

# Bayesian Phylogenetic Inference Using DNA Sequences: A Markov Chain Monte Carlo Method

Ziheng Yang and Bruce Rannala

Department of Integrative Biology, University of California, Berkeley

An improved Bayesian method is presented for estimating phylogenetic trees using DNA sequence data. The birth–death process with species sampling is used to specify the prior distribution of phylogenies and ancestral speciation times, and the posterior probabilities of phylogenies are used to estimate the maximum posterior probability (MAP) tree. Monte Carlo integration is used to integrate over the ancestral speciation times for particular trees. A Markov Chain Monte Carlo method is used to generate the set of trees with the highest posterior probabilities. Methods are described for an empirical Bayesian analysis, in which estimates of the speciation and extinction rates are used in calculating the posterior probabilities, and a hierarchical Bayesian analysis, in which these parameters are removed from the model by an additional integration. The Markov Chain Monte Carlo method avoids the requirement of our earlier method for calculating MAP trees to sum over all possible topologies (which limited the number of taxa in an analysis to about five). The methods are applied to analyze DNA sequences for nine species of primates, and the MAP tree, which is identical to a maximum-likelihood estimate of topology, has a probability of approximately 95%.

## Introduction

In an earlier paper, we proposed a Bayesian method for estimating phylogenetic trees (Rannala and Yang 1996) as an alternative to maximum likelihood (Felsenstein 1981). The method was an extension of the earlier work of Edwards (1970) on the problem of estimating phylogeny using gene frequency data from human populations. A birth–death process (Feller 1939; Kendall 1949) was used to specify the prior distribution of phylogenetic trees and ancestral speciation times, and a Markov process was used to model nucleotide substitution (Rannala and Yang 1996). The parameters of the birth–death process and the substitution model were estimated using maximum likelihood. These estimates were then used in place of the true parameters to evaluate the posterior probabilities of trees (a procedure known as empirical Bayesian analysis). For two sets of DNA sequences from several primate species that were analyzed, the Bayesian method generated the same best trees as were obtained by maximum-likelihood analyses, but the posterior probabilities for these trees were quite different from their bootstrap proportions and appeared to be less conservative (Rannala and Yang 1996). The method described in our earlier paper is only practical for analyzing data sets for a small number of species, as the calculations involve a sum over all tree topologies and the number of topologies increases rapidly with an increase in the number of species. As well, for each tree topology, a multi-dimensional integral over the ancestral speciation times is evaluated using numerical integration, and this calculation is not practical for more than about five species.

In this paper, we describe some refinements of the theory developed in our earlier paper that make the method practical for analyses of larger numbers of spe-

cies. We use Monte Carlo integration to evaluate more efficiently the integral over the ancestral speciation times for a given tree, and we avoid the need to sum over all topologies by evaluating the posterior probabilities of trees by using a Markov Chain Monte Carlo method. The model for the prior distribution of trees and speciation times has also been improved in two respects. First, species sampling by biologists is now considered. The birth–death process generates coalescent trees with internal branches longer, on average, than external branches. Taking species sampling into account reduces the internal branch lengths and results in a more realistic prior distribution of trees. Second, the birth and death rates of the prior distribution are treated as random variables and eliminated by integration. This approach, known as hierarchical Bayesian analysis (see Robert 1994), is expected to make the posterior probabilities more robust to violations of assumptions about the birth–death prior.

## Models and Estimation Theory

### The Data

Let  $s$  be the number of sequences (species) examined and  $n$  be the number of nucleotides in each sequence; insertions and deletions are ignored, and it is assumed that the sequences are aligned with gaps removed. We allow for species sampling so that  $S$  is the number of extant species sharing a most recent common ancestor (MRCA) and  $p = s/S$  is the fraction of these that are included in the study; the species included are assumed to be a random sample. The data can be represented as an  $s \times n$  matrix,  $\mathbf{X} = \{x_{ij}\}$ , where  $x_{ij}$  is the nucleotide at the  $j$ th site in the  $i$ th sequence. The  $j$ th column of the data matrix,  $\mathbf{x}_j = \{x_{1j}, \dots, x_{sj}\}'$ , will be the nucleotides among sequences at the  $j$ th site.

### The Labeled History

The sequences are descended through  $s - 1$  speciation events, which occurred at times  $t_1 > t_2 > \dots > t_{s-1}$  in the past (see fig. 1); we let  $\mathbf{t} = \{t_2, \dots, t_{s-1}\}$ . The time of the first bifurcation is set to one (i.e.,  $t_1 =$

Key words: molecular phylogeny, Bayesian estimation, Markov Chain Monte Carlo, nucleotide substitution, birth–death process.

Address for correspondence and reprints: Bruce Rannala, Department of Integrative Biology, University of California, Berkeley, California 94720-3140. E-mail: bruce@mws4.biol.berkeley.edu.

*Mol. Biol. Evol.* 14(7):717–724. 1997

© 1997 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

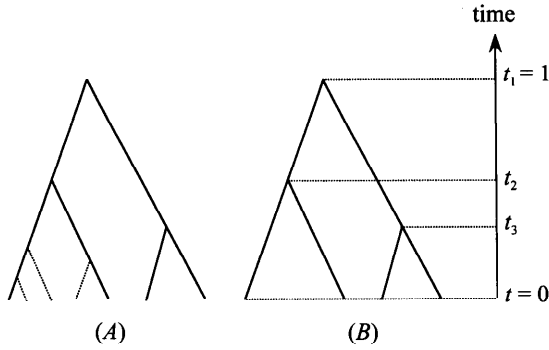


FIG. 1.—A labeled history of four species (B) sampled from a total of seven extant species (whose relationship is shown in A). Species that are not sampled are represented as dotted lines, and those that are sampled are represented as solid lines.

