From this we derived the posterior probability distribution of ancestral states over trees.

By fixing the sample of trees prior to running the trait-model Markov chain, our approach separates estimation of the phylogenetic tree from estimation of the model of trait evolution. Thus, we might have simultaneously estimated in a single Markov chain the posterior densities of the phylogenetic trees, their branch lengths, the parameters of the model of gene-sequence evolution, and the parameters of the model of trait evolution. However, traits are often selected for investigation in comparative studies because they evolve independently a number of times on the tree, a phenomenon known as homoplasy. The more homoplasy a character shows the less information it has about phylogenetic history. This may influence the way the Markov chain traverses the tree-space because combinations of trees and comparative results that return the highest likelihood may tend to favour less homoplasy. Including the comparative data in the MCMC sampling of trees could therefore risk distorting the estimate of the posterior density of trees.

The prior distributions.—Most of the controversies about the use of Bayesian methods revolve around the choice of prior distributions. This is because in some instances the particular form of the prior can dominate the posterior results. In other instances the data dominate the prior to such an extent that the posteriors derived from different priors are essentially equivalent. We placed a beta prior probability distribution on the rates of the model of trait evolution, allowing us to explore prior distributions of different 'informativeness.'

Two parameters, conventionally labeled α and β, jointly determine the beta distribution's mean and variance. Setting $\alpha = \beta = 1$ the beta distribution becomes uniform, corresponding to an uninformative prior. We allowed a uniform prior on the interval 0 to 100. For a second prior we first calculated the maximum likelihood estimate of the rate parameters of the model of trait evolution on each tree in our sample, using the program *Discrete* (Pagel, 1994). This gave for q_{GD} a mean of 5.82 ± 0.75 (mean $\pm$ standard deviation) and for q_{DG} 6.63 ± 1.17. We then chose the values of α and β to make the beta prior conform to these values. This is sometimes called an empirical Bayes estimator.

We chose the uniform prior and the maximum likelihood empirical Bayes prior deliberately to represent two extreme and probably unrealistic cases. Regarding the uniform prior, we do not believe a priori that all possible values of the rates are equally likely (for example, we do not believe that the rates are zero). Similarly, we do not have sufficient information to know in advance what the maximum likelihood values are for these data.

Thus, we derived a third prior, this one based upon the likelihood surface for the rate parameters. To find the likelihood surface we fixed the value of one rate parameter and then allowed the other rate parameter to vary from low to high values, calculating the likelihood of the data on the tree at each point. We then repeated this for the other rate parameter. We used the consensus tree from the $n = 500$ sample for these calculations. The typical likelihood surface is unimodal such that very low and very high values of the rate parameters returned poor likelihoods whereas intermediate values return higher likelihoods. The shape of their surfaces could be quite well characterized by a common beta distribution with a scaled mean of 8.7 and a variance of 29. This is also an empirical Bayes prior, but one that falls somewhere between the two extremes of the uniform and the maximum likelihood priors.

RESULTS

MCMC Sample of Phylogenetic Trees

Table 1 reports the autocorrelation coefficients in our MCMC sample for successive trees and for the parameters of the model of sequence evolution, indicating that

TABLE 1. Autocorrelations in the Markov Chain of trees. Entries show autocorrelations derived from trees sampled at intervals of 20,000 iterations in the Markov chain. The parameters κ and α are, respectively, the transition/transversion ratio from the HKY model of sequence evolution and the shape parameter of the gamma distribution.

	Log-likelihood	Tree length	κ	α
Autocorrelation, $r(n = 500)$	0.052	0.016	0.049	0.007

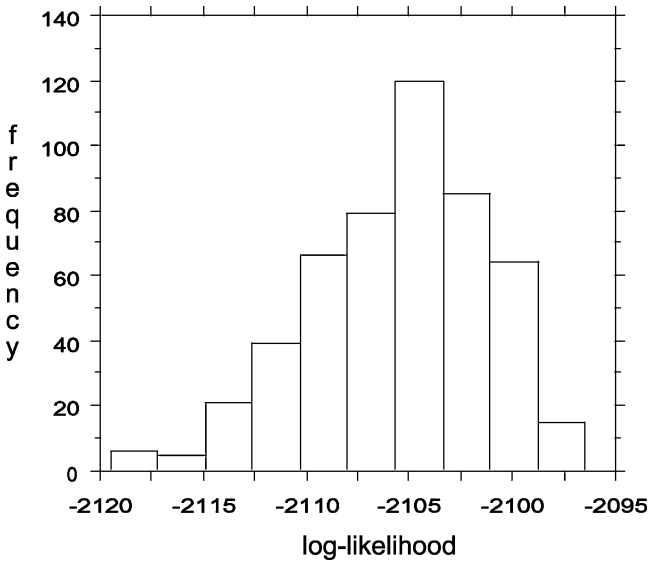

FIGURE 1. The frequency histogram of the $n = 500$ phylogenies sampled at intervals of 20,000 trees from the converged Markov Chain (see text). The left-skewed distribution has a mean log-likelihood of -2105 ± 4.30. The log-likelihoods span 23 log-units from a low of -2119 up to -2096.

both comprise independent samples of observations. If successive trees are not independent, a very large number may be required to estimate accurately the variance of their posterior probability distribution (Geyer, 1992).

