Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees

Bret Larget and Donald L. Simon

Department of Mathematics and Computer Science, Duquesne University

We further develop the Bayesian framework for analyzing aligned nucleotide sequence data to reconstruct phylogenies, assess uncertainty in the reconstructions, and perform other statistical inferences. We employ a Markov chain Monte Carlo sampler to sample trees and model parameter values from their joint posterior distribution. All statistical inferences are naturally based on this sample. The sample provides a most-probable tree with posterior probabilities for each clade, information that is qualitatively similar to that for the maximum-likelihood tree with bootstrap proportions and permits further inferences on tree topology, branch lengths, and model parameter values. On moderately large trees, the computational advantage of our method over bootstrapping a maximum-likelihood analysis can be considerable. In an example with 31 taxa, the time expended by our software is orders of magnitude less than that a widely used phylogeny package for bootstrapping maximum likelihood estimation would require to achieve comparable statistical accuracy. While there has been substantial debate over the proper interpretation of bootstrap proportions, Bayesian posterior probabilities clearly and directly quantify uncertainty in questions of biological interest, at least from a Bayesian perspective. Because our tree proposal algorithms are independent of the choice of likelihood function, they could also be used in conjunction with likelihood models more complex than those we have currently implemented.

Introduction

The traditional methods for phylogenetic inference select a single ''best'' tree, either according to some optimality criterion (maximum likelihood, maximum parsimony) or by a clustering algorithm (neighbor joining). Uncertainty may then be assessed by a subsequent procedure, such as the bootstrap. In contrast, a Bayesian approach to phylogeny reconstruction expresses the uncertainty in the phylogeny and in the parameters of the sequence mutation model with a posterior probability distribution. Summaries of parameters of interest, such as the tree topology, are described by their marginal posterior distributions. Equations which express the desired summaries are analytically intractable for even small phylogeny problems. The approach which has proven to be successful for many such intractable analytical Bayesian analyses is to use stochastic simulation to obtain a sample from the posterior distribution and to base inferences on this sample. (See Gelman et al. [1995] for an accessible and practical introduction to modern Bayesian methods.)

This paper describes a Bayesian approach to phylogeny reconstruction and introduces novel Markov chain Monte Carlo (MCMC) algorithms to solve the computational aspects of the problem. We demonstrate our methodology with two examples. While in many regards, the approach we advocate has a similar goal to an approach using maximum likelihood with bootstrapping, a Bayesian approach enjoys a substantial computational advantage in the examples we have studied. We make comparisons between the Bayesian computational approach described here and analysis by maximum likelihood with bootstrapping in the discussion.

We are aware of three groups who began working independently on a Bayesian approach to phylogenetic inference using MCMC at about the same time. The dissertation work of Mau (1996) led to additional papers. Mau and Newton (1997) make phylogenetic inferences with restriction site data. Mau, Newton, and Larget (1999) introduce the MCMC sampler that is the precursor to the tree proposal algorithms we introduce in this paper. These same authors apply their sampler in a Bayesian study of the coevolution of pocket gophers and their parasitic lice (Newton, Mau, and Larget 1999). A second disseration (Li 1996) also explores an MCMC approach to phylogeny reconstruction, but we are not aware of published papers in scientific journals resulting from this work. Rannala and Yang (1996) describe a Bayesian analysis for which the computational approach is suitable only for very small trees, and they employ MCMC on a somewhat larger tree in a subsequent paper (Yang and Rannala 1997). We compare our computational approach with the methods of these other authors in the discussion.

The MCMC papers on phylogenetic inference cited above all assume a molecular clock. This paper gives a different description of the tree proposal mechanism described in the papers authored by Mau, Newton, and Larget cited above that leads to a simpler introduction to the novel nonclock version presented here.

Furthermore, we introduce a completely different pair of algorithms which enjoy a computational advantage over the previously published methods. We demonstrate that a Bayesian approach to phylogeny reconstruction makes assessment of uncertainty computationally practical on trees far larger than those currently being assessed using maximum likelihood and bootstrapping.

Key words: Markov chain Monte Carlo, Metropolis-Hastings algorithm, phylogeny, tree reconstruction, Bayesian statistics.

Address for correspondence and reprints: Bret Larget, Department of Mathematics and Computer Science, Duquesne University, College Hall 440, Pittsburgh, Pennsylvania 15282.
E-mail: larget@mathcs.duq.edu.

Mol. Biol. Evol. 16(6):750–759. 1999

Materials and Methods

A Bayesian Approach

A Bayesian approach to phylogeny reconstruction requires a likelihood model for sequence evolution through a phylogenetic tree, prior distributions on trees and model parameters, and data. A tree $\psi = (\tau, \beta)$ is described by its tree topology τ and associated branch lengths β. The likelihood model $L(x|\omega)$ for observed data x may contain several parameters φ, where $\omega = (\psi, \varphi)$ represents a specific choice of tree topology, branch lengths, and model parameters. Parameter space $\Omega = (\Psi, \Phi)$ contains the sets of all possible trees Ψ and model parameters Φ. The tree topology is discrete, and its values partition Ω.

A fully Bayesian analysis models the prior uncertainty in ω with a joint prior distribution $p(\omega)$ for all the parameters in the space. The product of the likelihood function and the prior distribution, normalized to have volume 1 over Ω, is the joint posterior distribution upon which all inference is based, expressed as

$$p(\omega|x) = \frac{L(x|\omega)p(\omega)}{\int_\Omega L(x|\omega)p(\omega)\, d\omega}. \quad (1)$$

Notice that the numerator in equation (1) may be evaluated for any point ω, but that computing the denominator can be infeasible for even fairly small trees.

To find the posterior probability of a particular tree topology τ, we need to find the volume under $p(\omega|x)$ in the portion of the partition of Ω which corresponds to τ by integrating out all other parameters. A point $\omega = (\tau, \beta, \phi)$ may be broken into its component parts, and we have

$$p(\tau|x) = \frac{\int_B \int_\Phi L(x|\tau, \beta, \varphi)p(\tau, \beta, \varphi)\, d\varphi\, d\beta}{\sum_\tau \int_B \int_\Phi L(x|\tau, \beta, \varphi)p(\tau, \beta, \varphi)\, d\varphi\, d\beta}, \quad (2)$$

where B and Φ are the sets of all possible branch lengths and model parameter values, respectively. If we wanted to know the posterior probability that a group of taxa formed a monophyletic clade, we would sum the posterior probabilities of all topologies τ that satisfied this condition.