1), and parameters are then relative to this time scale (Edwards 1970). The relationship of the species is represented as a labeled history (denoted as  $\tau$ ), which is a tree of branching events (topology) with the nodes (ancestral speciation events) rank-ordered in time. A labeled history for  $s = 4$  sequences sampled from a total of  $S = 7$  extant species is shown in figure 1. A method was devised that assigns a unique integer index number to each distinct labeled history (see appendix A).

#### Speciation, Extinction, and Species Sampling

A linear birth–death process is used to model the dynamics of speciation and extinction and to specify the prior distribution of labeled histories and node times. The probability that a speciation event occurs in a particular lineage during an infinitesimal time interval  $\Delta t$  is  $\lambda \Delta t$ , the probability of an extinction event is  $\mu \Delta t$ , and the probability of two or more events is of order  $o(\Delta t)$ . The number of present-day species,  $S$ , is a random variable, and a subset,  $s$ , of the species is sampled so that each is included with probability  $\rho = s/S$ . Species sampling may be modeled as a mass-extinction event that occurs exactly at time present, with  $\rho$  being the probability that any particular species survives the extinction event (Nee, May, and Harvey 1994). The theory of generalized birth–death processes (Kendall 1948; Nee, May, and Harvey 1994) can be used to derive the probability that a lineage arising at time  $t$  in the past leaves one or more descendants in a present-day sample (using the notation of Nee, May, and Harvey [1994]) as

$$P(0, t) = \frac{\rho(\lambda - \mu)}{\rho\lambda + (\lambda(1 - \rho) - \mu)e^{(\mu - \lambda)t}}. \quad (1)$$

The probability that a lineage arising at time  $t$  in the past leaves exactly one descendant in the sample is

$$p_1(t) = \frac{1}{\rho} P(0, t)^2 e^{(\mu - \lambda)t}. \quad (2)$$

Using these results, the joint distribution of the node times ( $\mathbf{t}$ ), conditioned on  $t_1$ , may be obtained as

$$f(\mathbf{t} | s, t_1; \lambda, \mu) = (s - 2)! \prod_{j=2}^{s-1} \frac{\lambda p_1(t_j)}{v_{t_1}}, \quad (3)$$

where

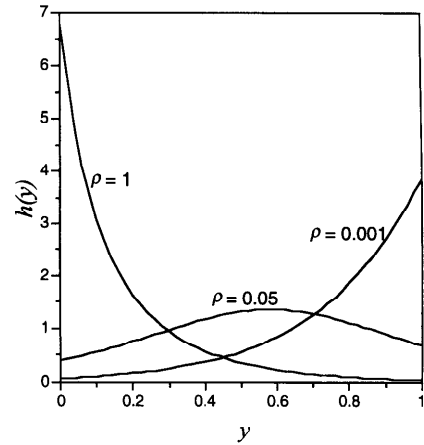


FIG. 2.—The kernel distribution,  $h(y)$  of equation (14), of the node times for different sampling proportions ( $\rho$ ). The node times ( $t_2, \dots, t_{s-1}$ ) represent the order statistics of  $(s - 2)$  independent random variables generated from this distribution. The birth and death rates are fixed at  $\lambda = 6.7$  and  $\mu = 2.5$ , which are the estimates of  $\lambda$  and  $\mu$  obtained for the primate mitochondrial DNA data with  $\rho = 9/150$  (see text). It is clear that increasing the sampling proportion decreases the expectation of the distribution of the node times and generates trees with longer internal branches.

$$v_{t_1} = 1 - \frac{1}{\rho} P(0, t_1) e^{(\mu - \lambda)t_1}. \quad (4)$$

The joint density of a particular labeled history,  $\tau$ , and set of speciation times,  $\mathbf{t}$ , is then

$$f(\tau, \mathbf{t} | s, t_1; \lambda, \mu) = f(\mathbf{t} | s, t_1; \lambda, \mu) f(\tau) \quad (5)$$

$$= \frac{2^{s-1}}{s!(s - 1)!} \prod_{j=2}^{s-1} \frac{\lambda p_1(t_j)}{v_{t_1}}. \quad (6)$$

where  $f(\tau) = 2^{s-1}/s!(s - 1)!$  is the probability associated with any particular labeled history (i.e., there are  $\xi(s) = s!(s - 1)!/2^{s-1}$  possible labeled histories, each having an equal probability, for a birth–death process). The limiting distribution when  $\lambda = \mu$  is

$$f(\tau, \mathbf{t} | s, t_1; \lambda, \mu) = \frac{2^{s-1}}{s!(s - 1)!} \prod_{j=2}^{s-1} \frac{1 + \rho\mu}{(1 + \rho\mu t_j)^2}. \quad (7)$$

If  $\rho = 1$ , equation (6) reduces to the density for a birth–death process with complete sampling (see eq. 8 of Rannala and Yang 1996). If  $\rho = 1$  and  $\mu = 0$ , the equation further reduces to the result for a Yule pure-birth process with complete sampling (see Edwards 1970). The effect of species sampling on the distribution of the node times under a birth–death process is shown in figure 2.

#### Model of Nucleotide Substitution

A continuous-time Markov process is used to model nucleotide substitution. The model used in J. Felsenstein's DNAML program (since 1984, PHYLIP version 2.6) will be used in this paper, although other substitution models are applicable as well (Rannala and Yang 1996). This model allows for different equilibrium nucleotide frequencies and transition/transversion rate bias. A parameter  $\kappa$  is used as the transition/transversion rate ratio, with  $\kappa = 0$  indicating no rate bias. The substitution rate matrix of the model is given in Kishino and Hase-

gawa (1989) and Rannala and Yang (1996). The parameter  $m$  is the substitution rate per site, measured by the number of substitutions per site from the root of the tree to the present. A molecular clock (rate constancy among lineages) is assumed in this paper, although that assumption may be relaxed. Substitutions are assumed to occur independently at different nucleotide sites; the conditional probability of observing the sequence data, given the labeled history ( $\tau$ ) and the node times ( $\mathbf{t}$ ), is then a product over sites

$$f(\mathbf{X}|\tau, \mathbf{t}; m, \kappa) = \prod_{j=1}^n f(\mathbf{x}_j|\tau, \mathbf{t}; m, \kappa), \quad (8)$$

where  $f(\mathbf{x}_j|\tau, \mathbf{t}; m, \kappa)$  is the conditional probability of observing the nucleotides at the  $j$ th site. The exact form of equation (8) depends on the tree topology (see, e.g., Felsenstein 1981).