Figure 1 shows the frequency histogram of the $n = 500$ phylogenies sampled from the converged Markov chain, forming a left-skewed distribution around a mean log-likelihood of -2105 ± 4.30 (mean $\pm$ SD). The frequency histogram shows that trees of with high or low log-likelihoods have a low posterior probability under the model of evolution, whereas trees of intermediate likelihood arise frequently. The long left-hand tail of the distribution identifies mostly unsampled trees that are possible under the model of evolution but of such low probability that the Markov chain has only a small chance of visiting them. Because they account for a small proportion of the probability density, they can be safely ignored when estimating parameters.

Figure 2 shows the consensus tree for the sample with the posterior probabilities of each node labeled below the line, and the amino acid at position 38 shown for each species. The tree confirms the by now familiar relationship of the cetacea to the Hippopotamus. The inner clade of ruminants, spanning the Chevrotain (*Tragulis javanicus*) to the cow (*Bos taurus*), is well supported, appearing in every tree in the sample. Other nodes within the ruminants, notably the Bovidae (goat, sheep, buffalo, cow) are less secure. The only substantial difference between our consensus tree and another recent tree of the Artiodactyls (Hassanin and Douzery, 2003), based on three mitochondrial and four nuclear gene segments, is that our ribonuclease tree more often places the Giraffe and Pronghorn inside the Roe and Hog deer species.

Estimates of the Rate Coefficients of the Model of Trait Evolution

Table 2 reports the mean and variance of the posterior distributions of the rate coefficients obtained from the three different beta prior probability distributions. Owing to the small amount of trait data, the priors exert a large influence on the posterior distributions. In such cases the choice of a prior must be carefully justified. The uniform or uninformative prior yields poor average log-likelihoods for the model of trait evolution and the rate posteriors have large variances and means. The maximum likelihood empirical Bayes prior unsurprisingly returns very narrow posterior distributions of the rates and the log-likelihood. The beta prior designed to mimic the likelihood surface of both rate parameters is less restrictive than the maximum likelihood prior. Nevertheless it returns quite similar average log-likelihoods to the ML prior, and the posterior distributions of the rates have smaller means and variances than the posteriors derived from the uniform prior. They also have smaller variances than their corresponding prior. These results suggest that the prior derived from the likelihood surface of the rates yields posteriors that produce good fits of the data to the model of trait evolution without imposing unrealistic prior restrictions on the rates. We therefore used this intermediate prior in all further analyses.

Figure 3 plots the 50,000 points that sample the joint posterior density of q_{DG} and q_{GD}, based upon the intermediate beta prior. The bivariate distribution displays the probability of each pair of rate coefficients under the model of trait evolution, given the trait data. The correlation between the pairs of rates is $r = 0.37$. For any value of one rate, there is a wide range of rates for the other parameter. This is one of the benefits of finding the joint posterior: it characterizes the range of plausible hypotheses about the values that the parameters of the model might take.

The marginal posterior distributions at the top and side of the Figure 3 illustrate that the bulk of the bivariate density of rates clusters around the two means. The mean and standard deviation of the instantaneous

TABLE 2. Posterior distributions derived from three different prior distributions. All priors are beta distributed. The uniform prior assumes that the two parameters of the beta distribution are $\alpha = \beta = 1$; the 'maximum likelihood' prior bases the beta distribution on the mean and variance across trees of the maximum likelihood values of the rate parameters; the 'likelihood surface' prior bases the beta distribution on the mean and variance of the likelihood surface of the rate parameters on the consensus tree.