Markov Chain Monte Carlo

The Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970) samples a dependent sequence of points in Ω, $\omega^{(0)}, \omega^{(1)}, \omega^{(2)}, \ldots$, such that after some point in the sequence, all subsequent sampled points are distributed approximately according to the posterior distribution. As a consequence, after discarding an initial portion of the sequence, the long-run frequencies of the sampled tree topologies are arbitrarily close to their posterior probabilities after sufficiently long simulations, by the Markov chain law of large numbers (Theorem 3 in Tierney 1994). Smith and Roberts (1993) give several advantages of a sample-based approach to Bayesian inference, including the ability to do graphical exploratory data analysis, inference, prediction, and model validation. Because Ω includes both tree and model parameter information, a sample to estimate the tree topology posterior probabilities of equation (2) can also provide inferences about branch lengths and model parameter values. Gelman et al. (1995, chapter 11) contains an accessible introduction to MCMC methods including the Metropolis-Hastings algorithm.

We begin with a Markov chain on Ω that proposes a move to state ω_2 from the current state ω_1 according to probability density function $q(\omega_1, \omega_2)$. The Metropolis-Hastings algorithm modifies these transition probability densities so that the resultant stationary distribution is the desired posterior distribution. In theory, any irreducible Markov chain may be modified by the Metropolis-Hastings algorithm so that long-run frequencies converge with probability one to the appropriate posterior probabilities. The art is in designing a Markov chain that rapidly traverses the posterior distribution so that inferences based on samples short enough to be computationally feasible will be sufficiently accurate.

The Metropolis-Hastings algorithm accepts a proposed new state ω^* from current state ω with probability

$$\min\left(1, \frac{p(\omega^*|x)q(\omega^*, \omega)}{p(\omega|x)q(\omega, \omega^*)}\right). \quad (3)$$

If the proposal is rejected, the current state is repeated in the sequence. Often, q is symmetric, and the Hastings ratio $q(\omega^*, \omega)/q(\omega, \omega^*)$ equals 1 and does not affect the acceptance probability. Notice that the posterior density appears only as a ratio in equation (3) so that the denominator in equation (1) cancels.

We will actually use a composition of two different basic update mechanisms to traverse Ω. Specifically, we begin with a randomly chosen initial tree and model parameter values, $\omega^{(0)} = (\psi^{(0)}, \varphi^{(0)})$, from some very dispersed distribution. Given the current state of $\omega^{(i)} = (\psi^{(i)}, \varphi^{(i)})$, a single cycle will consist of two stages. In the first stage, while keeping the current tree $\psi^{(i)}$ fixed, we propose new model parameters φ^* with a Markov chain q_1 on the space of model parameter values Φ which are either accepted ($\varphi^{(i+1)} = \varphi^*$) or rejected ($\varphi^{(i+1)} = \varphi^{(i)}$) with acceptance probability from equation (3). The second stage modifies the current tree $\psi^{(i)}$ in a sequence of steps while holding $\varphi^{(i+1)}$ fixed. One step of the second stage proposes a new tree ψ^* according to a Markov chain q_2 on Ψ which is accepted or rejected and repeats this process a fixed number of times. The tree $\psi^{(i+1)}$ is the result of a fixed number of Metropolis-Hastings steps according to q_2 from $\psi^{(i)}$.

MCMC Algorithms for Proposing New Trees

We describe two different algorithms for proposing new trees. Each algorithm has two versions, one which assumes a molecular clock and one which does not. The molecular clock version of the GLOBAL algorithm, which modifies all branch lengths and potentially changes the tree topology simultaneously, is equivalent to the algorithm presented in Mau, Newton, Larget (1999). Our

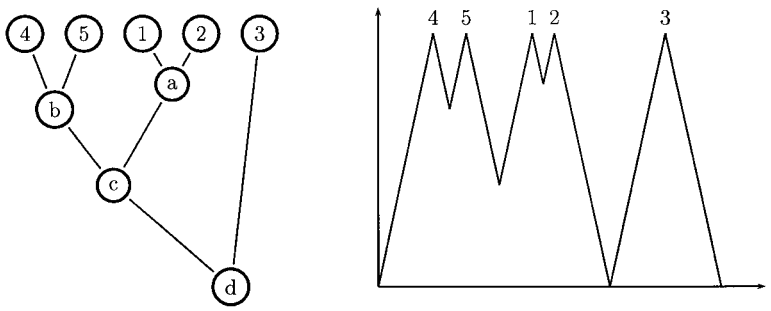

FIG. 1.—A tree and its representation under the molecular clock assumption. The peaks are all at the same height. The permutation of five leaf labels and the four ordered valley depths determine the tree completely. There are 2^4 equivalent ways to represent the same tree.

description of the algorithm is based on a different representation of the tree than is given in that paper. The LOCAL algorithm proposes changes to only small portions of the tree. Our experience is that a single algorithm is not sufficient for rapid mixing in all data sets.

A Tree Representation

At first, we shall consider the phylogeny to be a rooted binary tree. A binary tree with s leaves has $s - 1$ internal nodes including the root and may be drawn in 2^{s-1} equivalent ways, as there is an arbitrary decision to be made at each internal node on which subtree should be left and which should be right. For a given set of left/right choices there is a unique in-order traversal of the tree (e.g., Drozdek and Simon 1995, section 8.4). Each internal node is adjacent to two leaves in this traversal, the rightmost leaf of its left subtree and the leftmost leaf of its right subtree. Given an ordering of the nodes and the distances between adjacent nodes, the tree topology and branch lengths are uniquely determined.

If a molecular clock is assumed, the pair of distances to the adjacent leaves are equal for each internal node, and the branch lengths are all determined by $s - 1$ values. Otherwise, $2(s - 1)$ values are necessary.