### Hierarchical Bayesian Analysis

Two general approaches may be used to generate the posterior distribution when unknown parameters occur in the prior density: empirical Bayesian analysis and hierarchical Bayesian analysis (Berger 1985). Empirical Bayesian analysis replaces the unknown parameters with estimates; in our previous paper (Rannala and Yang 1996), maximum-likelihood estimates of the speciation and extinction rates of the birth–death prior were used. Hierarchical Bayesian analysis assigns second-level priors as densities for the unknown parameters of the prior. An integration is performed over the second-level priors to obtain a new prior that is completely specified. The posterior density is then generated in the usual manner. The potential advantages of hierarchical Bayesian analysis, especially with respect to the robustness of the posterior densities to the form of the prior, are discussed in Berger (1985) and Robert (1994).

In this paper, we use hierarchical Bayesian analysis to estimate the posterior distribution of phylogenetic trees. The speciation and extinction rates are generally unknown and may be assigned the prior densities  $f_\lambda(\lambda)$  and  $f_\mu(\mu)$ . The marginal prior density of  $\mathbf{t}$  is then

$$f(\mathbf{t}) = \int_0^\infty \int_0^\infty f(\mathbf{t}|\lambda, \mu) f_\lambda(\lambda) f_\mu(\mu) d\lambda d\mu. \quad (9)$$

The integral of equation (9) will usually have no analytical solution but may be evaluated using numerical methods (see below). The parameters  $\lambda$  and  $\mu$  are assigned uniform densities with the range parameters estimated (see below). The substitution rate ( $m$ ) and transition/transversion rate ratio ( $\kappa$ ) can often be estimated reliably using conventional maximum-likelihood methods, and these estimates appear to vary little among candidate trees. In this study, we use maximum-likelihood estimates of these parameters obtained for the maximum-likelihood tree, although these might also be assigned hierarchical priors. The parameters defining the equilibrium nucleotide frequencies are estimated using the observed frequencies.

### Posterior Distribution of Phylogenetic Trees

The posterior probability of the labeled history,  $\tau$ , conditional on the observed sequence data, can be calculated as

$$\begin{aligned} f(\tau|\mathbf{X}) &= \frac{f(\mathbf{X}, \tau)}{f(\mathbf{X})} \\ &= \frac{f(\mathbf{X}|\tau)f(\tau)}{f(\mathbf{X})}, \end{aligned} \quad (10)$$

where

$$f(\mathbf{X}|\tau) = \int_{t_2=0}^1 \cdots \int_{t_{s-1}=0}^{t_{s-2}} f(\mathbf{X}|\tau, \mathbf{t}) f(\mathbf{t}) dt_{s-1} \cdots dt_2, \quad (11)$$

and

$$f(\mathbf{X}) = \sum_{\tau} f(\mathbf{X}, \tau). \quad (12)$$

The conditional probability  $f(\mathbf{X}|\tau, \mathbf{t})$  is specified by the nucleotide substitution model (eq. 8), while the prior distribution of the node times,  $f(\mathbf{t})$ , is specified by the birth–death process (eq. 3 or 9).

In our previous paper (Rannala and Yang 1996), we used numerical integration to evaluate the integral of equation (11) over the random variables  $\mathbf{t}$ . This calculation is only practical for samples of about five or fewer sequences. We show below that the integral is more efficiently evaluated, for larger numbers of species, using Monte Carlo integration. The sum of equation (12) involves  $s!(s-1)!/2^{s-1}$  terms which will usually be too many to allow exact calculations. In this paper, we develop a Markov Chain Monte Carlo method (see Smith and Roberts 1993) for generating the posterior distribution of trees without explicitly evaluating equation (12).

### Monte Carlo Integration

Monte Carlo methods are used to integrate over the random variables  $\mathbf{t}$  as well as the parameters of the prior. Monte Carlo integration is increasingly efficient, by comparison with other methods of numerical integration, as the dimension of the integral (equal to  $s-2$  in our problem) increases (Fishman 1996, pp. 64–69). The integral of equation (11) can be approximated using the Monte Carlo estimator

$$f(\mathbf{X}|\tau) \approx \frac{1}{R} \sum_{j=1}^R f(\mathbf{X}|\tau, \tilde{\mathbf{t}}), \quad (13)$$

where  $R$  is the number of simulated replicates, and for each replicate, pseudorandom variables  $\tilde{\lambda}$  and  $\tilde{\mu}$  are generated from the prior densities  $f_\lambda(\lambda)$  and  $f_\mu(\mu)$ , and a vector of pseudorandom variables  $\tilde{\mathbf{t}}$  is generated from the density  $f(\mathbf{t}|\tilde{\lambda}, \tilde{\mu})$ . The Monte Carlo estimate of the density  $f(\mathbf{X}|\tau)$  is unbiased and consistent (Fishman 1996).

