

# Bayesian Selection of Continuous-Time Markov Chain Evolutionary Models

Marc A. Suchard,\* Robert E. Weiss,† and Janet S. Sinsheimer\*†‡

\*Department of Biomathematics, UCLA School of Medicine; †Department of Biostatistics, UCLA School of Public Health; and ‡Department of Human Genetics, UCLA School of Medicine

We develop a reversible jump Markov chain Monte Carlo approach to estimating the posterior distribution of phylogenies based on aligned DNA/RNA sequences under several hierarchical evolutionary models. Using a proper, yet nontruncated and uninformative prior, we demonstrate the advantages of the Bayesian approach to hypothesis testing and estimation in phylogenetics by comparing different models for the infinitesimal rates of change among nucleotides, for the number of rate classes, and for the relationships among branch lengths. We compare the relative probabilities of these models and the appropriateness of a molecular clock using Bayes factors. Our most general model, first proposed by Tamura and Nei, parameterizes the infinitesimal change probabilities among nucleotides (A, G, C, T/U) into six parameters, consisting of three parameters for the nucleotide stationary distribution, two rate parameters for nucleotide transitions, and another parameter for nucleotide transversions. Nested models include the Hasegawa, Kishino, and Yano model with equal transition rates and the Kimura model with a uniform stationary distribution and equal transition rates. To illustrate our methods, we examine simulated data, 16S rRNA sequences from 15 contemporary eubacteria, halobacteria, eocytes, and eukaryotes, 9 primates, and the entire HIV genome of 11 isolates. We find that the Kimura model is too restrictive, that the Hasegawa, Kishino, and Yano model can be rejected for some data sets, that there is evidence for more than one rate class and a molecular clock among similar taxa, and that a molecular clock can be rejected for more distantly related taxa.

## Introduction

Reconstruction of evolutionary relatedness among biological entities is a powerful tool in evolutionary biology and health care provision. For example, identification of bacterial pathogens and HIV strains by evolutionary relatedness may greatly increase the efficiency of therapeutic interventions (Rudolph et al. 1993; McCabe et al. 1995; Nerurkar et al. 1996; Relman et al. 1996; Crandall 1999). Incorrect evolutionary models and reconstruction methods may lead to inconsistent results or may include unrealistic constraints on the process, sacrificing model accuracy in favor of computational ease and speed (Rzhetsky and Sitnikova 1996; Swofford et al. 1996; Durbin et al. 1998).

Likelihood ratio tests for evolutionary models (for a review, see Huelsenbeck and Rannala 1997) can be remiss in that the topology space of evolutionary relatedness is discrete, data are sparse, parameter estimates may lie on the boundaries, and standard likelihood asymptotics may not apply (Navidi, Churchill, and von Haeseler 1991, 1993; Goldman 1993; Sinsheimer, Lake, and Little 1996; Lange 1997; Whelan and Goldman 1999). Using Markov chain Monte Carlo (MCMC) methods (Gilks, Richardson, and Spiegelhalter 1996) to approximate posterior distributions allows us to broach evolutionary model selection in a computationally feasible manner, with topology determination as an application of reversible jump MCMC (Green 1995). Although MCMC methods have previously been used in the reconstruction of evolutionary relatedness (Kuhner, Yamato, and Felsenstein 1995, 1998; Rannala and Yang 1996; Mau and Newton 1997; Yang and Rannala 1997;

Larget and Simon 1999; Mau, Newton, and Larget 1999; Li, Pearl, and Doss 2000), our methods differ from these in being fully Bayes with a proper, yet nontruncated and uninformative, prior in modeling assumptions, in likelihood computation, in proposal kernels, and in the range of hypotheses tested. The Bayesian hypothesis-testing approach we propose in this paper provides a framework to simultaneously infer evolutionary relationships and test a large set of modeling hypotheses, of which we illustrate only a few.

In the *Materials and Methods* section, we describe the data upon which evolutionary relatedness is determined and models for reconstructing evolutionary trees, we introduce a reversible jump MCMC approach to estimate these relationships, and we show that Bayes factor comparison of evolutionary models is possible using vague but proper priors and can be used without conditioning on a particular topology. To illustrate, in the *Results* section, we compare several hierarchical evolutionary models, examine the appropriateness of a molecular clock, and test the existence of multiple rate classes.

## Materials and Methods

### Evolutionary Relationships and Models *Data and Evolutionary Relationships*

We examined aligned deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) sequences to determine the relatedness among  $N$  organisms. Letting  $i$  index the organism and  $j$  index the site along a given sequence, each position in the data  $X_{ij}$  contained either a nucleotide base or an alignment gap. For simplicity, we first removed all insertion/deletion sites from these alignments to end up with ordered nucleotide sequences of length  $l$ , such that  $X_{ij} \in \{A, G, C, T/U\}$  for all  $i = 1, 2, \dots, N$  and  $j = 1, 2, \dots, l$ .

We assumed that nucleotide sites were independent and identically distributed (iid) within a set of sites

Key words: phylogenetics, Markov chain Monte Carlo, nested hypothesis testing, Bayes factors.

Address for correspondence and reprints: Janet S. Sinsheimer, Department of Human Genetics, UCLA School of Medicine, Los Angeles, California 90095-7088. E-mail: janet@sunlab.ph.ucla.edu.

*Mol. Biol. Evol.* 18(6):1001–1013, 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

evolving under the same evolutionary constraints (rate class  $r$ ). Consequently, the likelihood of observing a given pattern  $X_{1jr}X_{2jr}\cdots X_{Njr}$  within  $r$  was multinomially distributed, where the probability was determined by an unknown bifurcating topology  $\tau$  describing the evolutionary relatedness of the organisms, a set of branch lengths  $t_b \in T$  for  $b = 1, 2, \dots, 2N - 3$ , and a Markovian model for evolutionary change along this topology (Sinsheimer, Lake, and Little 1996). The set  $T$  did not necessarily maintain consistent definition between different topologies. As a result,  $N$ -taxon topologies were nonnested models, each supported on separate parameter spaces  $T^{(\tau)}$ .

### Models of Evolution

A popular class of evolutionary models are continuous-time Markov chain models, parameterized in terms of a  $4 \times 4$  infinitesimal rate matrix of nucleotide change  $Q$  and branch lengths that correspond to the expected number of changes between nodes per site. The matrix  $Q$  satisfies the condition  $Q\mathbf{1} = 0$ , leaving 12 nonnegative, off-diagonal parameters in the most general form of  $Q$ . The transition matrix

$$P(t) = e^{tQ} = \{p_{s_0s_1}(t)\} \quad (1)$$

defines the transition probabilities from state  $s_0$  to state  $s_1$  where  $s_0, s_1 \in (A, G, C, T/U)$  in time  $t$ . The data allow only for the estimation of the product  $tQ$ , so without loss of generality, we constrain  $\text{Trace}(Q) = -1$ .

We explore three nested evolutionary models which reduce the parameterization of  $Q$ . The most general, TN93 (Tamura and Nei 1993), allows for differing evolutionary rates between purine-to-purine transitions ( $\alpha$ ), pyrimidine-to-pyrimidine transitions ( $\gamma$ ), and transversions (purine-to-pyrimidine or pyrimidine-to-purine;  $\beta$ ) and allows the general stationary distribution of the nucleotides ( $\pi$ ) to vary subject to the constraints  $\pi_m \geq 0$ ,  $\sum \pi_m = 1$  for  $m \in (A, G, C, T/U)$  and detailed balance,  $Q\pi = \pi$ . The resulting infinitesimal rate matrix is

$$Q^{\text{TN93}} = \begin{pmatrix} - & \alpha\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha\pi_A & - & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & - & \gamma\pi_T \\ \beta\pi_A & \beta\pi_G & \gamma\pi_C & - \end{pmatrix}, \quad (2)$$

where the minus sign in each row represents minus the sum of the remaining elements in that row. Letting  $\text{Trace}(Q) = -1$  leads to  $\beta = [1 - \alpha(\pi_A + \pi_G) - \gamma(\pi_C + \pi_T)]/2$  and  $(\alpha, \gamma) \in [0, 1) \times [0, 1)$ . TN93 is a generalization of the HKY85 model (Hasegawa, Kishino, and Yano 1985), where  $\alpha = \gamma$ , and of the K80 model (Kimura 1980), where  $\pi_m = 1/4$  and  $\alpha = \gamma$ .