	Average $\pm$ SD of posterior distribution		
Prior	q_{GD}	q_{DG}	Log-likelihood
Uniform q_{GD}, q_{DG}: 0–100	11.47 ± 6.28	13.28 ± 9.13	-16.63 ± 1.50
Maximum likelihood q_{GD}: beta(5.82 ± 0.75) q_{DG}: beta(6.63 ± 1.17)	5.82 ± 0.90	6.63 ± 1.01	-15.22 ± 0.59
Likelihood surface q_{GD}, q_{DG}: beta(8.2 ± 5.39)	7.42 ± 3.48	8.05 ± 3.77	-15.82 ± 0.89

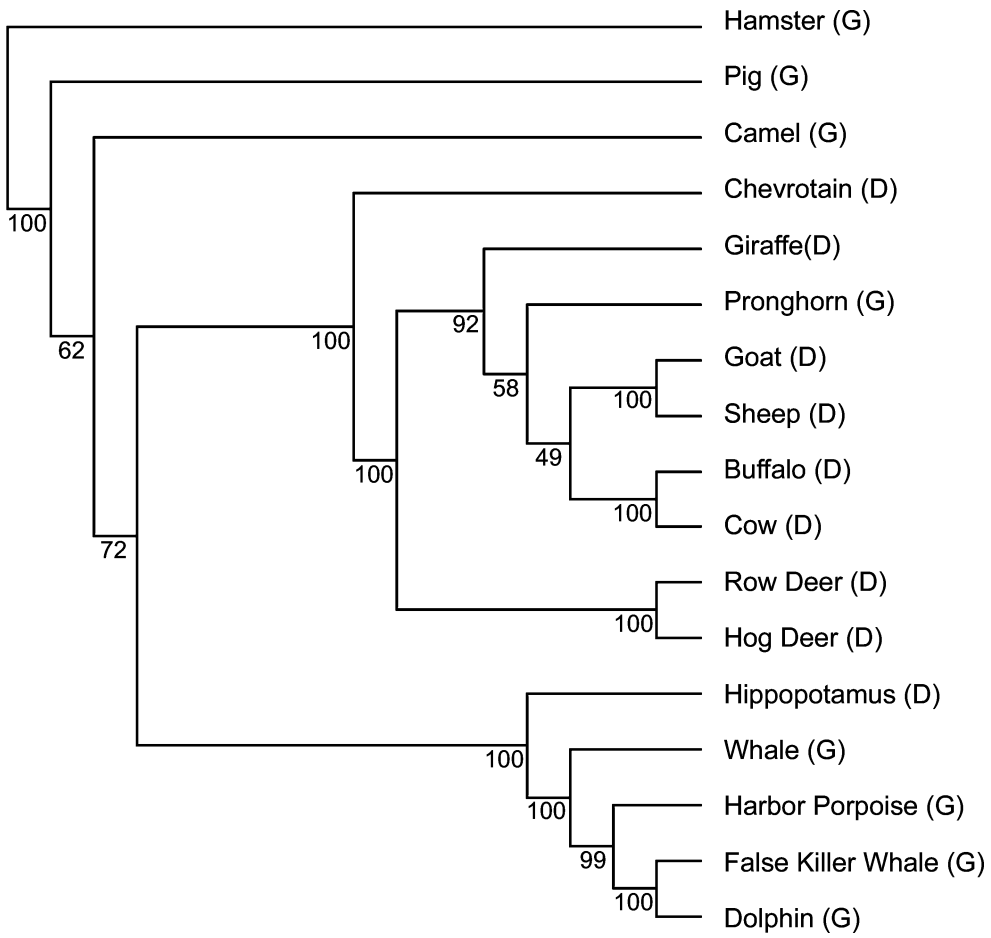

FIGURE 2. The consensus tree of the $n = 500$ trees sampled at intervals of 20,000 trees from the converged Markov chain. Autocorrelation between successive trees is 0.052 (Table 1). The numbers below the lines at the nodes are the Bayesian posterior probabilities, corresponding to the proportion of trees in the sample that have each particular node. The accession numbers and names for each species are as follows: Pronghorn (*Antilocapra Americana*, AJ271301); Lesser Malay Chevrotain (*Tragulus javanicus*, AJ271297); Water buffalo (*Bubalus bubalis*, AJ011843); Arabian camel (*Camelus dromedaries*, AJ006418); Pig (*Sus scrofa*, AJ005521); Harbor porpoise (*Phocoena phocoena*, AJ005523); False killer whale (*Pseudorca crassidens*, AJ005524); Hippopotamus (*Hippopotamus amphibious*, AJ005522); Cow (*Bos taurus*, S81740); Chinese long-tailed hamster (*Cricetulus longicaudatus*, X62945); Bottle-nosed dolphin (*Tursiops truncatus*, AY261462); Bowhead whale (*Balaena mysticetus*, AY261461); Roe deer (*Capreolus capreolus*, Y11672); Hog deer (*Axis porcinus*, Y11669); Goat (*Capra hircus*, S81742); Sheep (*Ovis aries*, S81741); Giraffe (*Giraffa camelopardalis*, S81739).

rate of gain of D (q_{GD}) is 7.42 ± 3.48, and the mean rate of loss of D (q_{DG}) is 8.05 ± 3.77 These posterior distributions can be well characterized by beta probability curves and Figure 4 plots them along with the likelihood-surface beta prior distribution. Relative to the prior, the posteriors concentrate their density among intermediate values, translating into narrower 95% credibility intervals: 95% equal-tailed credibility interval of the prior is 0.87 to 20.7; for the q_{GD} posterior 2.19 to 15.8; and for the q_{DG} posterior 2.09 to 16.6. In general the posteriors are 1.5 to 2 standard deviations narrower than the prior, showing that these data carry some information about their posterior distribution that is independent of the prior.