This tree representation, a permutation of the s taxa and a sequence of $s - 1$ or $2(s - 1)$ distances, may be visualized in a graph of the distance from the root in a depth-first walk through the tree, as in figures 1 and 2. A variation of this representation appears in Aldous (1993) and Aldous and Larget (1992). Durbin et al. (1998, pp. 206–210) describe this representation and give it the name ''traversal profile.'' Each taxon appears at a peak in the graph, and each internal node is a valley. The permutation of taxa is read across the tops of the peaks and the branch lengths and tree topology are determined by the $s - 1$ valley depths in the molecular clock case, or the $2(s - 1)$ left and right valley depths without a molecular clock.

GLOBAL with a Molecular Clock

For GLOBAL with a molecular clock (Mau and Newton 1997; Mau, Newton, and Larget 1999), first, one representation of the current tree is selected uniformly at random by choosing the left/right orientation of the two subtrees with equal probability for each internal node. Second, the $s - 1$ valley depths are simultaneously and independently modified by adding to each a small perturbation uniformly chosen between $-\delta_\psi$ and δ_ψ, keeping the depth between 0 and a specified maximum. If a proposed change would take a valley depth out of range, the excess is reflected back into the required interval. The resultant tree is either accepted or rejected by the Metropolis-Hastings algorithm. The size of δ_ψ affects the mixing properties and needs to be carefully chosen. We begin with a large value and halve it when the acceptance rate is low during an initial burn-in period. Changing the proposal mechanism based on the history of the chain violates the Markov property and can lead to invalid inferences from long-run frequencies (see Gilks, Roberts, and Sahu 1998). We leave the value of δ_ψ fixed while sampling trees for inference.

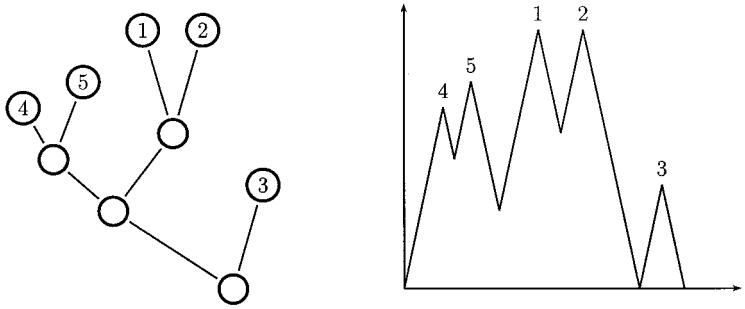

FIG. 2.—A tree and its representation without assuming the molecular clock. The permutation of five leaf labels and the eight left and right valley depths describe the tree completely. There are 2^4 equivalent representations for this rooted tree. A different rooting of the corresponding unrooted tree would be represented differently.

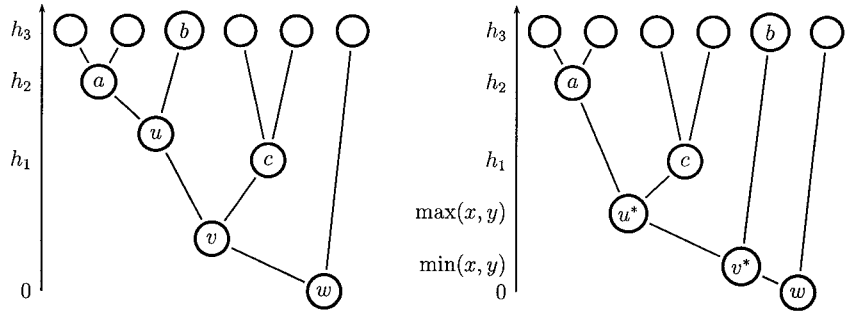

FIG. 3.—A tree before and after a local move which changes its tree topology. The labeled nodes are in the neighborhood of the randomly selected edge between u and v. Because $x < h_1$, there were three choices of tree topologies. A return move to the prior heights of u and v would result in a forced choice of tree topology, and is thus three times as likely, so the Hastings ratio is 3.

GLOBAL Without the Molecular Clock

For GLOBAL without the molecular clock, we perturb the $2(s - 1)$ left and right valley depths of the tree representation instead of the $s - 1$ valley depths as with the molecular clock. Because the likelihood models we use are reversible and do not distinguish between alternative rootings of the same unrooted tree, with a small probability we propose an alternative rooting (which is always accepted) in place of changing the tree.

LOCAL with a Molecular Clock

LOCAL modifies the tree only in a small neighborhood of a randomly chosen internal branch, leaving the remainder of the tree unchanged. We begin by choosing with equal probability one of the rooted tree's $s - 2$ internal edges (not joined to a leaf of the tree) from the current tree ψ. In figure 3, u and v are the nodes joined by the randomly chosen edge, and the other nodes involved in the proposal are labeled. LOCAL with a molecular clock will only change the branch lengths connecting these labeled nodes and possibly introduce a new tree topology.

First, we consider the case in which v is not the root of the tree. Leaving a, b, c, and w fixed, we pick new positions for nodes u and v. Let dist$(\cdot, \cdot)$ be the within-tree distance between any two nodes. Let h_1, h_2, and h_3 be the three distances dist(a, w), dist(b, w), and dist(c, w) in sorted order with $h_1 < h_2 < h_3$. LOCAL chooses x uniformly at random from the interval $[0, h_2]$ and y uniformly at random from $[0, h_1]$. Proposed nodes u^* and v^* will be distances $\max(x, y)$ and $\min(x, y)$ from w, respectively. If $\max(x, y) < h_1$, there are three possible tree topologies. One of the three children, a, b, and c, is randomly chosen to be joined to v^*, with the others becoming children of u^*. On the other hand, if $\max(x, y) > h_1$, the tree topology is forced, and the child node with the smallest height becomes a child of v^*. The Hastings ratio for this proposal is either 1/3, 1, or 3: If dist$(u, v) >$ dist(c, v) in the current tree and $\max(x, y) < h_1$, the Hastings ratio is 3; if dist$(u, v) <$ dist(c, v) in the current tree and $\max(x, y) > h_1$, the Hastings ratio is 1/3; otherwise, it is 1. When v is not the root, the overall height of the tree is unchanged.