Previous evolutionary reconstructions from nucleic acid data using some MCMC methods fix the stationary distribution at either an empirical estimate from the observed data (Li, Pearl, and Doss 2000) or at values determined by preliminary MCMC sampling (Mau, Newton, and Larget 1999). Like Larget and Simon (1999), we have not adopted either of these approaches. Empirical estimates give equal weight to all taxa and may

therefore be biased when taxon selection oversamples certain subgroups, while fixing parameters can lead to underestimation of the variance of other parameters. Instead, our MCMC approach samples all model parameters.

For all three models, the position of the root, the most recent common ancestor (MRCA) of all  $N$  taxa, is not estimable without further parameter restrictions (Felsenstein 1981), such as a molecular clock. If we can identify an outgroup taxon on the same branch as the root, a molecular clock among the remaining  $N - 1$  taxa is a nested submodel in our framework and thus can be tested. A molecular clock allows for a computationally advantageous parameterization (Mau, Newton, and Larget 1999) and reduces by half the number of branch lengths to be estimated.

We extend the HKY85 parameterization to a mixture model containing  $R$  infinitesimal rate matrices  $Q_r$  and  $R$  sets of branch lengths  $T_r$ , where  $r = 1, \dots, R$ ,  $R$  is the number of different site classes present in the data, and each site in the data is assigned a priori to belong to class  $r$ . We choose HKY85 as an example for comparison with previous work (Yang 1995; Larget and Simon 1999) and note that such mixture models are easily implemented for TN93 or K80 as well. This mixture model is applicable when the reading frame of the DNA sequence is known and rate assignments are based on codon position in the reading frame or when data from different genes known to evolve under different selective pressures are combined. The model is a generalization of the Bayesian computation of Larget and Simon (1999), in which they estimate multiple  $Q_r$  matrices but not multiple branch lengths, and of the work of Yang (1995), where he assumes that the branch lengths between different classes are scalar multiples. Multiple  $Q_r$  matrices allow for different transition/transversion ratios and stationary distributions across classes, and multiple  $T_r$  sets allow for varying rates of evolution both across classes and between species. Yang's (1995) scalar multiple branch lengths assume that the relative rates of evolution between classes are constant across species.

### Bayesian Computation

#### Priors

Priors must remain proper to estimate Bayes factors. We employ flat or vague but completely proper priors over the entire parameter space  $(\tau, \theta^{(\tau)})$ , where  $\theta^{(\tau)} = (\pi, \alpha, \gamma, T, \mu)$  and  $\mu$  is a hyperparameter to help define the prior for  $t_b$  in  $T$ . When we employ multiple site classes,  $\alpha = (\alpha_1, \dots, \alpha_R)$ ,  $\pi = (\pi_1, \dots, \pi_R)$ ,  $T = (T_1, \dots, T_R)$ , and  $\mu = (\mu_1, \dots, \mu_R)$ . We assume a prior for all parameters that is independent of topology such that  $q(\theta^{(\tau)} | \tau) = q(\theta)$  and that all components of  $\theta$  are, a priori, independent. For the TN93 model, we set

$$\tau \sim \text{Uniform over topologies,}$$

$$\pi \sim \text{Dirichlet}(1, 1, 1, 1),$$

$$q(\pi) = \Gamma(4) \quad \text{on } 0 \leq \pi_m \leq \sum \pi_m = 1,$$

$$\begin{aligned} \alpha &\sim \text{Uniform}[0, 1), & q(\alpha) &= 1\{0 \leq \alpha < 1\}, \\ \gamma &\sim \text{Uniform}[0, 1), & q(\gamma) &= 1\{0 \leq \gamma < 1\}, \\ t_b | \mu &\sim \text{Exponential}(\mu), & q(t_b | \mu) &= \frac{1}{\mu} e^{-(1/\mu)t_b}, \end{aligned}$$

$t_b \geq 0,$

and

$$\begin{aligned} \mu &\sim \text{Inv-gamma}(2.1, 1.1), \\ q(\mu) &= \frac{(1.1)^{(2.1)}}{\Gamma(2.1)} \mu^{-(2.1+1)} e^{-1.1/\mu}, \quad \mu > 0. \end{aligned} \quad (3)$$

Except for branch lengths, these prior probabilities are uninformative. The uniform distribution on  $\tau$  is over the discrete space of  $(2N - 5)!/2^{N-3}(N - 3)!$  possible topologies for  $N$  taxa (Felsenstein 1978). Branch lengths are iid given  $\mu$ . We take  $\mu$  to have expectation 1 and variance 10, as, a priori, we know little about its tendencies. The prior for  $t_b$ , supported on  $[0, \infty)$ , is vague but integrable. The usual Jeffreys' prior on  $t_b \in [0, \infty)$  is  $1/t_b$  (Jeffreys 1998). This prior is not integrable and precludes testing a molecular clock. The inverse-gamma density allows computation of the reciprocal moments of  $\mu$ , which are also needed to test for a molecular clock (see appendix B).

### Computation

We employ a Metropolis-within-Gibbs (Tierney 1994) sampler using reversible model jumping (Green 1995) among topologies. The parameter space dimension remains constant across topology models, although interpretation of portions of  $T$  is model-dependent. Each new state of the Markov chain is proposed via a Gibbs cycle. In each step within the cycle, a single parameter block is updated conditional on the remaining blocks using a Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970). We use the following update cycle:

$$\begin{aligned} &\tau, T | \alpha, \gamma, \pi, \mu \\ &\mu | \tau, T, \alpha, \gamma, \pi \\ &T | \tau, \alpha, \gamma, \pi, \mu, \quad \text{and} \\ &\alpha, \gamma, \pi | \tau, T, \mu. \end{aligned} \quad (4)$$

Previous MCMC approaches have updated  $\tau$  and  $T$  simultaneously (Larget and Simon 1999; Mau, Newton, and Larget 1999); however, these approaches consider, at most, proportional rates of evolution between rate classes. Here, we include an extra  $T$ -only block to improve mixing within different sets of branch lengths for each class. We give our transition kernels for each Metropolis-Hastings step in appendix A. Where possible, we employ transition kernels that are symmetric to decrease computational complexity and are supported on the same bounded or discrete space as the kernel's underlying parameters to increase acceptance probabilities.

Each run of our MCMC chain consists of 500,000 full update cycles, and we disregard the first 100,000 steps as burn-in. For the starting state, we draw  $\tau, \mu, T | \mu, \alpha,$  and  $\gamma$  directly from the prior distributions and

set  $\pi$  equal to observed nucleotide frequencies in each class. We estimate functions of the chain's posterior by subsampling every 40 steps after burn-in. Multiple chains are run to insure adequate convergence. We use  $D = \sum t_b$ , representing the total divergence between all taxa,  $\mu, \alpha, \gamma,$  and  $\pi,$  to assess convergence within and across topologies. These parameters retain their interpretation as the sampler moves between topologies and may be used effectively to monitor how well the MCMC sampler is performing (Brooks and Guidici 1999).

We calculate the likelihood of the data given  $\tau, T, \alpha, \gamma,$  and  $\pi$  by integrating out the unknown states of the internal nodes using the pruning algorithm of Felsenstein (1981).

### Model Comparisons

We make comparisons among models using Bayes factors (Kass and Raftery 1995). The Bayes factor in favor of model  $M_1$  against model  $M_0,$  given the data  $Y,$  can be expressed as

$$B_{10} = \frac{f(Y|M_1)}{f(Y|M_0)}, \quad (5)$$

where  $f(Y|M_i) = \int f(Y|\theta_i, M_i)q(\theta_i) d\theta_i,$   $\theta_i$  are the parameters under model  $M_i,$   $f(\cdot)$  are sampling densities, and  $q(\cdot)$  are priors. Different densities are distinguished by their arguments in a common abuse of notation.