Applying the average values of the rate coefficients to the average tree length of 0.91 ± 0.9 implies about 6 gains and 7 losses over the roughly 60 million years Jermann et al. (1995) suggest that this tree spans. However, this is only a very rough approximation as the continuous-time Markov model allows more than one change per branch and long branches may conceal multiple changes.

Within- and between-tree variance in the rate parameters.— Trees can vary in topology and branch lengths and so may yield different posterior distributions of the rate coefficients and log-likelihoods of the model of trait evolution. On the other hand, if the posterior distributions (as discovered by the Markov chain) are the same in every tree, then the between-tree component of variance will be zero, and the contribution of phylogenetic uncertainty to the results is unimportant.

We analyzed the 50,000 observations in Figure 3 by a simple one-way analysis of variance (ANOVA), grouping the data into 500 bins corresponding to the 500 different phylogenetic trees in our MCMC sample (Table 3). Because the 500 trees are a probability sample, this partitioning of the data automatically weights different trees according to their posterior probabilities in the universe of trees. The small but significant F-ratios for

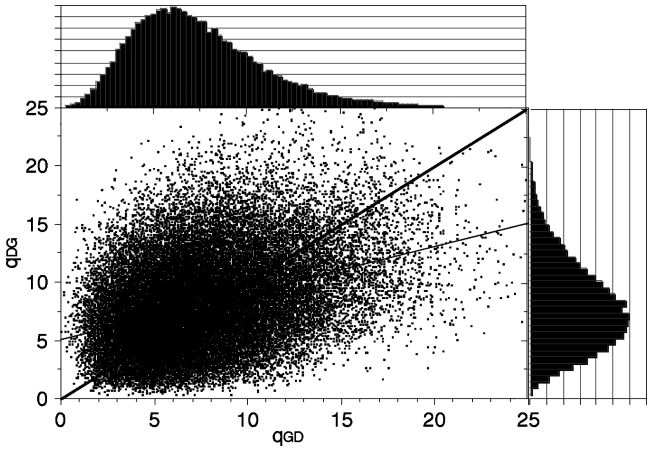

FIGURE 3. The joint posterior density of the rate parameters (q_{GD}, q_{DG}) from the model of trait evolution. The approximately bivariate normal distribution is drawn from 50,000 pairs of rate coefficients sampled from a converged Markov chain implementing the continuous-time Markov model of trait evolution and moving among trees (see text). The parameter q_{GD} measures the instantaneous rate of gain of aspartic acid (D) at position 38 of the ribonuclease gene, whereas q_{DG} measures the rate of loss. The marginal distributions of q_{GD}, and q_{DG} are plotted on their respective axes. The mean instantaneous rate of gain of D (q_{GD}) was 7.42 ± 3.48, and the mean rate of loss of D (q_{DG}) was 8.05 ± 3.77. The mean q_{DG}/q_{GD} ratio for paired values is 1.29 ± 1.29. The steeper diagonal line conforms to the 1:1 ratio and the shallower line is the regression line: $q_{DG} = 5.09 + 0.40\ (q_{GD})$; $r = 0.37$.

the rates show that the means of the posterior distributions of rate coefficients vary somewhat among trees. The much larger *F*-ratio for the log-likelihood indicates that the model of trait evolution fits some trees much better than others. All three of these results demonstrate why accounting for phylogenetic uncertainty by integrating analyses over trees is important in comparative studies: the results from one tree may differ from those of another.