In the second case, v is the root of the tree, and there is no node w. LOCAL randomly changes the distances between v and the children a, b, and c and chooses a new location for u. Let h_1, h_2, and h_3 be the three distances dist(a, v), dist(b, v), and dist(c, v) in sorted order with $h_1 < h_2 < h_3$. We let $h_1^* = h_1 \times e\lambda_1^{(U-0.5)}$, where U is a uniform(0, 1) random variable, and λ_1 is a tuning parameter. We let $h_i^* = h_i + h_1^* - h_1$ for $i = 2, 3$ be the proposed distances of a, b, and c to the proposed root v^*. We then place u^* at a height x above v^*, chosen uniformly at random between 0 and h_2^*. The relative sizes of x and h_1^* determine whether the tree topology is forced or randomly chosen from three possibilities as above. The Hastings ratio is $r \times (h_1^*/h_1)$, where r is 1/3, 1, or 3, as above. LOCAL with a molecular clock is very similar in character to the method for rearranging trees in Kuhner, Yamato, and Felsenstein (1995). Their method differs from ours in the manner in which new branching points are proposed.

LOCAL Without the Molecular Clock

For LOCAL without the molecular clock, we randomly pick one of the $s - 3$ internal edges of the unrooted tree, designating its two nodes u and v. The other two neighbors of u are randomly labeled a and b, and v's two other neighbors are randomly labeled c and d with equal probability. Set $m =$ dist(a, c). Our proposal changes m by multiplying edge lengths on the path from a to c by a random factor. We then detach either u or v with equal probability and reattach it along with its unchanged subtree to a point chosen uniformly at random on the path from a to c. Specifically, $m^* = m \times e\lambda_2(U_1^{-0.5})$, where U_1 is a uniform(0, 1) random variable and λ_2 is a tuning parameter. Let $x =$ dist(a, u) and $y =$ dist(a, v) be distances in the current tree. If u is chosen to move, the proposal sets $x^* = U_2 \times m^*$ and $y^* = y \times m^*/m$. If v is chosen to move, $x^* = x \times m^*/m$ and $y^* = U_2 \times m^*$. In both cases, U_2 is a uniform(0, 1) random variable. If $x^* < y^*$, the tree topology does not change while dist$(a, u^*) = x^*$, dist$(u^*, v^*) = y^* - x^*$, and dist$(v^*, c) = m^* - y^*$. If $x^* > y^*$, the tree topology does change as u^* becomes a neighbor of c and v^* becomes a neighbor of a while dist$(a, v^*) = y^*$, dist$(v^*, u^*) = x^* - y^*$, and dist$(u^*, c) = m^* - x^*$. The Hastings ratio in this case is $(m^*/m)^2$. An example using this proposal mechanism is shown in figure 4.

MCMC Algorithms for Updating Parameters

When a parameter, such as κ in the HKY85 model (Hasegawa, Kishino, and Yano, 1985), is restricted in a

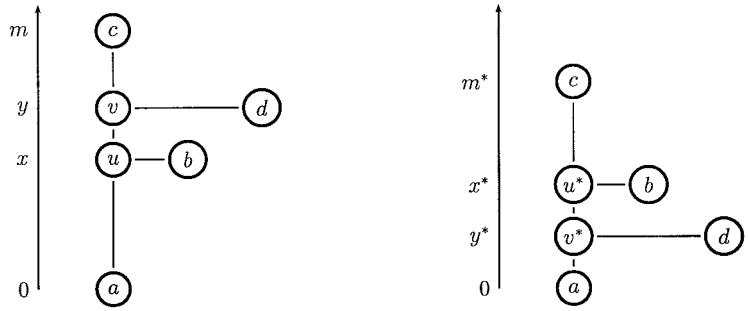

FIG. 4.—A tree before and after a non-molecular-clock local move which changes its tree topology. The distance from *a* to *c* was modified by a randomly chosen factor. The subtrees extending from *u** and *v** through *b* and *d*, respectively, did not change. The node *v* was randomly chosen to move, while the relative position of *u* on the path from *a* to *c* did not change.

range [0, *M*], we propose a new value $\kappa^* = \kappa + U$, where *U* is chosen uniformly at random between $-\delta_\kappa$ and δ_κ, reflecting the excess back into the range should κ^* be negative or exceed *M*. The Hastings ratio is 1 for this proposal distribution.

When a set of parameters is constrained to sum to a constant, we propose a new set of values according to a Dirichlet distribution centered at the current parameter values. Specifically, if the current values are $z = (z_1, z_2, \ldots, z_k)$, where $\sum_i z_i = c$, we let $z^* = cY$, where *Y* is randomly chosen from a Dirichlet distribution with parameters $(\alpha z_1, \alpha z_2, \ldots, \alpha z_k)$, with α a tuning parameter. The higher α is, the more likely the proposed parameter values are to remain close to their current values. (Dirichlet random variables are generated by normalizing independent gamma random variables by their sum. See Johnson and Kotz [1972, chapter 40] for details.) The Hastings ratio for this proposal distribution is a ratio of two Dirichlet densities, and we do not report its complicated expression.

The entire parameter proposal chain is obtained by independently proposing new parameter values for all parameters with these two types of proposals and accepting or rejecting the combined proposal in a single Metropolis-Hastings step. It is critical for proper mixing that the tuning parameters δ_κ and α be chosen well for good acceptance rates.