If model  $M_0$  is nested within another model  $M_1$  such that the parameter space of  $M_1$  is  $\theta_1 = (\omega, \phi)$  and the parameter space of  $M_0$  is  $\theta_0 = (\omega_0, \phi)$  where  $\omega_0$  is a known constant with  $q_0(\phi) \propto q_1(\omega = \omega_0, \phi),$  then  $B_{10}$  may be estimated via posterior simulation of  $M_1$  using the Savage-Dickey ratio (Verdinelli and Wasserman 1995),

$$\frac{1}{B_{10}} = B_{01} = \frac{p(\omega = \omega_0 | Y, M_1)}{q(\omega = \omega_0 | M_1)}, \quad (6)$$

where  $q(\omega = \omega_0 | M_1)$  is the prior and  $p(\omega = \omega_0 | Y, M_1)$  is the posterior of  $\omega,$  both evaluated at  $\omega_0.$  There exist several methods for estimating the posterior density of  $\omega$  from an MCMC simulation, including nonparametric kernel density estimation methods and multivariate normal approximations. The priors induced by restrictions of the infinitesimal rate matrices and by restrictions of the number of classes are derivable. The restrictions and induced priors are presented in the next two sections. In contrast, the derivation of the priors induced by the molecular-clock restrictions is more involved. Analytical results for some situations are presented in appendix B, and a general numerical approximation is presented in the *Induced Priors on Coalescent Height Differences* section, below.

### Restrictions on Evolutionary Rates

The HKY85 model is a restriction of TN93, and the K80 model is a restriction of HKY85. To test the appropriateness of the restricted models, we generate a posterior sample of the joint  $(\alpha, \gamma, \pi)$  using our MCMC sampler under our most general TN93 model. We then

estimate the Bayes factors in favor of TN93 against HKY85 and in favor of TN93 against K80. We then generate a posterior sample of the joint  $(\alpha, \pi)$  under HKY85 to estimate the Bayes factor in favor of HKY85 against K80.

We approximate the posterior densities of  $(\alpha, \gamma, \pi)$  and  $(\alpha, \pi)$  using a normal approximation with the estimated posterior mean and posterior covariance evaluated at the joint restriction  $\alpha = \gamma$  and at  $(\pi_m = 1/4, \alpha = \gamma)$  (in the former case) and at  $\pi_m = 1/4$  (in the latter case). We directly calculate the appropriate prior densities at these restrictions. When testing  $\alpha = \gamma$ , we recall that this restriction is equivalent to  $\alpha - \gamma = 0$  and that the difference of two Uniform[0, 1] random variables is triangularly distributed on  $[-1, 1]$  with a density of 1 at the restriction (Feller 1971). We then form the respective Bayes factors using equation (6).

### Multiple Classes

In data sets putatively containing multiple site classes, we estimate the Bayes factors in favor of multiple  $Q_r$  matrices under HKY85 by first generating a posterior sample of  $(\alpha_1, \dots, \alpha_R, \pi_1, \dots, \pi_R)$  using our MCMC sampler. We approximate the posterior density at the restriction  $(\alpha_1 = \dots = \alpha_R, \pi_1 = \dots = \pi_R)$  using a normal approximation based on the posterior mean and posterior covariance of the sample reparameterized as

$$\psi = (\alpha_1 - \alpha_R, \dots, \alpha_{R-1} - \alpha_R, \pi_1 - \pi_R, \dots, \pi_{R-1} - \pi_R), \quad (7)$$

where the elements of  $\pi_r$  are  $\pi_{m,r}$ ,  $r \in (1, \dots, R)$ ,  $m \in (A, G, C, U/T)$ , evaluated at  $\psi = (0, \dots, 0)$ .

We form the Bayes factor by dividing this posterior density estimate by the prior evaluated at the joint restriction. The induced prior equals  $q(\alpha_1 - \alpha_R, \dots, \alpha_{R-1} - \alpha_R) \times q(\pi_1 - \pi_R, \dots, \pi_{R-1} - \pi_R)$  by prior independence. We identify that

$$(\alpha_1 - \alpha_R, \dots, \alpha_{R-1} - \alpha_R) = (\alpha_1, \dots, \alpha_{R-1}) - (\alpha_R, \dots, \alpha_R) = U - W, \quad (8)$$

where  $U$  and  $W$  are two independent, multidimensional, random variables. We evaluate  $q(U - W)$  at  $U - W = 0$  using the convolution integral for the difference of two random variables. This results in

$$q(\alpha_1 - \alpha_R = 0, \dots, \alpha_{R-1} - \alpha_R = 0) = 1 \quad (9)$$

given our prior on  $\alpha$ . This calculation does not depend on the number of classes  $R$ .

Following a similar derivation,

$$q(\pi_1 - \pi_R = 0, \dots, \pi_{R-1} - \pi_R = 0) = 6^{R-1}. \quad (10)$$

### Molecular-Clock Restrictions

To test the appropriateness of a molecular clock, we condition on the posterior mode topology, identify a known outgroup, and reparameterize the branch lengths in terms of coalescent height differences,  $\Delta_{ij}$ . These parameters measure the difference in the sums of the

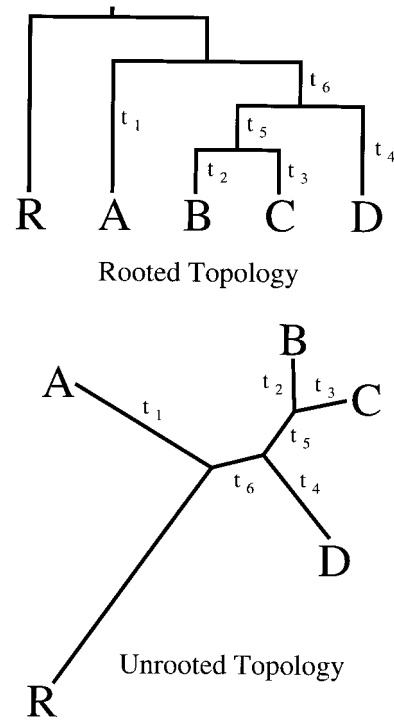


FIG. 1.—Topology used for simulation under a molecular clock. Taxon  $R$  is assigned as the outgroup (having diverged before the remaining taxa) to allow rooting at an arbitrary position along taxon  $R$ 's branch.

branch lengths between two contemporary taxa  $i$  and  $j$  and their MRCA. Under a molecular clock,  $\Delta_{ij} = 0$ .

Given the outgroup  $R$  in the topology illustrated in figure 1, a molecular clock constrains

$$\begin{aligned} \Delta_{AB} = 0 &= t_1 - (t_2 + t_5 + t_6), \\ \Delta_{AC} = 0 &= t_1 - (t_3 + t_5 + t_6), \\ \Delta_{AD} = 0 &= t_1 - (t_4 + t_6), \\ \Delta_{DB} = 0 &= t_4 - (t_2 + t_5), \\ \Delta_{DC} = 0 &= t_4 - (t_3 + t_5), \quad \text{and} \\ \Delta_{BC} = 0 &= t_2 - t_3. \end{aligned} \quad (11)$$

Each constraint may be considered marginally as a diagnostic to identify portions of the topology which violate or support a molecular clock, or all constraints may be examined jointly. The ability to simultaneously consider each constraint marginally is an advantage of our framework over previous molecular-clock tests and allows for testing of a local molecular clock (Hillis, Mable, and Moritz 1996; Huelsenbeck, Larget, and Swoford 2000) within a subset of taxa.

Taken jointly, half of the constraints in equation (11) are redundant, so we can reduce the joint restriction dimensionality by employing a conditioning argument to show that

$$\Pr(\Delta_{AB} = 0, \Delta_{AC} = 0, \Delta_{AD} = 0, \Delta_{DB} = 0, \Delta_{DC} = 0, \Delta_{BC} = 0) = \Pr(\Delta_{AB} = 0, \Delta_{DB} = 0, \Delta_{BC} = 0). \quad (12)$$

It is straightforward to extend the conditioning argument to larger trees.