It is straightforward to simulate from the joint density  $f(\mathbf{t}|\lambda, \mu)$  by taking advantage of the property that this is equivalent to the density of the order statistics of  $s-2$  independent and identically distributed random

variables (see Rannala 1997) with common density (see eq. 3)

$$h(y) = \frac{\lambda p_1(y)}{v_1}. \quad (14)$$

The procedure is to generate a set of  $s - 2$  pseudorandom variables  $y_2, y_3, \dots, y_{s-1}$  from the density  $h(y)$  and order these so that  $y_{(s-1)} < y_{(s-2)} < \dots < y_{(2)}$ . The node times are then assigned using the relation  $\hat{t}_i = y_{(i)}$  for  $i = 2, \dots, s - 1$ . Pseudorandom variables with density  $h(y)$  may be generated using the inverse transformation method according to the following procedure: (1) generate a uniform (0, 1) random variable  $U$ ; (2) obtain an observation from  $h(y)$  using the following transformation for  $\lambda \neq \mu$ :

$$y = \frac{\log\{\phi - U\rho\lambda\} - \log\{\phi - U\rho\lambda + U(\lambda - \mu)\}}{\mu - \lambda}, \quad (15)$$

where

$$\phi = \frac{\rho\lambda(e^{(\mu-\lambda)} - 1) + (\mu - \lambda)e^{(\mu-\lambda)}}{e^{(\mu-\lambda)} - 1}. \quad (16)$$

For the case  $\lambda = \mu$ , we instead use the transformation:

$$y = \frac{U}{1 + \lambda\rho(1 - U)}. \quad (17)$$

Because  $f(\mathbf{X}|\tau, \mathbf{t})$  is often very small, its logarithm was calculated instead. Scaling factors were used to avoid overflows (or underflows) in the sum of equation (13), as  $f(\mathbf{X}|\tau, \mathbf{t})$  varies widely among replicates. The standard error  $\sigma_p$  of the estimate of  $p = f(\mathbf{X}|\tau)$  was calculated, and the accuracy of the Monte Carlo integration was fixed in advance by terminating the replicates when  $\sigma_p/p < \delta$ , where  $\delta$  is a prespecified value.

### Markov Chain Monte Carlo

To evaluate the posterior distribution of phylogenetic trees, we used a Markov Chain Monte Carlo (MCMC) method (see Hastings 1970; Smith and Roberts 1993; Fishman 1996). MCMC methods are useful for generating a probability distribution  $\pi = \{\pi_i\}$ ,  $i = 0, 1, \dots$ , when  $\pi_i$  is not easily calculated directly, but a function of the  $\pi$  such as  $\pi_i/\pi_j$  may be calculated directly. A simple algorithm for this purpose is the Metropolis-Hastings algorithm (Hastings 1970). There are two components to this algorithm: (1) a potential transition (from state  $i$  to state  $j$ ) is chosen using a nominating transition probability function  $q_{ij}$ ; (2) the chain moves to state  $j$  with probability  $\alpha_{ij}$  and remains in  $i$  with probability  $1 - \alpha_{ij}$ . The transition probabilities of the chain are then

$$p_{ij} = \begin{cases} q_{ij}\alpha_{ij}, & \text{if } i \neq j \\ 1 - \sum_l q_{il}\alpha_{il}, & \text{if } i = j. \end{cases} \quad (18)$$

The  $\alpha_{ij}$  are chosen so that the chain  $\mathbf{P} = \{p_{ij}\}$  has  $\pi$  as its stationary distribution and satisfies

$$\pi\mathbf{P} = \pi. \quad (19)$$

A simple form for  $\alpha_{ij}$  is

$$\alpha_{ij} = \min\left\{\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right\}. \quad (20)$$

A sufficient condition for this chain to generate  $\pi$  as its stationary distribution is that  $\{q_{ij}\}$  be irreducible and aperiodic (see Smith and Roberts 1993).

In the context of phylogenetic inference, we are interested in generating the posterior distribution of phylogenetic trees so that the states of the chain are possible labeled histories and  $\pi_i = f(\tau = i|\mathbf{X})$ . The ratios of posterior probabilities used in the calculations are of the form

$$\frac{\pi_j}{\pi_i} = \frac{f(\mathbf{X}|\tau = j)f(\tau = j)/f(\mathbf{X})}{f(\mathbf{X}|\tau = i)f(\tau = i)/f(\mathbf{X})} \quad (21)$$

$$= \frac{f(\mathbf{X}|\tau = j)}{f(\mathbf{X}|\tau = i)}. \quad (22)$$

The density  $f(\mathbf{X})$  is eliminated so that the calculation does not involve the sum of equation (12). The approach is useful for generating the posterior distribution of  $\tau$  even when the number of species in the sample is moderately large. The probability  $f(\mathbf{X}|\tau)$  was estimated by Monte Carlo integration (see equation 13) using two methods. The exploratory method recalculates  $f(\mathbf{X}|\tau)$  at each step with a low level of accuracy. The main method calculates each  $f(\mathbf{X}|\tau)$  only once, but with greater accuracy, and retains this value so that it can be reused the next time this labeled history appears.

### Transitions Between Labeled Histories

A convenient choice for the nominating function  $q_{ij}$  is a stochastic representation of the nearest neighbor interchange (NNI) algorithm for generating transitions between rooted binary trees. This algorithm generates two neighboring topologies for each internal branch, and a rooted tree of  $s$  species has  $2(s - 2)$  neighbors (see fig. 3). We assign an equal probability to each of the neighboring topologies. The NNI algorithm modifies the topology but ignores the ordering of the nodes (i.e., labeled history). To modify the algorithm so that the chain moves between labeled histories, we assign an equal probability to each of the possible labeled histories for a nominated topology. This involves enumerating and recording all the labeled histories for that topology, and an algorithm was devised for this purpose (see appendix B). We also allow the chain to move, with probability  $\beta$ , to another labeled history that belongs to the current tree topology if the topology has more than one labeled history. The transition probability  $q_{ij}$  is then

$$q_{ij} = \begin{cases} \frac{\beta}{h_j - 1}, & \text{if no topology change occurs} \\ \frac{1 - \beta}{2(s - 2)h_j}, & \text{if topology change occurs,} \end{cases} \quad (23)$$