Results and Discussion

Examples

A Primate Phylogeny

For our first example, we reanalyze the primate mitochondrial DNA sequences studied by Yang and Rannala (1997). The data set is distributed with the PAML (Phylogenetic Analysis by Maximum Likelihood, Yang 1997) package and represents segments of the mitochondrial genomes of nine primates. The data originally appeared in Hayasaka, Gojobori, and Horai (1988). The sequences each contain 888 sites, with segments from two protein-coding genes and three tRNA genes. In our analysis, we use F84 with a molecular clock, the model of nucleotide base substitution in the DNAMLK program in PHYLIP (Felsenstein 1995). The parameterization of the instantaneous rate matrix we use is

$$\theta\begin{bmatrix} \cdot & (1+\kappa/\pi_R)\pi_G & \pi_C & \pi_T \\ (1+\kappa/\pi_R)\pi_G & \cdot & \pi_C & \pi_T \\ \pi_A & \pi_G & \cdot & (1+\kappa/\pi_Y)\pi_T \\ \pi_A & \pi_G & (1+\kappa/\pi_Y)\pi_C & \cdot \end{bmatrix} \tag{4}$$

Instead of modeling all sites equally, we allow different parameter values for each of four different site categories. In the protein-coding regions, we have different parameters for each codon position. A fourth category is for the tRNA genes.

We assume a uniform prior $p(\psi)$ on all clocklike trees whose total height is less than 100. (The exact choice of this constant has no effect on the simulations, provided it is large enough.) Because each category requires six parameters in its instantaneous rate matrix, φ is a vector of 24 parameters. Its flat prior, independent of the prior on trees, is

$$p(\varphi) = p(\theta_1, \ldots, \theta_4) \prod_{i=1}^{4} p(\kappa_i) p(\pi_{i,A}, \pi_{i,G}, \pi_{i,C}, \pi_{i,T}), \tag{5}$$

where $p(\kappa_i)$ is the uniform density between 0 and 100 for each *i*. The density $p(\theta_1, \ldots, \theta_4)$ is determined so that $(w_1\theta_1, \ldots, w_4\theta_4)$ has a flat Dirichlet distribution where w_i is the proportion of all sites in category *i*. This constraint avoids confounding with branch lengths. We also assume that $p(\pi_{i,A}, \pi_{i,G}, \pi_{i,C}, \pi_{i,T})$ is a flat Dirichlet density for each category *i*. Because our prior $p(\varphi)$ is constant over the set of permissible parameter values, the ratio of posterior densities in equation (3) is simply the ratio of likelihoods.

Before our runs for inference, we conducted several short runs to find good initial values for model parameters and tuning parameters, and we report these values in the caption of figure 5. We completed four separate runs from randomly selected initial trees and obtained consistent results. Each run consisted of 2,000 cycles with GLOBAL with a molecular clock and no parameter updating during which δ_ψ was dynamically lowered to 0.00625. This was followed by 2,000 cycles with one parameter proposal and one tree proposal using LOCAL with a molecular clock to complete burn-in. We continued the same sequence of cycles, subsampling every

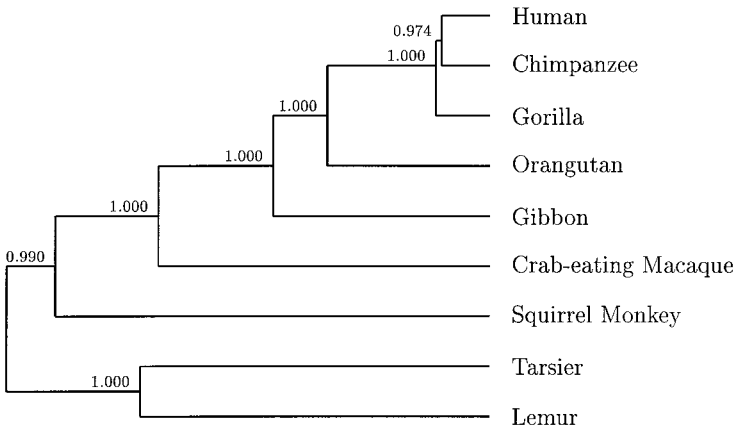

FIG. 5.—Primate tree topology with clade posterior probabilities. The number above and to the left of each internal node is the posterior probability that the taxa in the corresponding subtree form a clade in the true tree (although perhaps with a different tree topology than that shown) based on observed frequencies in 40,000 sampled trees from four combined runs. Analysis of the sample also yields posterior probabilities for each subtree topology, which we do not report. The branch lengths are approximately scaled to the means of their posterior estimates. Initial model parameters values were $\kappa = (1.16, 2.37, 8.07, 2.24)$, $\theta = (0.89, 0.29, 2.32, 0.39)$, with π values equal to observed frequencies in the data. Initial tuning parameter values were $\lambda_\theta = \lambda_\pi = 3{,}000$, $\lambda_1 = 2 \log 2$, $\delta_\psi = 0.2$, and $\delta_\kappa = 0.1$.

tenth tree from the next 100,000 cycles for inference. Examination of trace plots of the log-likelihood and the observed consistency between runs suggests that the burn-in period was sufficiently long. The log-likelihood for sampled trees in each run varied around a mean of −4,887.5. The combined samples yield a posterior probability of 0.964 for the best tree topology, with an estimated Monte Carlo standard error of 0.005 determined from the four independent samples. The tree topology is the same as that found in Yang and Rannala (1997). Their posterior probabilities for the same tree topology varied between 0.951 and 0.958 under different priors. The discrepancy in our values is caused primarily by the fact that we modeled categories of sites differently. Figure 5 summarizes the uncertainty in the tree topology. Our approach allows inferences on model parameters and tree branch lengths, including 95% credible regions (Bayesian confidence intervals) displayed in table 1 for the parameters in the model.

A Whale Phylogeny

The traditional whale phylogeny places one group of toothed whales (Odontoceti), the sperm whales (Physeteridae), as a sister group to dolphins and as more distantly related to baleen whales (Mysticeti). This traditional view has been challenged (Milinkovitch, Meyers, and Powell 1994; Milinkovich, Orŕa and Meyers 1995) and the assertion made that sperm whales and baleen whales were sister groups on the basis of analysis of molecular sequences. Another set of authors (Árnason, Gretarsdottir, and Gullberg 1993; Árnason and Gullberg 1994) found evidence for yet another tree topology with dolphins and baleen whales being the most closely related. Adachi and Hasegawa (1995) found that the choice of outgroup made a considerable difference in the bootstrap proportions in favor of each competing tree topology. Their analysis found support for the tree proposed by Milinkovitch, Meyers, and Powell (1994) and Milinkovitch, Orŕa and Meyers (1995). A Bayesian reanalysis of the cytochrome *b* data set of 1,140 aligned base pairs can attach a posterior probability to each of the three competing hypotheses.