### Induced Priors on Coalescent Height Differences

For two- or three-taxon rooted topologies and for any arbitrary pair of taxa considered marginally in an  $N$ -taxon topology given an outgroup, we derive exact expressions for the ordinate of the induced prior  $q(\Delta_{ij} = 0)$  in appendix B. To consider constraints jointly for  $N > 3$ , we estimate  $\hat{\eta}$ , the ordinate of the induced joint prior  $q(\Delta_{ij} = 0)$ , via simulation. We draw  $n = 50,000$  samples of  $\mu$  and  $T|\mu$  directly from our priors and form the appropriate  $\Delta_{ij}$  from  $T$ . We calculate  $\hat{\eta}$  using the multi-dimensional density estimator

$$\hat{\eta} = \frac{1}{n\omega_d} \sum_{k=1}^n 1_{\{\|\Delta_{ij}^{(k)}\| \leq w\}}, \quad (13)$$

where  $\Delta_{ij}^{(k)}$  is the  $k$ th sample,  $d$  is the dimension of  $\Delta_{ij}$ ,  $w$  is the radius of a hypersphere in  $\mathfrak{R}^d$ ,  $\omega_d = \pi^{d/2} w^d / \Gamma[(d/2) + 1]$  is its volume, and  $\|\cdot\|$  is the Euclidean norm. For each simulation, we fix  $w$  at its smallest value such that the hypersphere contains at least  $\sqrt{n}$  of the simulation sample (Loftsgaarden and Quesenberry 1965). If the true density  $q(\Delta_{ij})$  is locally linear in the neighborhood of 0, then this method is unbiased. For  $N = 2$  taxa,  $q(\Delta)$  does not satisfy the locally linear condition, as the mode of  $\Delta$  is 0 (see appendix B). For  $N \geq 3$ , the mode is no longer centered at 0, and the approximation's bias decreases.

Conditioning on  $\omega_d$  and recalling that  $\hat{\eta}$  is the sum of independent Bernoulli random variables, the finite sample variance is approximated by

$$\text{Var}(\hat{\eta}) = \frac{\hat{\eta}(0)(1 - \omega_d \hat{\eta}(0))}{N\omega_d}. \quad (14)$$

As a diagnostic for these simulations and the estimator, we compare  $\hat{\eta}$  with the analytic results developed in appendix B for  $n = 2$  and 3. For  $n = 2$ , we calculate  $\hat{\eta} = 1.0 \pm 0.1$  (estimate  $\pm$  SD), and the exact result equals 0.955. For  $n = 3$ ,  $\hat{\eta} = 0.87 \pm 0.02$ , while the exact result equals 0.897. To evaluate the estimator in higher dimensions, we drew 50,000 samples from a 12-dimensional multivariate  $N(1, I)$ , where  $1 = (1, \dots, 1)^t$  and  $I$  is the identity matrix, and obtained  $\hat{\eta} = 3.5 \times 10^{-8} \pm 0.5 \times 10^{-8}$ , while the theoretical density is  $4.0 \times 10^{-8}$ . These results return the theoretical densities to within the same order of magnitude and show only small simulation error or bias.

## Results

To make our inference methods more concrete, we examined four data sets: (1) simulated data, (2) representative organisms from across all living kingdoms (Tree of Life [TOL]), (3) primates, and (4) different HIV isolates. Each of these data sets illustrates different aspects of Bayesian inference. The simulated data demonstrated that a molecular clock will be accepted when it is actually present. The primate data were used to test for multiple rates and to test restrictions on the in-

tesimal rate matrix without conditioning on topology. The TOL data, the primates, and the HIV isolates demonstrated the versatility of the Bayesian method in testing the molecular-clock hypothesis. The TOL data also demonstrated that our MCMC implementation is practical for as many as 15 taxa.

### Simulated Data

To insure that our methods would support a molecular clock if one were present, we simulated sequences of length 1,500 under a molecular clock for four contemporary taxa ( $A, B, C, D$ ) and an outgroup ( $R$ ) using the topology in figure 1. We imposed a molecular clock by assigning branch lengths so that the evolutionary distances from MRCA and contemporary taxa were equal. The approximate posterior density of  $\Delta_{ij} = 0$  was 1,106.3. A  $\log_{10} B_{10}$  value of 0 implies that both models are equally likely, while values greater than 2 represent very strong evidence in support of the general model and values less than  $-2$  represent very strong evidence in support of the restricted model (Kass and Raftery 1995). The induced prior for  $\Delta_{ij}$  in this topology was 0.52, yielding a  $\log_{10} B_{10}$  value of  $-3.32$ , which favors a molecular clock.

### Tree of Life

The TOL data set consisted of 15 16S ribosomal RNA sequences (rRNA) (Lake 1988). There were 1,039 aligned nucleotides after removal of gaps, and  $\pi_{\text{obs}} = (0.2408, 0.3157, 0.2464, 0.1971)'$ . The species were drawn from four major classes of living organisms: eukaryotes, eubacteria, halobacteria, and eocytes, and also included the chloroplastic sequence from a eukaryote, *Zea mays* (chl.). Figure 2 (left) shows the modal topology ( $86\% \pm 3\%$ , posterior probability mean  $\pm$  SD determined from 10 independent chains) and the conditional posterior mean branch lengths estimated under the TN93 model. The model correctly clustered the eukaryotes, eocytes, halobacteria, and eubacteria into their appropriate monophyletic groups (clades) based on organism morphology and clustered the chloroplastic sequence in the eubacterial clade. This result is consistent with the endosymbiotic hypothesis of the origins of eukaryotic cellular organelles (Margulis 1981) and has been demonstrated previously using rRNA (Lake 1988; Bhattacharya and Medlin 1995). Table 1 lists the marginal posterior means and standard errors of  $\alpha, \gamma, \pi, \mu,$  and  $D$  under TN93, HKY85, and K80.

### Primates

The primate data comprised a portion of the mitochondrial DNA from a human, a chimpanzee, a gorilla, an orangutan, a gibbon, a macaque, a squirrel monkey, a tarsier, and a lemur (Brown et al. 1982; Hayasaka, Gojobori, and Horai 1988) and had previously been analyzed using MCMC methodology (Yang and Rannala 1997; Larget and Simon 1999). There were 888 sites after removal of alignment gaps, and  $\pi_{\text{obs}} = (0.3219, 0.1076, 0.3044, 0.2660)'$ . Figure 3 illustrates the two