where  $h_j$  is the number of distinct labeled histories for

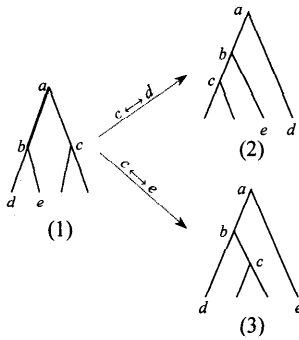


FIG. 3.—Nearest-neighbor interchange (NNI) algorithm for a rooted binary tree topology. The NNI algorithm generates two neighbors for each interior branch. Consider an interior branch  $a-b$ , where  $a$  is the ancestral node and  $b$  is the descendant node. Node  $c$  is the other descendant of  $a$ , and nodes  $d$  and  $e$  are descendants of  $b$ . The two neighbors of tree 1 are generated by interchanging node  $c$  with node  $d$  (tree 2), and node  $c$  with node  $e$  (tree 3). If any of nodes  $c$ ,  $d$ , or  $e$  is an ancestral node, then the entire subtree down that node is moved during the interchange. Similarly, the interior branch  $a-c$  also generates two neighbors, so that tree 1 (or any rooted binary tree of four species) has four neighbors by NNI. Note that all trees are connected by this tree perturbation algorithm; that is, it is possible to reach any particular tree from an initial tree by repeated application of NNI. Also, trees 2 and 3 are themselves neighbors, so that tree 1 can reach tree 2 by either one or two steps. The Markov chain generating transitions among trees that are nominated by the NNI algorithm is therefore irreducible and aperiodic.

the topology of labeled history  $j$ . It can be shown that the Markov chain  $\{q_{ij}\}$  is irreducible and aperiodic, as required (see fig. 3).

#### Calculating Posterior Probabilities for Candidate Trees

For a set of candidate labeled histories generated by the MCMC algorithm outlined above, posterior probabilities can be calculated to a higher degree of accuracy. The posterior probability of topology  $\tau = i$  is  $f(\tau = i, \mathbf{X})/f(\mathbf{X})$ , and this simplifies to  $f(\mathbf{X}|\tau = i)/\sum_{\tau} f(\mathbf{X}|\tau)$ , where the summation is taken over all possible labeled histories. If a set is found that includes all labeled histories with nonnegligible probability, a reasonable approximation for the posterior probabilities is obtained if the above summation is instead performed over this reduced set of labeled histories. A useful approach is then to generate a set of candidate trees from the Markov chain using a moderate degree of accuracy for the Monte Carlo integration and then use the labeled histories of this set, with a high degree of accuracy in the integrations, to calculate more refined estimates of the posterior probabilities for the candidate trees. If some trees with nonnegligible probability are not included in the set of candidate trees, the above approximation will overestimate the posterior probabilities of the trees in this set. The resulting probabilities may then be interpreted as relative probabilities for topologies in the set; these are still useful quantities for comparing trees. Another possibility is to calculate the ratio of the posterior probabilities for a particular pair of trees. This is known as the Bayes factor (Robert 1994). Candidate trees might also be obtained using other phylogenetic methods, such as maximum-likelihood, maximum-parsimony, or distance-based methods. Of course, these methods are not

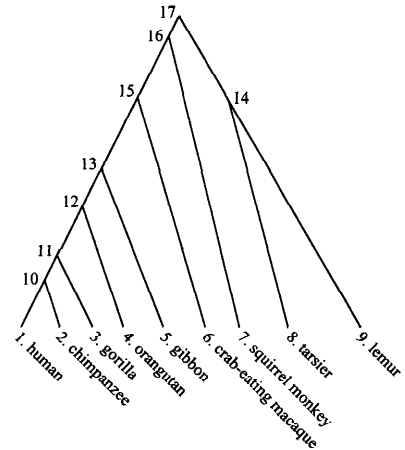


FIG. 4.—Maximum-likelihood tree of nine primate species estimated from the mitochondrial DNA sequences (888 bp). The log likelihood for this tree is  $\ell = -5250.37$  with transition/transversion rate ratio estimated to be  $\hat{\kappa} = 1.63 \pm 0.15$ . Node times are proportional to their estimated values, and the ordering of the nodes corresponds to that of the ancestral speciation times so that the tree also gives the labeled history. In the Bayesian analysis, this same labeled history has the highest posterior probability.

certain to generate the set of all trees with nonnegligible probability, and the results should then not be interpreted as posterior probabilities.

#### Phylogeny of the Primates

To illustrate the method, we analyzed DNA data consisting of a segment of the mitochondrial genomes of human, chimpanzee, gorilla, orangutan, gibbon, crab-eating macaque, squirrel monkey, tarsier, and lemur (Hayasaka, Gojobori, and Horai 1988). The sequence consists of parts of two protein-coding genes and three tRNA genes. There are 888 sites in the sequence after removal of alignment gaps. The transition/transversion rate ratio is estimated to be  $\kappa = 1.63$ , and the mutation rate (the average number of substitutions per site) is  $m = 0.24$ . The observed nucleotide frequencies are 0.2660 (T), 0.3044 (C), 0.3219 (A), and 0.1076 (G).

Estimates of the parameters of the birth-death process, used as a prior, were obtained using maximum-likelihood estimates of node times for the maximum-likelihood tree (fig. 4). The estimates of the node times are  $t_2 = 0.8911$ ,  $t_3 = 0.7329$ ,  $t_4 = 0.6966$ ,  $t_5 = 0.5075$ ,  $t_6 = 0.4026$ ,  $t_7 = 0.2272$ , and  $t_8 = 0.1738$ , with  $t_1 = 1$ . These estimates were treated as observations to estimate the speciation and extinction rates using equation (3) as the likelihood function. The estimates are  $\hat{\lambda} = 6.7$  and  $\hat{\mu} = 2.5$ , if we use the estimate of Wolfheim (1983) that there are approximately 150 species of extant primates in total ( $\rho = 9/150 = 0.06$ ), or  $\hat{\lambda} = 8.2$  and  $\hat{\mu} = 4.1$  if we use the estimate of Kavanagh (1984) that there are approximately 185 species of extant primates in total ( $\rho = 9/185 = 0.049$ ).