The data set contains sequences from 14 species of whales (2 dolphins, 1 sperm whale, and 11 baleen whales) and 17 artiodactyles (hippopotamus, 6 camels, pig, peccary, cow, sheep, goat, black-tailed deer, giraffe,

Table 1
The Central 95% of the Posterior Distribution of Each Likelihood Model Parameter in Each of the Four Rate Categories, as Determined by the Four Combined MCMC Runs and the Observed Mean Values

	Coding Genes						tRNA Genes	
	First Position		Second Position		Third Position			
	95% C.R.[a]	Mean	95% C.R.	Mean	95% C.R.	Mean	95% C.R.	Mean
κ.....	(0.73, 1.64)	1.14	(1.94, 5.47)	3.36	(5,69, 15.45)	9.81	(1.48, 4.01)	2.50
θ.....	(0.81, 1.17)	0.98	(0.19, 0.40)	0.28	(2.04, 2.45)	2.25	(0.28, 0.53)	0.39
π_A ...	(0.33, 0.42)	0.37	(0.12, 0.20)	0.16	(0.33, 0.42)	0.37	(0.28, 0.39)	0.34
π_G ...	(0.10, 0.16)	0.13	(0.07, 0.13)	0.10	(0.04, 0.06)	0.04	(0.11, 0.18)	0.14
π_C ...	(0.23, 0.32)	0.27	(0.28, 0.37)	0.32	(0.37, 0.43)	0.40	(0.17, 0.26)	0.22
π_T....	(0.19, 0.26)	0.22	(0.37, 0.48)	0.42	(0.16, 0.20)	0.18	(0.25, 0.36)	0.31

NOTE.—Because we assume flat priors, the means are likely to be close to the maximum-likelihood estimates.
[a] C.R. = credible region.

fallow, pronghorn, and chevrotain). Using data from all 31 taxa, we used the likelihood model HKY85 (Hasegawa, Kishino, and Yano 1985) with different parameter values for each codon position and did not assume a molecular clock. Experimentation on short runs led us to choose these initial parameters: $\kappa = (6.0, 3.8, 10.0)$, $\theta = (0.29, 0.10, 2.61)$, π given by observed frequencies, $\delta_\psi = 0.2$ (dynamically reduced to about 0.0016), $\delta_\kappa = 0.2$, $\lambda_\theta = 2{,}000$, and $\lambda_\pi = 4{,}000$. Burn-in consisted of 5,000 cycles of GLOBAL with a molecular clock and no parameter updating followed by 10,000 cycles each with one parameter proposal and one tree proposal using LOCAL without a molecular clock. We continued the same cycle sequence and subsampled every tenth tree from the next 500,000 cycles for inference. We repeated the simulation four times with different random seeds, obtaining consistent results. Our inferences are based on the combined sample of 200,000 saved tree topologies.

We attach posterior probabilities to each of the three hypotheses by simply counting the number of tree topologies of each type in the sample. We find that the most probable hypothesis is the tree (dolphins, (sperm whales, baleen whales)) with posterior probability 99.3% (198,630/200,000). The traditional tree (baleen whales, (dolphins, sperm whales)) has posterior probability 0.6% (1,185/200,000), while the other alternative has posterior probability 0.1% (185/200,000). This analysis draws the same general conclusion as that drawn by Adachi and Hasegawa (1995). A host of other inferences could be drawn from our sample of 200,000 trees, but we refrain. The validity of the inferences depends on the validity of the likelihood model, prior distributions, and data.

The fact that out posterior probabilities are much more extreme than the bootstrap proportions in Adachi and Hasegawa (1995) is not caused by fundamental differences between the two approaches. Rather, it results from our use of more data to locate the root of the whale tree: we use 17 outgroup taxa in a single analysis instead of one or two.

In both examples, we used GLOBAL during burn-in and LOCAL for our inference runs. LOCAL is about three times as fast as GLOBAL per cycle on the data sets in this paper and seems to mix through tree topologies equally well. Were we interested in estimating branch lengths, GLOBAL may be a better choice, because it updates all branch lengths with each proposal.

Comparison with Maximum Likelihood and Bootstrapping

In this paper, we are interested in assessing uncertainty in a Bayesian framework. The alternative method most similar to this is maximum-likelihood estimation and bootstrapping (Felsenstein 1985). The application of maximum-likelihood has become increasingly widespread for trees without too many taxa. A practical concern is that maximum-likelihood estimation depends on heuristic optimization that is not guaranteed to converge to the true optimal trees.

The computational effort in bootstrapping the maximum-likelihood procedure on a problem with many taxa can be so onerous that practitioners may be inclined instead to choose an alternative method. One computationally tractible approximate likelihood-based approach is quartet puzzling (Strimmer and von Haeseler 1996), which provides some numerical means to assess uncertainty, although a theory which guides an objective interpretation of these numerical values has yet to be developed. In a similar vein, although bootstrapping has been proven to be a valid method of assessing uncertainty in a variety of situations (see Efron and Tibshirani 1993 for many examples), there is considerable debate over the proper way to interpret bootstrap proportions in a phylogeny reconstruction (Felsenstein and Kishino 1993; Berry and Gascuel 1996; and Newton 1996). From our perspective, an interesting interpretation is that of Efron, Halloran, and Holmes (1996). These authors conclude that the most reasonable interpretation of bootstrap proportions, as applied in Felsenstein (1985), are as Bayesian posterior probabilities with a uniform prior. While refuting the claim made by others of systematic bias in Felsenstein's application of the bootstrap, they argue that a two-stage bootstrap requiring at least 20 times the computational effort that Felsenstein's bootstrap requires is necessary for proper frequentist statistical inference.

There is a considerable difference in the time we needed for our analyses and the time required using the bootstrap and maximum likelihood to achieve similar standard errors in the estimated probabilities. In our analysis, each of the four runs required about 100 min of CPU time on a 300-MHz Pentium II PC operating under Solaris x86. The best tree topology with all 31 taxa (not shown) had a posterior probability of 0.413 with an estimated Monte Carlo standard error of 0.013 (determined from our four independent samples).