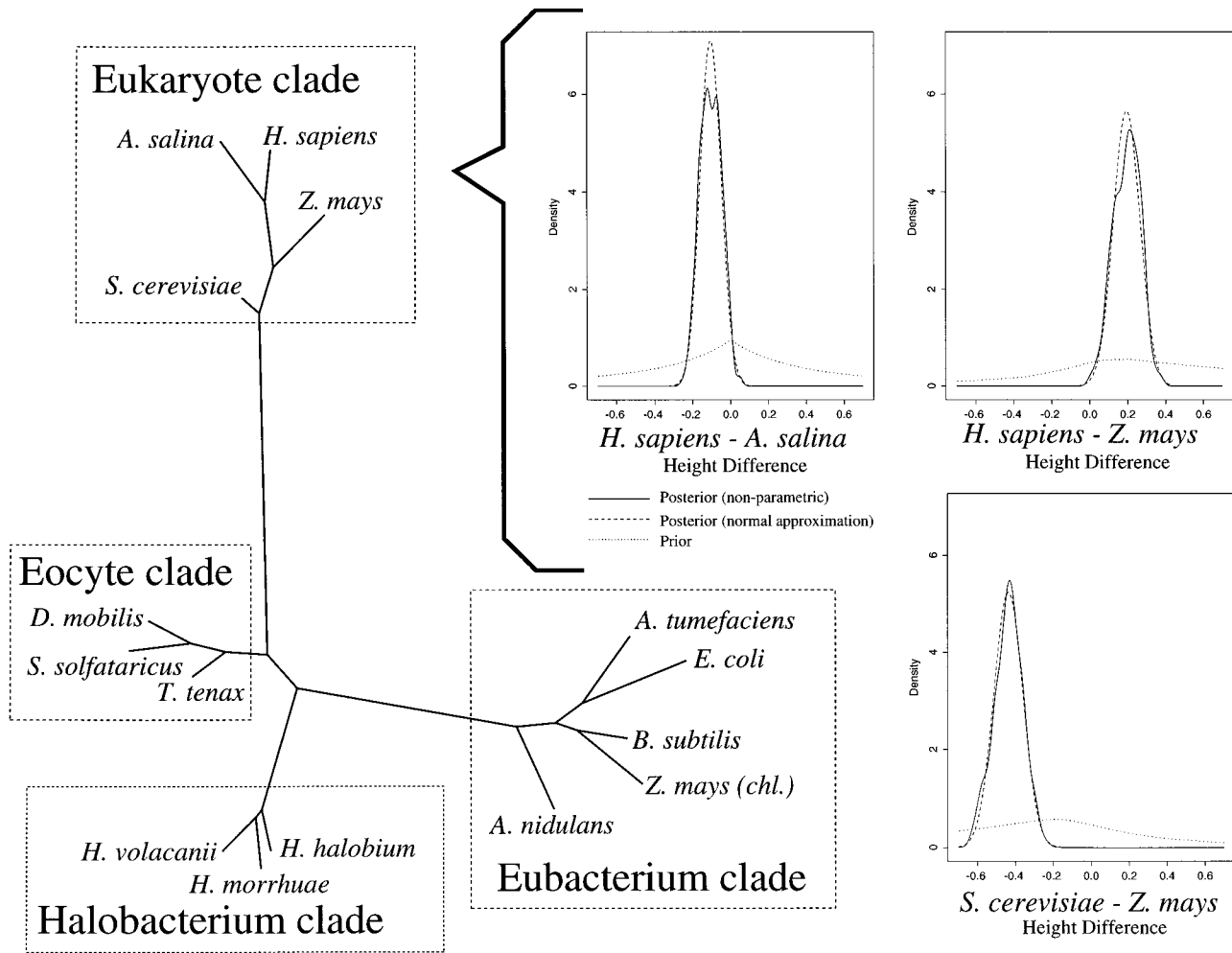


FIG. 2.—The Tree of Life modal ( $87\% \pm 3\%$ ) topology under TN93 (left). Branch lengths are drawn to scale. Plots of the marginal posterior (solid line), normal approximation to the posterior (dashed), and prior (dotted) for the three molecular-clock height differences ( $\Delta_{ij}$ ) within the Eukaryote clade are shown (right). The data do not support a molecular-clock restriction, as the posterior densities are less than the prior density at  $\Delta_{ij} = 0$ .

dominant topologies seen in the posterior of all models and the conditional posterior mean branch lengths under the TN93 model. In both topologies, the sampler properly clusters the apes, monkeys, and prosimians. Table 1 gives evolutionary parameters and divergence estimates and their standard errors.

The posterior distribution of topologies was model-dependent, with the relationship between humans, chimps, and gorillas varying. Under TN93, humans and chimps were topographically the most closely related among the three species ( $90\% \pm 3\%$ ). Similarly, under HKY85, the posterior mean was  $84\% \pm 2\%$  using one rate class and  $92\% \pm 3\%$  using four rate classes. Under K80, the posterior mean was  $90\% \pm 3\%$ . However, under an even more restrictive model proposed by Jukes and Cantor (1969) (JC69) in which  $\alpha = \gamma = \beta$  and  $\pi_m = 1/4$ , we found that chimps and gorillas were topographically the most closely related ( $88\% \pm 5\%$ ). Unconditional on topology, the distance (expected number of changes per site) between humans and chimps was  $0.41 \pm 0.05$  (posterior mean  $\pm$  posterior SD) and the distance between chimps and gorillas was  $0.52 \pm 0.05$

under TN93. Under JC69, these distances were  $0.40 \pm 0.05$  and  $0.45 \pm 0.05$ , respectively.

## HIV

The HIV data contained the complete HIV genomes of two subtype D isolates, eight subtype B isolates, and one ADI subtype recombinant, MAL (Korber et al. 1997). The subtype B isolates JRCSF and JRFL were collected from the same patient. There were 7,969 sites after removing all gaps in the aligned genomes, and  $\pi_{\text{obs}} = (0.3698, 0.2365, 0.1708, 0.2229)'$ . Figure 4 displays the two topologies that account for virtually 100% of the posterior. These topologies were drawn with their conditional posterior mean branch lengths under TN93. One internal branch within the subtype B clade was approximately zero. At zero, the two shown topologies become equivalent. The sampler placed JRFL and JRCSF as nearest neighbors and correctly clustered the D and B subtypes. Table 1 gives the estimated evolutionary parameters and divergence.

**Table 1**  
**Parameter Estimates for the Tree of Life (TOL),**  
**Primates, and HIV Under the TN93, HKY85, and K80**  
**Models**

	TN93	HKY85	K80
<b>TOL</b>			
$\mu$ . . . .	0.356 (0.070)	0.349 (0.066)	0.354 (0.070)
$D$ . . . .	8.900 (0.225)	8.757 (0.211)	8.768 (0.212)
$\alpha$ . . . .	0.400 (0.020)	0.530 (0.012)	0.529 (0.012)
$\gamma$ . . . .	0.702 (0.033)	= $\alpha$	= $\alpha$
$\pi_A$ . . . .	0.246 (0.009)	0.225 (0.008)	1/4
$\pi_G$ . . . .	0.309 (0.010)	0.298 (0.009)	1/4
$\pi_C$ . . . .	0.242 (0.008)	0.263 (0.008)	1/4
$\pi_T$ . . . .	0.203 (0.008)	0.224 (0.008)	1/4
<b>Primates</b>			
$\mu$ . . . .	0.421 (0.111)	0.425 (0.107)	0.396 (0.101)
$D$ . . . .	5.664 (0.198)	5.755 (0.197)	5.276 (0.176)
$\alpha$ . . . .	0.580 (0.044)	0.675 (0.017)	0.656 (0.016)
$\gamma$ . . . .	0.744 (0.036)	= $\alpha$	= $\alpha$
$\pi_A$ . . . .	0.323 (0.012)	0.313 (0.011)	1/4
$\pi_G$ . . . .	0.110 (0.008)	0.103 (0.007)	1/4
$\pi_C$ . . . .	0.290 (0.011)	0.298 (0.011)	1/4
$\pi_T$ . . . .	0.277 (0.011)	0.286 (0.010)	1/4
<b>HIV</b>			
$\mu$ . . . .	0.136 (0.031)	0.133 (0.030)	0.129 (0.029)
$D$ . . . .	1.639 (0.030)	1.565 (0.019)	1.532 (0.024)
$\alpha$ . . . .	0.633 (0.016)	0.696 (0.009)	0.528 (0.006)
$\gamma$ . . . .	0.806 (0.026)	= $\alpha$	= $\alpha$
$\pi_A$ . . . .	0.374 (0.005)	0.368 (0.005)	1/4
$\pi_G$ . . . .	0.232 (0.004)	0.228 (0.004)	1/4
$\pi_C$ . . . .	0.177 (0.004)	0.182 (0.003)	1/4
$\pi_T$ . . . .	0.217 (0.004)	0.221 (0.004)	1/4

NOTE.—Posterior means and standard deviations of the branch length hyperparameter ( $\mu$ ), the total divergence ( $D$ ), the infinitesimal rate parameters ( $\alpha$  and  $\gamma$ ), and the stationary distribution ( $\pi$ ) are shown. The  $\alpha$  and the fractions in the final two columns indicate fixed-model restrictions.