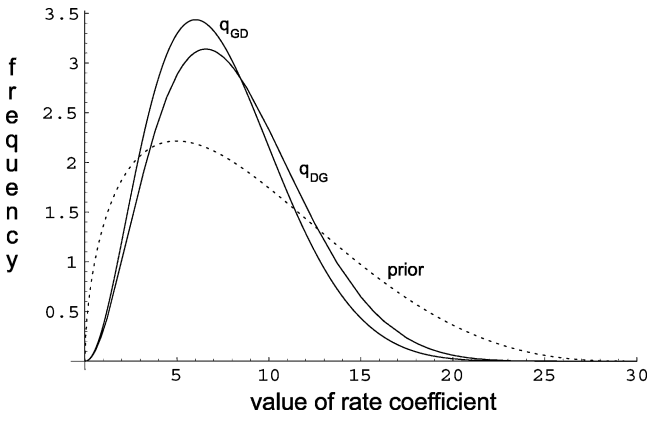

FIGURE 4. Plot of the prior distribution of rate coefficients and the two posterior distributions, represented as beta probability distributions: prior (mean = 8.7, variance = 29); q_{GD} posterior (mean = 7.42, variance = 12.11); q_{DG} posterior (mean = 8.05, variance = 14.21). Both posterior distributions concentrate more of their results in intermediate values. The 95% equal-tailed credibility interval of the prior is 0.87 to 20.7; for the q_{GD} posterior 2.19 to 15.8; and for the q_{DG} posterior 2.09 to 16.6.

TABLE 3. Estimates of the between- and within-tree components of variance in ancestral state reconstructions. The MS between has 499 *df* and the MS within has 50,000 − 499 or 49,501 *df* (see text).

Measure	Mean square between trees	Mean square within trees	F = MSB/MSW	*P*-value
q_{GD}	34.2	11.9	2.87	<0.001
q_{DG}	65.7	13.7	4.81	<0.001
Log-likelihood	32.85	0.45	73.35	<0.0001

Reconstructed Ancestral States

Figures 5 shows the consensus phylogeny with the panels on the left depicting the estimated posterior probabilities of glycine at eight selected nodes. The numbers above the panels are written as *p(G|n)* to denote that the distributions are derived from only those trees in which the node exists, and correspond to the estimator of Equation 5. Because the trait is binary, the probability of aspartic acid is just $p(D \mid n) = 1 - p(G \mid n)$. Beneath each panel we combine the information on the posterior probability of the ancestral state with the posterior probability of the node, by finding their product, following the approximation to Equation 6. That is, we find $p(D \mid n) * p(n)$ and $p(G \mid n) * p(n)$. The sum of these two combined probabilities is always equal to the node's posterior probability. The remainder of the probability, summing to 1, corresponds to the probability that the node does not exist. Thus, for any node we can write three probabilities that sum to 1.0 corresponding to the probabilities of the alternative states given that the node exists plus the probability that the node does not exist:

$$1.0 = p(D \mid n) * p(n) + p(G \mid n) * p(n) + (1 - p(n))$$

For traits with more than two states the same equality holds, but there will be a probability for each of the alternative states, given that the node exists.

Beginning with node 1, the reconstructions suggest that with probability 0.81 ± 0.11, glycine was the ancestral state of the Artiodactyls. Individual reconstructed probabilities at this node range from 0.43 to 0.95. An analysis of variance of the probabilities, using tree as the grouping factor, gives a sense of the contribution of phylogenetic uncertainty to inferences of this ancestral state. The ANOVA returns a highly significant *F*-statistic of 82.72 (mean square between groups = 0.55, mean square within groups = 0.007), $P \ll 0.001$.

The reconstructions indicate that glycine probably remained ancestral up through nodes 2, 3, and 4. However, the breadth of the posterior distributions at these nodes coupled with the uncertainty about the nodes themselves means that other evolutionary scenarios cannot be ruled out. The uncertainty about the true tree quite correctly limits what we can be said about the ancestral states.

Allowing glycine to be ancestral, aspartic acid evolved at least once in the branch leading to Hippopotamus.