To estimate a proportion close to 0.413 to the same degree of accuracy requires about 1,400 independent observations. In some sense, this means that the information from our highly dependent sample of 200,000 trees is equivalent to about 1,400 independent draws from the posterior. Ignoring differences between bootstrap proportions and Bayesian posterior probabilities, the information in our sample is about as good as 1,400 bootstrap replicates. We achieved this in less than 7 h of CPU time on a desktop PC. In the process, we also obtained estimates and assessments of uncertainty of all parameters in the model.

In contrast, a single run of DNAML from the PHYLIP package with three rate categories required over 180 min of CPU time on the same computer. Assuming the genuine maximal tree is found in about 3 h, bootstrapping 1,400 times would require 175 days of CPU time, nearly half a year.

The difference in computational requirements is easily explained. Maximum likelihood requires extensive computation to produce a single tree. To assess uncertainty using the bootstrap, at least 100 equally challenging computational problems must be solved. Computation used in one optimization is not used in subsequent optimizations. In contrast, in a Bayesian analysis by Metropolis-Hastings MCMC, once burn-in has been

reached, every tree evaluation adds information useful for assessing uncertainty.

While many practitioners adhere to or avoid Bayesian methods on purely philosophical grounds, we advocate a Bayesian approach for phylogenetic inference because, unlike bootstrapping, it quantifies uncertainty in questions of biological interest in a directly interpretable manner (from a Bayesian perspective) and does this with much less computational effort. When deciding between three unnested hypotheses, such as the three candidate whale phylogenies, a Bayesian analysis attaches a posterior probability to each. Classical hypothesis testing, in contrast, is awkward to apply when there is no natural null model.

Comparison with the Computational Approach of Yang and Rannala

We reanalyzed the primate data using the program `mcmctree` in PAML to make efficiency comparisons between our two computational approaches. The results of these comparisons are striking. Running our code on the computer described above, we found that each run required 9 min of CPU time, for a total of 36 min of CPU time, to obtain our estimates of the tree topology and the parameters. We ran `mcmctree` on the same computer using the empirical Bayes option (which is faster than the hierarchical option) with the same initial parameter estimates described in Yang and Rannala (1997) and $\delta_1 = 0.01$ to obtain Monte Carlo standard errors twice as large as those we obtained. The program required 10.9 h of CPU time to complete. To halve the standard errors would have required more time.

While there are minor differences between the Bayesian models we use and those used by Yang and Rannala (1997), our methods of Metropolis-Hastings sampling are quite different. The computational method described in Yang and Rannala (1997) uses MCMC as a means to generate a collection of labeled histories which will subsequently be evaluated individually by Monte Carlo integration. Because they evaluate the posterior probability ratios of labeled histories to only limited precision during the MCMC portion of their computational approach, the sampled labeled histories are not a valid sample from their posterior, and the advantages of a sample-based approach to Bayesian inference mentioned earlier are not available. The Monte Carlo integration estimates of posterior labeled history probabilities found by Yang and Rannala are accurate and valid. Their approach is simply substantially more expensive computationally than the method we present here, at least for this example.

In our approach, the state space includes the tree topology, branch lengths, and model parameters. The calculation of an acceptance probability of a proposed tree sums over the unknown data at the internal nodes, a process that is rapid and accurate with the pruning algorithm (Felsenstein 1983). In contrast, determining the relative posterior probabilities between two labeled histories requires integrating over all possible branch lengths in addition to summing over unknown data at the internal nodes of each tree. This is a highly expensive computation to perform with great accuracy.

Extending Yang and Rannala's Approach to the Larger Example

We could not apply Yang and Rannala's (1997) approach to the whale/artiodactyl data. Although there is no inherent reason why their method cannot work on larger trees, the distributed code is restricted to trees with 10 or fewer taxa. This artificial restriction could presumably be removed. A second complication was our use of a model without the molecular clock. Because the MCMC algorithm of Yang and Rannala operates on labeled histories which are not defined for unrooted trees, some algorithmic change would be necessary to handle this model change. Inference using Yang and Rannala's computational approach to ascertain the posterior probabilities of the 175 tree topologies in the 99% credible region for the tree topology found in our analysis would require at least 175 separate 30-dimensional Monte Carlo integrations, a considerably more difficult computation than that for the 14 separate 8-dimensional Monte Carlo integrations their code computed for the primate data set. Assuming their MCMC algorithm could find all tree topologies with nonnegligible posterior probabilities, it is also unknown how much computational time would be required to achieve the accuracy we demonstrate running our algorithms for less than 7 h.

Conclusions

We have demonstrated that a Bayesian approach to phylogenetic inference has substantial advantages over the approach of maximum likelihood and bootstrapping for large trees and that the Metropolis-Hastings algorithms we introduce are superior to other published computational algorithms, at least on the small number of data sets we have considered. We do, however, mention several issues of which a user of these methods ought to be aware.

Pitfalls of MCMC

Just as there is reason to question whether a maximum-likelihood algorithm truly finds a global maximum, there is reason to question whether an MCMC algorithm correctly identifies and measures the posterior probabilities of the collection of highly probable tree topologies. A common difficulty is for a particular simulation run to get stuck in one region of the parameter space and fail to visit other regions where the posterior is of comparable size or even higher, leading to substantial bias in inference. A related difficulty appears when transitions between islands of high posterior probability occur at very low rates. Without providing an ironclad guarantee that these potential pitfalls have been avoided, obtaining consistent results from several repeated long runs from randomly chosen disparate initial trees is a minimal criterion for reliable results of the computation. A more stringent test is to check if the true tree topology is captured with substantial posterior probability on simulated data. Mau, Newton, and Larget (1999) conduct a simulation study with a 32-taxon tree

using GLOBAL with a molecular clock that successfully passes this test. However, the algorithms in this paper have failed to give consistent results when applied to simulated data from a tree with 64 taxa (data not shown).