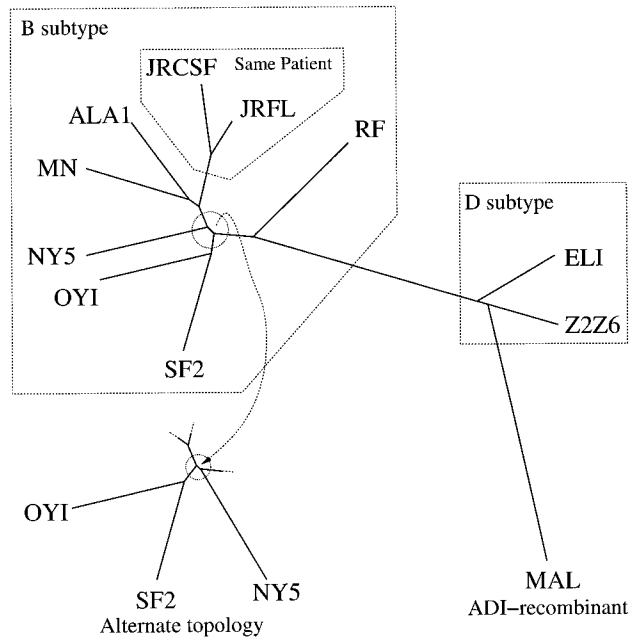


FIG. 4.—Two topologies account for 100% of the posterior for HIV under TN93. Branch lengths are drawn to scale. The two topologies converge as the circled internal branches approach zero.

**TN93, HKY85, and K80 Comparison**

The  $\log_{10}$  Bayes factors for all examples and models are given in table 2. TN93 was strongly supported by the TOL and HIV examples when comparing restrictions of  $\alpha$ ,  $\gamma$ , and  $\pi$ ; all  $\log_{10} B_{10}$  values were  $\geq 3$ . Support for TN93 over HKY85 was less conclusive when restrictions of  $\alpha$  and  $\gamma$  were compared for the primate example, for which the  $\log_{10} B_{10}$  value was 0.3. HKY85 was strongly supported over K80 when restrictions of  $\pi$  were compared.

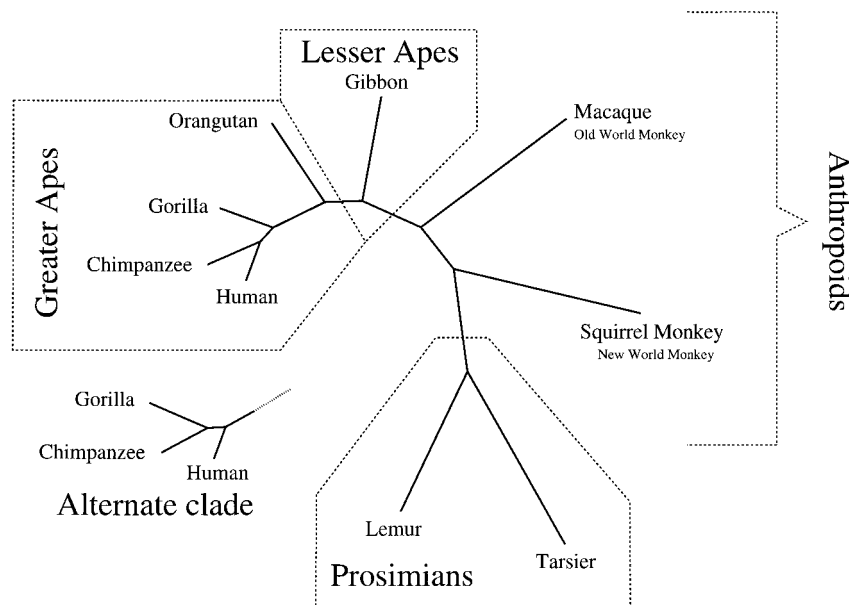


FIG. 3.—The two dominant topologies for primates under TN93 using one rate class. The complete displayed topology has a posterior probability of  $90\% \pm 3\%$ , while the alternate clade accounts for the remaining 10%. Branch lengths are drawn to scale.

**Table 2**  
**Log<sub>10</sub> Bayes Factors in Favor of the More General Evolutionary Model Against a Nested Model for the Tree of Life (TOL), Primates, and HIV**

Data Set	Log <sub>10</sub> $B_{10}$	HKY85	K80
TOL .....	TN93	8	11
	HKY85	0	3
Primates .....	TN93	0.3	93
	HKY85	0	107
HIV .....	TN93	4	163
	HKY85	0	167

### Multiple Classes in the Primates

The mitochondrial sequences in the primate data set comprised the coding region for two individual protein subunits with known reading frames and a transfer RNA (tRNA) portion (Brown et al. 1982; Hayasaka, Gjobori, and Horai 1988). Following Yang (1995), we divided the data into four classes, one for the tRNA (194 nt in length) and the remaining three for the first, second, and third codon positions in the protein subunits with lengths 232, 231, and 231 nt. In table 3, we report the posterior means and the posterior standard deviations of the parameters  $\alpha$ ,  $\pi$ ,  $\mu$ , and  $D$  for each class under HKY85. The total divergence  $D$  serves as a surrogate for reporting all branch lengths  $T$ . The posterior of the ratio  $D_i/D_j$  estimates the relative rates of evolution between classes  $i$  and  $j$ . Between the first and second codon positions, we found a posterior mean ratio of 0.49 (0.06 SD), between the first and third we found a posterior mean ratio of 10.2 (2.2), and between the first position and tRNA we found a posterior mean ratio of 0.62 (0.07). These estimates are comparable to those determined by Yang (1995) and include measures of uncertainty. An approximate 10-fold increase in mutation was observed in the third codon position compared to the first, consistent with the increased redundancy of the genetic code in the third position. Furthermore, the evolutionary rate parameters  $\alpha$  and stationary distributions  $\pi$  between classes were quite disparate. The log<sub>10</sub> Bayes factor in favor of multiple  $\alpha$  and  $\pi$  across all four classes was 43.7.

### Testing a Molecular Clock

We examined the appropriateness of a molecular clock conditional on the modal topology under TN93. We chose the eocyte clade as the outgroup for TOL, as using this clade offered the least support against a molecular clock. The ADI recombinant was assumed to be the outgroup for the HIV example. For the primate example, the lemurs represented the outgroup. In table 4, we list the posterior means and standard deviations of minimum sufficient sets of molecular-clock constraints for these examples.

There were 9 constraints for HIV, 7 for the primates, and 11 for the TOL mode topologies. We further give the number of nodes traversed between the taxa and their MRCA, each  $\Delta_{ij}$ 's marginal prior density evaluated at 0 as determined by these numbers of nodes

**Table 3**  
**Parameter Estimates When Fitting Four Site Classes Using the Primate Example Under the HKY85 Model**

	PROTEIN READING FRAME—CODON POSITION			
	First	Second	Third	tRNA
$\mu$ ...	0.42 (0.11)	0.25 (0.07)	4.16 (1.62)	0.30 (0.08)
$D$ ...	5.28 (0.37)	2.54 (0.22)	65.0 (18.5)	3.22 (0.29)
$\alpha$ ...	0.61 (0.03)	0.79 (0.04)	0.92 (0.02)	0.74 (0.04)
$\pi_A$ ..	0.37 (0.02)	0.18 (0.02)	0.38 (0.02)	0.34 (0.03)
$\pi_G$ ...	0.13 (0.01)	0.12 (0.02)	0.05 (0.01)	0.14 (0.02)
$\pi_C$ ...	0.27 (0.02)	0.30 (0.02)	0.39 (0.02)	0.22 (0.02)
$\pi_T$ ...	0.23 (0.02)	0.40 (0.03)	0.18 (0.01)	0.30 (0.03)

NOTE.—Posterior means and standard deviations of the branch length hyperparameter ( $\mu$ ), the total divergence ( $D$ ), the infinitesimal rate parameter ( $\alpha$ ), and the stationary distribution ( $\pi$ ) are shown for each class.

(appendix B), and log<sub>10</sub>  $B_{10}$  against a molecular clock for each two-taxon comparison. Examined univariately, all  $\Delta_{ij}$ 's supported a molecular clock within the B subtype clade for HIV, with the exception of the JRCSF-JRFL constraint. Between clades, a molecular clock was strongly rejected (SF2-ELI constraint, log<sub>10</sub>  $B_{10}$  = 10.8). Within the primates, a molecular clock was weakly supported by each constraint within the anthropoids (apes and monkeys) (all log<sub>10</sub>  $B_{10}$   $\leq$  -0.5) but rejected between the anthropoids and prosimians (log<sub>10</sub>  $B_{10}$  = 1.0).