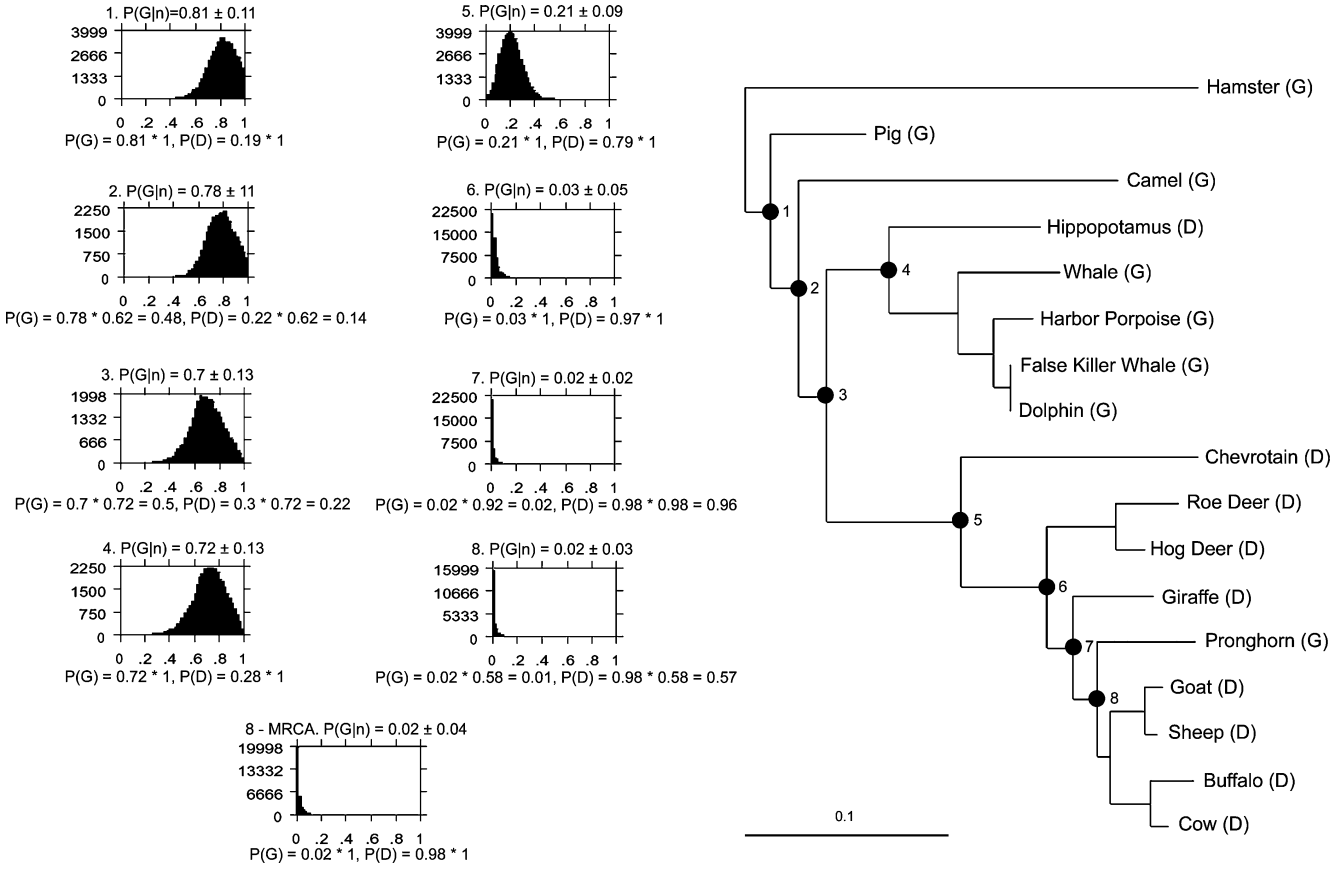

FIGURE 5. The consensus tree of Figure 2 with panels showing the posterior densities of the reconstructed ancestral states at eight selected nodes as numbered on the tree. The branch lengths of this tree are the maximum likelihood values obtained from applying to the consensus topology the HKY + Γ model using four rate categories. Above each panel $p(G \mid n)$ reports the conditional probability of G for trees that have the node (Equation 5). Under each panel, we report $p(G)$ and $p(D)$ from Equation 6 by weighting $p(G \mid n)$ and $p(D \mid n) = 1 - p(G \mid n)$ by the posterior probability of the node. This value combines the uncertainty about the ancestral state with the uncertainty about the node itself. $p(G)$ and $p(D)$ must sum to the posterior probability of the node, $p(n)$. This means that for a proportion $1 - p(n)$ of trees at each node the ancestral state cannot be calculated because the node does not exist. The lower right hand panel labeled "mrca" reports the posterior probability for glycine obtained when we reconstruct the ancestral state of the most recent common ancestor to the Pronghorn-Cow clade in each tree (see text).

Between nodes 3 and 5, the ancestral state posterior distribution drops to a mean of about 0.21 and barely overlaps 0.5, suggesting that aspartic acid (D) was probably gained near the origin of the ruminants. The high posterior probability of node 5 makes this conclusion more secure. The ancestral state posterior distributions for nodes 6 and 7 have means near zero and narrow confidence limits, suggesting that D was most likely retained as ancestral throughout this clade, eventually being lost in the lineage leading to Pronghorn. Both of these nodes also have high posterior probabilities.

A Most-Recent-Common-Ancestor Approach

The conditional estimator of Equation 5 reconstructs node 8 as aspartic acid with 0.98 probability, but the low posterior probability of the node (0.58) weakens this inference. Owing to the uncertainty of where the Pronghorn-Cow clade falls in any given tree, and whether it falls in amongst other species, we cannot say with confidence what this group's likely ancestral state was. This is important if we wish to regard the G in Pronghorn as a loss of aspartic acid.