Model Misfit

The standard errors we report in this paper measure the likely disparity between the theoretical probabilities expressed in equation (2) and the numerical values our simulations provide. They are calculated assuming that the underlying likelihood model is a sufficiently good explanation for the observed data. In this paper, we have modeled obvious data features by allowing different parameter values for different codon positions, for example. Still, data sequences generated by our best fitted model would likely differ considerably from genuine data regarding composition of amino acids, locations of stop codons, and other biologically relevant features. More effort is needed to incorporate more biological understanding into the likelihood models in common use. Users of MCMC in phylogentic inference ought to be concerned more about proper modeling than about computational issues in the simulations.

A Promising Future

Despite these caveats, we are quite confident that the algorithms we have developed represent a major step forward in likelihood-based phylogenetic analysis. We reemphasize that the tree proposal algorithms presented in this paper interact with the likelihood model solely through the acceptance probabilities; they require no modification should a Bayesian approach be extended to problems with different types of data or likelihood models.

We have developed the software package Bayesian Analysis in Molecular Biology and Evolution (BAMBE, Simon and Larget 1998), written in ANSI C, which is publicly available on our Web site (http://www.mathcs.duq.edu/larget/bambe.html). Windows and Macintosh versions are under development.

Acknowledgments

Support for this research has been provided by NSF grant DBI-9723799. The authors thank Masami Hasegawa for providing the data used in the whale example. We used algorithms for generating random variables from software libraries at StatLib (http://www.stat.cmu.edu/) and NetLib (http://www.netlib.no/). A discussion with Jeff Thorne led to small changes in some of our MCMC algorithms. Michael Newton provided advice on implementing the Dirichlet distribution proposal mechanism and made welcome editorial comments. We thank Bob Mau, John Huelsenbeck, and two anonymous referees, whose comments led to an improved structure of and greater accuracy in the article. The NSF supported B.L.'s time at the Isaac Newton Institute for Mathematical Sciences at Cambridge University during preparation of the initial submission.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1995. Phylogeny of whales: dependence of the inference on species sampling. Mol. Biol. Evol. **12**:177–179.

ALDOUS, D. 1993. The continuum random tree III. Ann. Prob. **21**:248–289.

ALDOUS, D., and B. LARGET. 1992. A tree-based scaling exponent for random cluster models. J. Phys. A Math. Gen. **25**:L1065–L1069.

ÁRNASON, U., S. GRETARSDOTTIR, and A. GULLBERG. 1993. Comparison between the 12S rRNA, 16S rRNA, NADH1 and COI genes of sperm and fin whale mitochondrial DNA. Biochem. Syst. Ecol. **21**:115–122.

ÁRNASON, U., and A. GULLBERG. 1994. Relationship of baleen whales established by cytochrome b gene sequence comparison. Nature **367**:726–728.

BERRY, V., and O. GASCUEL. 1996. On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. Mol. Biol. Evol. **13**:999–1011.

DROZDEK, A., and D. SIMON. 1995. Data structures in C. PWS publishing company, Boston.

DURBIN, R., S. EDDY, A. KROGH, and G. MITCHESON. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, England.

EFRON, B., E. HALLORAN, and S. HOLMES. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. **93**:13429–13434.

EFRON, B., and R. J. TIBSHIRANI. 1993. An introduction to the bootstrap. Chapman and Hall, London.

FELSENSTEIN, J. 1983. Statistical inference of phylogenies (with discussion). J. R. Stat. Soc. A **146**:246–272.

———. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

———. 1995. PHYLIP (phylogeny inference package). Version 3.572c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

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

GELMAN, A., J. CARLIN, H. STERN, and D. RUBIN. 1995. Bayesian data analysis. Chapman and Hall, London.

GILKS, W., G. ROBERTS, and S. SAHU. 1998. Adaptive Markov chain Monte Carlo through regeneration. J. Am. Stat. Assoc. **93**:1045–1054.

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

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

HAYASAKA, K., T. GOJOBORI, and S. HORAI. 1988. Molecular phylogeny and evolution of primate mitochondrial DNA. Mol. Biol. Evol. **5**:626–644.

JOHNSON, N., and S. KOTZ. 1972. Distributions in statistics. Continuous multivariate distributions. Wiley, New York.

KUHNER, M., J. YAMATO, and J. FELSENSTEIN. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140**:1421–1430.

LI, S. 1996. Phylogenetic tree construction using Markov chain Monte Carlo. Ph.D. dissertation, Ohio State University, Columbus.

MAU, B. 1996. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Ph.D. dissertation, University of Wisconsin, Madison.

MAU, B., and M. A. NEWTON. 1997. Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. J. Comp. Graph. Stat. **6**:122–131.

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

METROPOLIS, N., A. ROSENBLUTH, M. ROSENBLUTH, A. TELLER, and E. TELLER. 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. **21**:1087–1092.

MILINKOVITCH, M. C., A. MEYERS, and J. R. POWELL. 1994. Phylogeny of all major groups of cetaceans based on DNA sequences from three mitochondrial genes. Mol. Biol. Evol. **11**:939–948.

MILINKOVITCH, M. C., G. ORTÍ, and A. MEYERS. 1995. Novel phylogeny of whales revisited but not revised. Mol. Biol. Evol. **12**:518–520.

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. *In* F. SEILLIER-MOSEIWITCH, T. P. SPEED, and M. WATERMAN, eds. Statistics in molecular biology. Monograph Series of the Institute of Mathematical Statistics (in press).

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

SIMON, D., and B. LARGET. 1998. Bayesian analysis in molecular biology and evolution (BAMBE). Version 1.01 beta. Department of Mathematics and Computer Science, Duquesne University, Pittsburgh, Pa.

SMITH, A., and G. ROBERTS. 1993. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J. R. Stat. Soc. B **55**:3–23.

STRIMMER, K., and A. VON HAESELER. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

TIERNEY, L. 1994. Markov chains for exploring posterior distributions. Ann. Stat. **22**:1701–1762.

YANG, Z. 1997. Phylogenetic analysis by maximum likelihood (PAML). Version 1.3. Department of Integrative Biology, University of California at Berkeley.

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

MIKE HENDY, reviewing editor

Accepted February 15, 1999