Also in table 4, we calculate the joint posterior using a multivariate normal approximation and prior densities via simulation evaluated at 0, and by taking the ratio of these values, we report the joint log<sub>10</sub>  $B_{10}$  for each data set. The TOL and HIV examples offered strong support against a molecular clock (log<sub>10</sub>  $B_{10}$  = 31.8 and 12.3, respectively), while the primates offer weaker support against a molecular clock (log<sub>10</sub>  $B_{10}$  = 1.3).

We examined three subsets of our examples identified as interesting by the marginal diagnostics: (1) the eight subtype B isolates, (2) the anthropoids, and (3) the eukaryotes. Table 4 displays the corresponding joint log<sub>10</sub> posterior and prior densities and log<sub>10</sub>  $B_{10}$  for these subsets. As an illustration, figure 2 (right) plots the marginal posterior and prior distributions of the three coalescent height differences among the eukaryotes. The eukaryotes continued to offer strong support against a molecular clock (log<sub>10</sub>  $B_{10}$  = 14.0), while the much more closely related B subtype isolates and anthropoids offered strong support in favor of a local molecular clock (log<sub>10</sub>  $B_{10}$  = -3.7 and -2.4, respectively).

### Discussion

We propose a reversible jump MCMC algorithm for sampling from the posterior distribution of topologies and other parameters used to model the relatedness among organisms. Individual topologies are separate statistical models. While evolutionary parameters retain definition across these models, some branch lengths do not. For a fixed number of organisms, the dimension of the parameter space spanned by the branch lengths within a topology model remains constant, making reversible model jumps convenient.

**Table 4**  
**Molecular-Clock Estimates for the Tree of Life (TOL), Primates, and HIV**

MOLECULAR-CLOCK HEIGHT DIFFERENCE	POSTERIOR			Traversed Nodes	PRIOR	
	Mean	(SD)	Log <sub>10</sub> $f(\Delta_{ij} = 0   Y)$		Log <sub>10</sub> $p(\Delta_{ij} = 0)$	Log <sub>10</sub> $B_{10}$
TOL data set, eocyte clade as outgroup						
<i>Homo sapiens</i> – <i>Artemia salina</i> . . . . .	–0.108	(0.063)	0.16	1, 1	–0.02	–0.2
<i>H. sapiens</i> – <i>Zea mays</i> . . . . .	0.195	(0.070)	–0.91	2, 1	–0.32	0.6
<i>Saccharomyces cerevisiae</i> – <i>Z. mays</i> . . . . .	–0.443	(0.079)	–6.22	1, 2	–0.32	5.9
<i>Aspergillus nidulans</i> – <i>Z. mays</i> (chl.) . . . . .	–0.270	(0.097)	–1.07	1, 3	–0.62	0.5
<i>Z. mays</i> (chl.)– <i>Bacillus subtilis</i> . . . . .	0.146	(0.060)	–0.45	1, 1	–0.02	0.4
<i>B. subtilis</i> – <i>Escherichia coli</i> . . . . .	–0.323	(0.079)	–2.98	2, 2	–0.32	2.7
<i>E. coli</i> – <i>Agrobacterium tumefaciens</i> . . . . .	0.122	(0.070)	0.09	1, 1	–0.02	–0.1
<i>Halobacterium volacanii</i> – <i>Halococcus morrhuae</i> . . . . .	–0.016	(0.047)	0.90	1, 1	–0.02	–0.9
<i>H. volacanii</i> – <i>Halobacterium halobium</i> . . . . .	0.071	(0.055)	0.49	2, 1	–0.32	–0.8
<i>S. cerevisiae</i> – <i>H. halobium</i> . . . . .	0.705	(0.128)	–6.14	2, 3	–0.45	5.7
<i>H. halobium</i> – <i>A. nidulans</i> . . . . .	–0.643	(0.122)	–5.11	2, 2	–0.32	4.8
Joint log-densities . . . . .			–31.66		0.18	31.8
Eukaryotes only . . . . .			–14.18		–0.18	14.0
Primate data set, lemur as outgroup						
Human–chimpanzee . . . . .	–0.070	(0.054)	0.51	1, 1	–0.02	–0.5
Chimpanzee–gorilla . . . . .	0.078	(0.062)	0.47	2, 1	–0.32	–0.8
Gorilla–orangutan . . . . .	0.076	(0.078)	0.50	2, 1	–0.32	–0.8
Orangutan–gibbon . . . . .	0.105	(0.089)	0.35	2, 1	–0.32	–0.7
Gibbon–macaque . . . . .	–0.047	(0.113)	0.51	2, 1	–0.32	–0.8
Macaque–squirrel monkey . . . . .	0.160	(0.132)	0.16	2, 1	–0.32	–0.5
Squirrel monkey–tarsier . . . . .	0.398	(0.141)	–1.28	2, 1	–0.32	1.0
Joint log-densities . . . . .			–1.61		–0.33	1.3
Anthropoids only . . . . .			1.89		–0.46	–2.4
HIV data set, MAL as outgroup						
SF2-OYI . . . . .	–0.013	(0.011)	1.27	1, 1	–0.02	–1.3
MN-ALA1 . . . . .	0.011	(0.010)	1.34	1, 1	–0.02	–1.4
JRCSF-JRFL . . . . .	0.030	(0.008)	–1.44	1, 1	–0.02	1.4
MN-JRCSF . . . . .	–0.004	(0.012)	1.50	2, 2	–0.32	–1.8
NY5-MN . . . . .	–0.017	(0.012)	1.08	1, 3	–0.62	–1.7
SF2-RF . . . . .	–0.011	(0.012)	1.33	3, 1	–0.62	–2.0
SF2-NY5 . . . . .	0.032	(0.014)	0.20	2, 2	–0.32	–0.5
ELI-ZZZ6 . . . . .	0.028	(0.010)	0.01	1, 1	–0.02	0.0
SF2-ELI . . . . .	0.276	(0.036)	–11.42	4, 2	–0.62	10.8
Joint log-densities . . . . .			–12.39		–0.08	12.3
B subtypes only . . . . .			3.64		–0.04	–3.7

NOTE.—Posterior means and standard deviations of molecular-clock height differences ( $\Delta_{ij}$ ), posterior and prior log<sub>10</sub> ordinates for  $\Delta_{ij} = 0$ , and log<sub>10</sub> Bayes factors against a molecular clock restriction conditional on mode topology under TN93 are shown. Traversed nodes are the numbers of nodes connecting the two taxa and their most common recent ancestor. The final rows for each example considers  $\Delta_{ij}$  jointly.

In allowing the sampler to explore the posterior across topology models, we overcome a shortfall of traditional analysis used to compare different continuous-time Markov evolutionary models. One can use a likelihood ratio test by maximizing the likelihood of the general model and the likelihood of the restricted model conditional on the same topology; however, the topologies that maximize the likelihood may differ under the two models. Then, general and restricted evolutionary models are no longer nested, and formal inference under a likelihood ratio test is no longer possible. In effect, our reversible jump MCMC sampler integrates out the nonnested portions of the parameter space. The Bayesian approach also allows us to effectively incorporate the uncertainty in the topology into the variance of the parameter estimates. Frequentist inference is forced to condition on topology and therefore underestimates the uncertainty.