Two different analyses were performed using these estimates. In the empirical Bayesian analysis, the estimates of  $\lambda$  and  $\mu$ , described above, were treated as parameters of the prior distribution of nodes times ( $t$ ). In the hierarchical Bayesian analysis, the estimates were

**Table 1**  
**Comparison of 10 Labeled Histories Under Different Models**

Labeled History	Tree Topology	HBA ( $S = 185$ )		HBA ( $S = 150$ )		EBA ( $S = 185$ )		EBA ( $S = 150$ )	
		$\ell$	$\pi$	$\ell$	$\pi$	$\ell$	$\pi$	$\ell$	$\pi$
1 . . . . .	((((((12)3)4)5)6)7)(98))	-5,261.89	0.745	-5,261.74	0.706	-5,260.98	0.713	-5,260.99	0.710
2 . . . . .	((((((12)3)4)5)6)7)(98))	-5,263.17	0.208	-5,262.80	0.243	-5,262.06	0.243	-5,262.05	0.247
3 . . . . .	((((((12)3)4)5)6)7)(98))	-5,268.03	0.002	-5,267.73	0.002	-5,267.34	0.001	-5,267.35	0.001
4 . . . . .	((((((12)3)4)5)6)((98)7))	-5,265.07	0.031	-5,264.75	0.035	-5,264.18	0.029	-5,264.20	0.029
5 . . . . .	((((((12)3)4)5)6)((98)7))	-5,269.86	0.000	-5,269.77	0.000	-5,269.04	0.000	-5,269.11	0.000
6 . . . . .	((((((12)3)4)5)6)((98)7))	-5,269.21	0.001	-5,269.19	0.000	-5,268.49	0.000	-5,268.52	0.000
7 . . . . .	((((((1(23)4)5)6)7)(89))	-5,268.05	0.002	-5,267.89	0.002	-5,267.23	0.001	-5,267.23	0.001
8 . . . . .	((((((1(23)4)5)6)7)(89))	-5,269.67	0.000	-5,269.22	0.000	-5,268.25	0.001	-5,268.27	0.000
9 . . . . .	((((((12)3)4)5)6)((98)7))	-5,266.15	0.011	-5,265.91	0.011	-5,265.15	0.011	-5,265.18	0.011
10 . . . . .	((((((12)3)4)5)6)((98)7))	-5,270.25	0.000	-5,270.22	0.000	-5,269.97	0.000	-5,270.04	0.000

NOTE.—For each topology, the species are numbered as follows: 1 (human), 2 (chimpanzee), 3 (gorilla), 4 (orangutan), 5 (gibbon), 6 (crab-eating macaque), 7 (squirrel monkey), 8 (tarsier), and 9 (lemur). The tree topologies of the first three labeled histories are shown in figure 4. The speciation time for node 14 is between those of nodes 13 and 15 for labeled history 1, between nodes 15 and 16 for labeled history 2, and between nodes 16 and 17 for labeled history 3. Orderings of the nodes for other labeled histories are not given, as these labeled histories have negligibly small posterior probabilities. The logarithm of the probability  $f(\mathbf{X}|\tau)$  is denoted as  $\ell$ , and the posterior probability is denoted as  $\pi$ . HBA, hierarchical Bayesian analysis; EBA, empirical Bayesian analysis;  $S$ , number of species sharing a most recent common ancestor.

used as the means of the uniform priors for  $\lambda$  and  $\mu$ . For example, with the sampling proportion  $\rho = 0.06$ ,  $\lambda$  was generated from the uniform density  $U(0, 2 \times 6.7)$ , and  $\mu$  is generated from the uniform density  $U(0, 2 \times 2.5)$ . The probability of a transition to an alternative labeled history was fixed at  $\beta = 0.15$ , and a low accuracy level ( $\delta = 0.5$ ) was used for the Monte Carlo integration in the initial MCMC analysis used to identify candidate labeled histories.

There are  $\xi(9) = 57,153,600$  possible labeled histories for nine species. We used several different initial labeled histories for the MCMC analysis and found that the chain converged rather quickly to the same subset of labeled histories, independent of the initial labeled history; after about 200 steps, the chain circulates among only a few labeled histories. A total of  $10^5$  replicates were performed for each separate MCMC analysis. The initial 200 replicates of each run were discarded, and the labeled histories visited during the remainder of the MCMC analysis were collected. The posterior probability distribution for these trees was recalculated using a higher accuracy level. The results are summarized in table 1. For all the analyses, one labeled history, identical to that obtained by a maximum-likelihood analysis (fig. 4), has a posterior probability between 70% and 74%. Two alternative labeled histories for this topology were also visited in the MCMC analysis, and one of these has substantial posterior probability (20%–25%). The tree topology shown in figure 4 then has a total posterior probability of about 95%–96%. The bootstrap proportion for the likelihood tree, calculated from these data using the RELI approximation method of Kishino and Hasegawa (1989), is 68%. Consistent with the findings of our previous paper, the posterior probability is greater than the bootstrap proportion. Seven additional labeled histories generated from three different tree topologies were visited during the MCMC analysis, but their posterior probabilities are negligible in comparison to the best topology.

It is noteworthy that the posterior probabilities of different labeled histories do not change much among

the different analyses. Although the sampling proportion ( $\rho$ ) has a major effect on estimates of the birth and death rates, its effect on the posterior probability is relatively minor. Part of the reason is that the birth and death rate parameters are estimated from the same set of node times, so the prior distribution of the node times is expected to be similar for the two sampling proportions used. Furthermore, a comparison of the results for the empirical Bayesian analysis with those for the hierarchical Bayesian analysis suggests that adding a second-level prior for the birth and death rates does not change the posterior probabilities substantially. The reason appears to be that most information concerning phylogeny derives from the data, and the prior density has much less effect. This result agrees with our previous finding, based on two smaller data sets, that the method is robust to variations in the prior (Rannala and Yang 1996).