To work around the limitation of poorly supported nodes, we can for each tree identify the most recent common ancestor to a group of species and reconstruct the state of that node, then combine this information across trees. For some of the trees, this node will include the species of interest plus some others, whereas for other trees the node will uniquely define the group of interest. The reconstructed ancestral state using the 'mrca' approach for node 8 is also 0.98, but now this 'floating' node appears by definition in every tree, giving us confidence that this group's ancestral state was aspartic acid. The posterior density of the most-recent-common-ancestor node reconstruction is shown in the bottom panel of Figure 5. This technique can strengthen one's confidence in the ancestral state of a group that occupies a weakly supported node in the tree.

DISCUSSION

Statistical models applied to data collected across species can be used to discover and characterize the underlying evolutionary processes that gave rise to observed patterns of species diversity (Pagel, 1997, 1999a). That is, one can infer the trends, ancestral states, and patterns that most likely held in the past, given what is observed now. The methods we have described here implement statistical models designed to detect such processes, and allow investigators to estimate the posterior probability distributions of the parameters that characterize them. The posterior distributions account for phylogenetic uncertainty in the sense that the posterior probability distribution for any given parameter includes all of the variability (uncertainty) attributable to fitting the model to different trees. Applying these methods to the ribonuclease data, we find evidence that glycine was ancestral but that aspartic acid probably evolved somewhere in the lineage leading to Hippopotamus, and then again around the origin of the ruminants. However, we find a considerable range of possibilities at earlier nodes in the tree, emphasizing that different evolutionary scenarios could have occurred.

A key difference between methods designed to account for uncertainty and 'optimizing' or parsimony methods is the shift away from single 'best' trees or optimal reconstructions of ancestral states or rates and correlations, and towards estimating one's confidence in alternative hypotheses on the data. In the likelihood approaches that we have described, the observed data are treated as fixed (Edwards, 1972), and we allow the hypotheses on those data to vary. Each pair of gain and loss parameters in our model of trait evolution (Fig. 3) is a hypothesis on the trait data. In the case of the glycine/aspartic acid trait data, these posterior distributions do not pin down the precise rates of gains and losses with much certainty. That is, a range of hypotheses about the rates is consistent with the data. However, the shape of the posterior distribution tells us in which hypotheses we should have the most confidence.

A similar picture emerges for the ancestral states. Some nodes admit only a narrow range of probabilities and are reconstructed with a high degree of certainty. For others, the picture is variable. This situation is often worsened when the confidence in the existence of the ancestral node itself is taken into account because a node's posterior probability puts an upper limit on the confidence in the ancestral state. However, we show how to use a 'most-recent-common-ancestor' approach to examine an alternative hypothesis for nodes that are poorly supported. Given a set of species whose common ancestor's ancestral state is of interest, it is straightforward to find a node on each tree in the MCMC sample that includes those species. In some trees the node will include only the species of interest, whereas in others it will include these species and others; but there will be a node to reconstruct in every tree. The ancestral state of this node can be reconstructed and its posterior density examined. We have shown that this technique can improve confidence in the posterior probability of an ancestral state.

The posterior distributions, in addition to pointing towards what is most probable, can also be used to put probabilities on explicit alternative hypotheses on the data. We find, for example, that the rate of loss of aspartic acid (q_{DG}) slightly exceeds the rate at which it is gained (q_{GD}). How probable is a model hypothesizing equal rates of gain and loss? This can be read directly from the joint posterior of the rate coefficients, where we find that approximately 10.3% of the pairs fall within ± 0.5 units of each other. There is no strong tendency, then, for the model of trait evolution to be drawn towards a 1:1 ratio of the rates. Neither is there a tendency to be drawn towards the 1:2 ratio corresponding to the numbers of gains of G ($n = 1$) versus gains of D ($n = 2$) that may seem apparent on the tree. Only 10.2% of pairs have a ratio between 0.4 and 0.6. This result emphasizes that mapping traits to count changes can be misleading, and that long branches in particular may conceal more than one change in a trait.

An attractive feature of the likelihood approach for reconstructing ancestral states is that probabilities are calculated by summing over all possible histories at each node (Pagel, 1994), without ever actually assigning any ancestral state. Only the observed data are fixed. In a Bayesian-MCMC setting this means that the posterior distributions of the rate parameters of the model of trait evolution and the posterior distributions of the ancestral states at the internal nodes (the reconstructed ancestral states) are, by definition, all consistent with the observed data.

The Bayesian approach we describe is easily generalized to investigate other evolutionary processes. For example, to investigate correlated evolution of two binary traits, Pagel (1994) developed a likelihood ratio statistic that compares two models of evolution. In one, the traits are allowed to evolve independently on the tree. In the other, they evolve in a dependent or correlated fashion. Evidence for correlated evolution takes the form of the dependent model fitting the data better, the evidence of which is a larger likelihood. In a Bayesian context we may wish to derive the posterior density of the independent model and use it as our prior for the dependent model, asking whether the dependent posterior distribution has shifted away from the prior. If it does this provides evidence to favor the dependent model. Pagel and Meade (in press) apply this MCMC approach to investigating correlated evolution in an anthropological context.