## Discussion

The exact form of the nominating function  $q_{ij}$  in the Markov chain is arbitrary and it is useful to consider modifications to  $q_{ij}$  that increase the rate of convergence of the chain to the stationary distribution or reduce the amount of computation by avoiding the need to perform numerical integration on trees with very small probabilities. For many data sets, most trees are very unlikely, having vanishingly small posterior probability, and only a small fraction of all possible trees are actually visited by the chain. The efficiency of the method may then be improved by preferentially nominating trees with larger posterior probabilities rather than assigning each neighbor an equal probability as we have done in this paper. One possibility is to alter the transition probabilities to different adjacent topologies (under the NNI perturbation) at each step, weighting these according to some additional optimality criterion that is relatively inexpensive to calculate and is positively correlated with the posterior probability. One possible criterion is the tree length (number of changes) obtained using the maxi-

mum-parsimony algorithm. Another is the conditional log probability  $\log\{f(\mathbf{X}|\tau, \hat{\mathbf{t}})\}$  calculated using approximate estimates of the node times  $\hat{\mathbf{t}}$ . Such modifications do not affect the stationary distribution but may reduce the amount of computation required to generate the distribution. One difficulty with this approach is that the optimality criteria need to be calculated for all labeled histories that the chain may reach at each step, so the algorithm becomes more complicated. Another difficulty is that some possible criteria, such as the tree length under maximum parsimony, do not discriminate among labeled histories for the same topology, although different labeled histories may have quite different posterior probabilities.

Importance sampling techniques are also potentially useful for improving the efficiency of the calculation of  $f(\mathbf{X}|\tau)$  by Monte Carlo integration (see eq. 13). Rather than sampling the node times  $\mathbf{t}$  from the prior distribution, we might sample from another distribution which is concentrated in the region where the integrand is large, and then rescale the calculation (Fishman 1996, pp. 257–259). We noticed in our analysis of the primate data set that more Monte Carlo replicates were needed to achieve the same level of accuracy when a second-level prior was used for the distribution of  $\lambda$  and  $\mu$  than when these were treated as parameters, apparently due to the fact that a second-level prior may generate values of  $\mathbf{t}$  that more often deviate from the empirical distribution. In such cases, importance sampling may be very effective in reducing the cost of the integration.

### Program Availability and Performance

The method described in this paper has been implemented in the *mcmctree* program, written in ANSI C. The program is included in the PAML (Phylogenetic Analysis by Maximum Likelihood) package, which is distributed by Z. Yang and can be obtained by anonymous ftp from our ftp site at [mw511.biol.berkeley.edu/pub](http://mw511.biol.berkeley.edu/pub) or from the Indiana molecular biology ftp site at [ftp.bio.indiana.edu/molbio/evolve](http://ftp.bio.indiana.edu/molbio/evolve). The current version of the program allows a maximum of 10 species to be analyzed on 32-bit machines (compilers).

### Acknowledgments

Stanley Sawyer and two anonymous reviewers provided many helpful comments. Support for this project was provided in part by NIH grant GM40282 to Montgomery Slatkin and by a Natural Sciences and Engineering Research Council (NSERC) of Canada postdoctoral fellowship to B.R.

### APPENDIX A

#### Indexing Labeled Histories

To allow a compact representation, we need a system to index all possible labeled histories for a given number of species. The system should allow us to obtain the index for any labeled history, or the labeled history for any index. Note that a systematic enumeration of labeled histories also leads to an indexing system. We

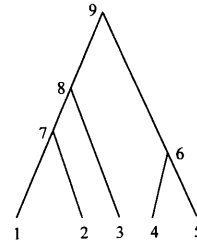


FIG. 5.—An example tree of five species. This is treated as a labeled history in appendix A and as a rooted tree topology in appendix B.

can enumerate all labeled histories by tracing the coalescence events among lineages backward in time. Let the  $k$ th coalescence event join two of  $k + 1$  lineages to produce  $k$  lineages (see fig. 1). There are  $\binom{k+1}{2} = (k + 1)k/2$  possible pairs that may coalesce at this event. The number of distinct labeled histories for  $s$  species is then  $\xi(s) = \prod_{k=1}^{s-1} \binom{k+1}{2} = s!(s - 1)!/2^{s-1}$ , and our index ranges from 0 to  $\xi(s) - 1$ .

A labeled history is indexed by recording which pair of lineages is joined at the  $k$ th coalescence event,  $k = s - 1, \dots, 1$ . A variable  $z_k$  is used for this purpose, which ranges from 0 to  $(k + 1)k/2 - 1$ . If lineages  $i_k$  and  $j_k$  (with  $1 \leq j_k < i_k \leq k + 1$ ) are joined at the  $k$ th coalescence, we let

$$z_k = (i_k - 2)(i_k - 1)/2 + (j_k - 1). \quad (24)$$

Note that  $z_k$  is the index for the  $(i_k, j_k)$ th element of a lower-diagonal matrix when the matrix is arranged into a vector by adjoining rows. The lineage formed by joining  $i_k$  and  $j_k$  is labeled  $j_k$ , and lineage  $k + 1$  is labeled  $i_k$  (if  $i_k < k + 1$ ). At the next step, we calculate  $z_{k-1}$  and again trace the coalescence process backward, repeating this procedure until the last pair of lineages is joined. The index of the labeled history is then

$$I = \sum_{k=1}^{s-1} z_k \eta_k(s), \quad (25)$$

where  $\eta_k(s) = \prod_{j=k+1}^{s-1} \binom{j+1}{2}$  and  $\eta_{s-1} = 1$ . This is the index of the  $(z_1, z_2, \dots, z_{s-1})$ th element in an  $(s - 1)$ -dimensional matrix of the form  $\binom{s}{2} \times \dots \times \binom{s}{2}$  when the elements of the matrix are arranged into a vector by adjoining rows. A sample calculation for the labeled history of figure 5 is given below:

$k$	$i_k$	$j_k$	$z_k$	Renumber Taxa	$\eta_k(s)$
4	5	4	9	6 $\rightarrow$ 4	1
3	2	1	0	7 $\rightarrow$ 1, 4 (6) $\rightarrow$ 2	10
2	3	1	1	8 $\rightarrow$ 1	60
1	2	1	0		

The index of this labeled history is  $I = 9 \times 1 + 0 \times 10 + 1 \times 60 = 69$ . It is also possible to recover the labeled history for a given index  $I$ . The coefficients  $z_k$ ,  $k = (s - 1), \dots, 1$ , are uniquely determined by equation (25) as  $z_k = \text{mod}([I/\eta_k(s)], \binom{k+1}{2})$ ,  $k = (s - 1), \dots, 1$ , where  $[a]$  denotes the integral part of  $a$  and  $\text{mod}(a, b)$  denotes the remainder when  $a$  is divided by  $b$ . From  $z_k$  the coefficients  $i_k$  and  $j_k$  can be recovered (using eq. 24) as

$$i_k = \lfloor (\sqrt{1 + 8z_k} + 3)/2 \rfloor, \quad (26)$$

$$j_k = z_k - (i_k - 1)(i_k - 2)/2 + 1. \quad (27)$$

The labeled history can then be constructed using the  $i_k$  and  $j_k$ ,  $k = (s - 1), \dots, 1$ .

#### APPENDIX B

##### Enumerating Labeled Histories for a Topology

Here, we outline an algorithm for enumerating all the possible labeled histories for a particular rooted tree topology. Starting with the root of the tree, which represents the first speciation event, we enumerate all possible interior nodes that are candidates for the next speciation event, choosing each in turn and repeating the procedure until all candidate nodes at each speciation event are enumerated (visited). Let  $D_k$  be the set of candidate interior nodes for the  $k$ th speciation event and let the node chosen be  $d_k$ ,  $k = 1, \dots, s - 1$ . The set of candidate nodes for the next speciation event,  $D_{k+1}$ , includes all elements of  $D_k$  except for  $d_k$  and the descendant interior nodes of node  $d_k$  (if these exist).

For the rooted tree of figure 5, the first speciation event is at node 9; that is,  $D_1 = \{9\}$  and  $d_1 = 9$ . The root has two descendant nodes, so that  $D_2 = \{8, 6\}$ . There are then two possibilities for  $d_2$ . (1) Suppose that we choose  $d_2 = 8$ . To form  $D_3$ , we keep 6 in  $D_2$  and replace 8 with its descendant node 7, so that  $D_3 = \{7, 6\}$ . Letting  $d_3 = 7$  (so that  $d_4 = 6$ ) or  $d_3 = 6$  (so that  $d_4 = 7$ ) generates two orderings: 9-8-7-6 and 9-8-6-7. (2) If we choose  $d_2 = 6$ , then  $D_3 = \{8\}$  and  $D_4 = \{7\}$  so that  $d_3 = 8$  and  $d_4 = 7$ , leading to the ordering 9-6-8-7. The tree topology then has three possible orderings of the interior nodes (labeled histories): 9-8-7-6, 9-8-6-7, and 9-6-8-7.

Note that simply selecting a node from the set  $D_k$ ,  $k = 1, \dots, s - 1$ , with uniform probability is not guaranteed to produce a labeled history such that each ordering has equal probability. For example, choosing  $d_2 = 8$  or  $d_2 = 6$  with equal probability in the above example will lead to probabilities  $1/4$ ,  $1/4$ , and  $1/2$  for the three orderings, respectively, instead of  $1/3$  each as expected if probabilities are uniform. To assign equal probabilities for all possible labeled histories, we therefore record the possible orderings enumerated by the above algorithm and then select one at random.

#### LITERATURE CITED

- BERGER, J. O. 1985. Statistical decision theory and Bayesian analysis. Springer-Verlag, New York.
- EDWARDS, A. W. F. 1970. Estimation of the branch points of a branching diffusion process. *J. R. Stat. Soc. B* **32**:155–174.
- FEIJER, W. 1939. Die grundlagen der volterraschen theorie des kampfes ums dasein in wahrshcheinlichkeits theoretischen behandlung. *Acta Biotheor.* **5**:1–40.
- FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**:368–376.
- FISHMAN, G. S. 1996. Monte Carlo: concepts, algorithms, and applications. Springer-Verlag, New York.
- HASTINGS, W. K. 1970. Monte Carlo sampling methods using Markov chains and their application. *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.
- KAVANAGH, M. 1984. A complete guide to monkeys, apes and other primates. Viking Press, New York.
- KENDALL, D. G. 1948. On the generalized birth-and-death process. *Ann. Math. Stat.* **19**:1–15.
- . 1949. Stochastic processes and population growth. *J. R. Stat. Soc. B* **11**:230–264.
- KISHINO, H., and M. HASEGAWA. 1989. Evaluation of maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**:170–179.
- NEE, S., R. M. MAY, and P. H. HARVEY. 1994. The reconstructed evolutionary process. *Philos. Trans. R. Soc. Lond. B* **344**:305–311.
- RANNALA, B. 1997. Gene genealogy in a population of variable size. *Heredity* (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.
- ROBERT, C. P. 1994. The Bayesian choice: a decision-theoretic motivation. Springer-Verlag, New York.
- SMITH, A. F. M., and G. O. ROBERTS. 1993. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Stat. Soc. B* **55**:3–23.
- WOLFHEIM, J. H. 1983. Primates of the world. University of Washington Press, Seattle.

STANLEY A. SAWYER, reviewing editor

Accepted March 19, 1997