Bayesian approaches to phylogenetic inference and comparative studies reveal a much fuller sense of the information in the data than optimizing approaches and should be widely applied. It is now easily possible to work with trees of hundreds of organisms using inexpensive desktop computers.

ACKNOWLEDGMENTS

We thank Chris Simon for inviting this article, a preliminary version of which was presented in the Molecular Phylogenetics Symposium

at the XIXth International Congress of Genetics, Melbourne, 2003. This work is supported by grant numbers 45/G14980 and 45/G19848 to MP from the Biotechnology and Biological Sciences Research Council (UK). Jeff Thorne provided valuable comments on a previous draft. The software *BayesMultiState* can be downloaded from www.ams.reading.ac.uk/zoology/pagel.

REFERENCES

Cunningham, C. 1999. Some limitations of ancestral character-state reconstruction when testing evolutionary hypotheses. Syst. Biol. 48:665–674.

Edwards, A. W. F. 1972. Likelihood. The Johns Hopkins University Press.

Efron, B., E. Halloran, and S. Holmes. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA 93:13429–13434.

Felsenstein, J., and H. Kishino. 1993. Is there something wrong with the bootstrap? A reply to Hillis and Bull. Syst. Biol. 42:193–200.

Galtier, N., N. Tourasse, and M. Gouy. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science 283:220–221.

Geyer, C. J. 1992. Practical Markov chain Monte Carlo. Stat. Sci. 7:473–511.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. Introducing Markov chain Monte Carlo. Pages 1–19 *in* Markov Chain Monte Carlo in Practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). Chapman and Hall, London.

Hassanin, A., and E. J. P. Douzery 2003. Molecular and morphological phylogenies of Ruminantia, and the alternative position of the Moschidae. Syst. Biol. 52:206–228.

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

Hibbett D. S., L. B. Gilbert, and M. J. Donoghue. 2000. Evolutionary instability of ectomycorrhizal symbioses in basidiomycetes. Nature 407:506–508.

Holden, C. J. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. Proc. R. Soc. Lond. B 269:793–799.

Huelsenbeck, J. P., and J. P. Bollback. 2001. Empirical and hierarchical Bayesian estimation of ancestral states. Syst. Biol. 50:351–366.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

Jermann, T. M., J. G. Opitz, J. Stackhouse, and S. A. Benner. 1995. Reconstructing the evolutionary history of the artiodactyl ribonulcease superfamily. Nature 374:57–59.

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. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.

Lutzoni, F., M. Pagel, and V. Reeb. 2001. Major fungal lineages derived from lichen-symbiotic ancestors. Nature 411:937–940.

Mau, B., M. Newton, and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092.

Mooers, A. Ø., and D. Schluter. 1999. Support for one and two rate models of discrete trait evolution. Syst. Biol. 48:623–633.

Newton, M. 1996. Bootstrapping phylogenies: Large deviations and dispersion effects. Biometrika 83:315–328.

Newton, M., B. Mau, and B. Larget. 1999. Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. Pages 143–162 *in* Statistics in molecular biology (F. Seillier-Moseiwitch, T. P. Speed, and M. Waterman, eds.). Institute of Mathematical Statistics, Cambridge, UK.

Nielsen, R. 2002. Mapping mutations on phylogenies. Syst. Biol. 51:729–739.

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. Scripta 26:331–348.

Pagel, M. 1999a. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Pagel, M. 1999b. The maximum likelihood approach to reconstructing ancestral character states on phylogenies. Syst. Biol. 48:612–622.

Pagel, M., and F. Lutzoni. 2002. Accounting for phylogenetic uncertainty in comparative studies of evolution and adaptation. Pages 148–161 *in* Biological evolution and statistical physics (M. Lässig and A. Valleriani, eds.). Springer–Verlag, Berlin.

Pagel, M., and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53:571–581.

Pagel, M., and A. Meade. In press. Bayesian estimation of correlated evolution across cultures: A case study of marriage systems and wealth transfer at marriage. To appear *in* The evolution of cultural diversity: A phylogenetic approach (R. Mace, C.J. Holden, and S. Shennan, eds.). University College London Press, London.

Rannala, B., and Z. Yang. 1996. Probability distributions of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

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.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876–4882.

Wilson, I., and D. Balding. 1998. Genealogical inference from microsatellite data. Genetics 150:499–510.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

First submitted 20 October 2003; reviews returned 5 January 2004; final acceptance 10 June 2004
Associate Editor: Jeffrey Thorne