The TOL example offers strong evidence against the universal appropriateness of a molecular clock; however, the anthropoids and subtype B isolates demonstrate that a local molecular clock for closely related taxa is a reasonable model. This finding is quite insensitive to prior choice. A molecular clock was originally employed in MCMC methods for evolutionary reconstruction to reduce computation (e.g., Mau and Newton 1997; Yang and Rannala 1997; Mau, Newton, and Larget 1999; Li, Pearl, and Doss 2000), but numerous examples of restriction violations exist (Ayala, Barrio, and Kwiatkowski 1996; Leitner et al. 1996; Hillis, Mable, and Moritz 1996; Simon et al. 1996; Holmes, Pybus, and Harvey 1999; Richman and Kohn 1999). Larget and Simon (1999) show that eliminating the molecular clock by doubling the number of estimable branch lengths does not produce an intractable problem; we extend the computation to allow for multiple sets of branch lengths that

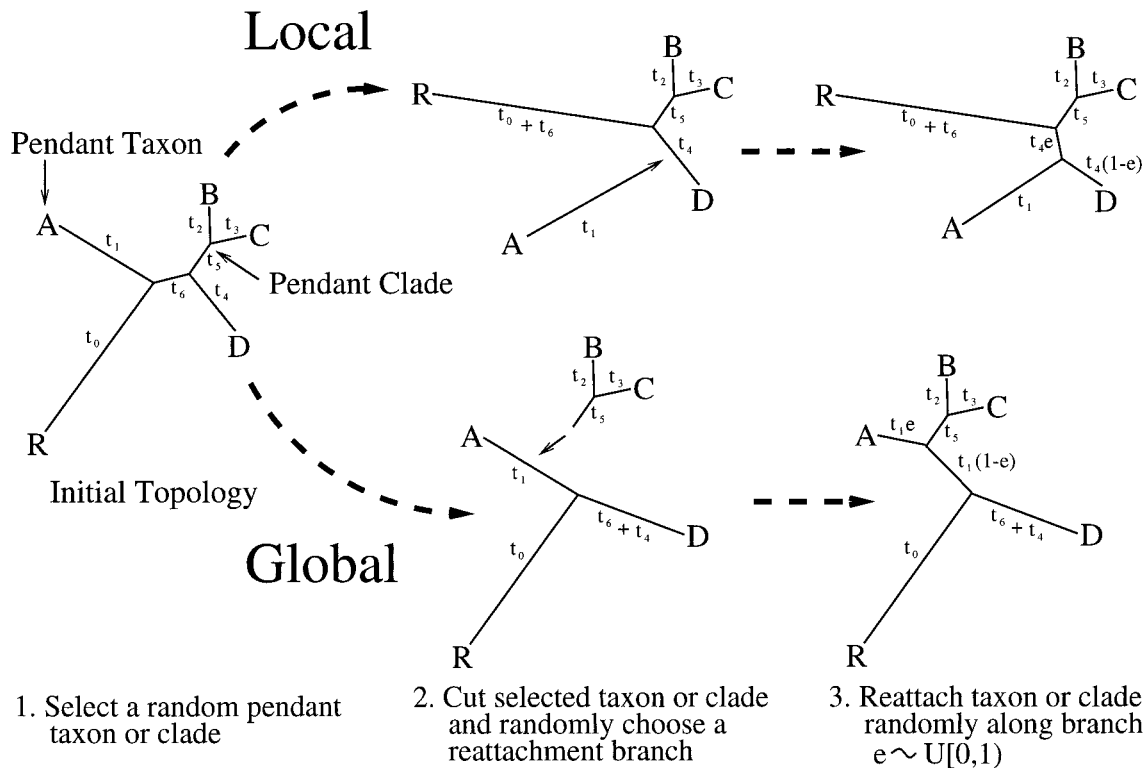


FIG. 5.—Local and global pendant leaf algorithms.

are not constrained by a molecular clock. In doing so, we further the frameworks of Thorne, Kishino, and Painter (1998) and Huelsenbeck, Larget, and Swofford (2000) in several important ways. Thorne, Kishino, and Painter (1998) introduce a Bayesian approach that does not impose a molecular clock by first assuming that the true relationship of the taxa under study is known with complete certainty and employing empirical Bayes priors that use the data twice. However, they do not provide a statistical test of the appropriateness of a molecular clock. Huelsenbeck, Larget, and Swofford (2000) continue to assume that the true relationship is known and formulate a likelihood ratio test that may become difficult to interpret when the data are sparse. We overcome both shortfalls by providing a framework to statistically test the appropriateness of a molecular clock while not having to condition on a known a priori topology and simultaneously make inference about the appropriate parameterization of the infinitesimal rate matrix. Additionally, pairwise diagnostic Bayes factors we propose allow the researcher to conveniently identify portions of an evolutionary history that violate a molecular clock and portions that support a local molecular clock. On the other hand, Huelsenbeck, Larget, and Swofford (2000) allow for the estimation of divergence times, while our approach does not eliminate the confounding of time and evolutionary rate.

To provide both large and small jumps among so many branch lengths, we choose a 50/50 mixture of two transition kernels—the first updating all branch lengths simultaneously using a reflective normal driver with small variance, and the second randomly selecting and

updating one branch length using a driver with large variance. This mixture removed initially poor convergence in the HIV data set that had small branch lengths as compared with the other two examples. We find quick convergence and sufficient mixing for up to at least 15-taxon topologies without a molecular clock.

#### APPENDIX A

##### Transition Kernels

##### Joint Tree and Branch Length Proposals

We propose new trees and branch lengths using a mixture of three transition kernels, local pendant leaf, global pendant leaf, and leaf permutator, that perform both small and large scale rearrangements with mixing probabilities of 1/3 each.

##### Pendant Leaves (local and global variants)

Pendant leaf-type algorithms are described by Li, Pearl, and Doss (2000). We employ two variants, one (local) that moves a single taxon at a time from its present location in a topology to a new location forming a new topology, and one (global) that moves an entire clade of taxa at a time. Figure 5 illustrates these rearrangements.

These drivers are asymmetric. Let  $t_y$  be the length of the newly selected branch before the pendant leaf or leaves bisect it, and let  $t_x$  be the length of the original branch bisected by the pendant leaf or leaves; then, we choose a random location uniformly along  $t_y$ . The ratio of the densities of the transition kernel going from state  $x$  to state  $y$  over that from state  $y$  to state  $x$  is  $t_x/t_y$ .

### Leaf Permutor

This algorithm randomly selects an internal node ( $I$ ) and then randomly permutes all leaves which are descendants of  $I$ . This driver is symmetric, but it is not ergodic for  $N \geq 6$  taxa. However, our mixture of the three drivers remains ergodic as long as one member of the mixture is itself ergodic. Li, Pearl, and Doss (2000) show that pendant leaf algorithms are ergodic.

### Branch Length, Stationary Distribution, Evolutionary Parameters, and Hyperparameter Proposals

We propose a new set  $T$  using a 50/50 mixture of two drivers. The first is a multivariate normal driver of dimension  $2N - 3$ , and the second randomly selects and updates one  $t_b$  from  $T$  using a univariate normal driver. As all branch lengths are restricted to be nonnegative, both drivers update their respective elements of  $T$  by reflecting about 0, such that  $t_b^* = |t_b + \epsilon_j|$ , where  $\epsilon_j \sim N(0, \sigma_{T,j}^2)$  for  $j = 1, 2$ . We set  $\sigma_{T,1}^2 = 2.5 \times 10^{-4}$  and  $\sigma_{T,2}^2 = 0.1$ .

We propose a new  $\pi$  using a trivariate normal driver with diagonal variance  $5 \times 10^{-5}$  and constrain  $\sum \pi_m = 1$ , such that  $\pi_m^* = |\pi_m + \epsilon_{\pi,m}|$  for  $m \in (A, G, C)$ , where  $\epsilon_{\pi,m} \sim N(0, \sigma_{\pi}^2)$ , and  $\pi_{T/U}^* = 1 - \pi_A^* - \pi_G^* - \pi_C^*$ . We reject all proposals in which  $\pi_m^* \notin [0, 1]$  for all  $m \in (A, G, C, T/U)$ . More efficient drivers for  $\pi$  exist, and we are evaluating several different possibilities. We also propose a new  $\alpha$  and a new  $\gamma$  using two independent normal drivers with variance  $5.0 \times 10^{-5}$  that are reflected about both ends of the support space  $[0, 1]$ . Finally, we propose a new  $\mu$  using a normal driver with variance 0.1 and, again, we reflect about 0 such that  $\mu^* = |\mu + \epsilon|$ . We choose all variances to allow for 20%–30% acceptance rates in each Metropolis-Hastings proposal (Gelman, Roberts, and Gilks 1996).

#### APPENDIX B

### Analytic Solutions of the Marginal Priors

Here, we determine the exact distribution of  $\Delta_{ij}$  for two-taxon rooted topologies, the joint prior density  $q(\Delta_1, \Delta_2)$  at  $(0, 0)$  for three-taxon rooted topologies, and the univariate, marginal density  $q(\Delta_{ij})$  at 0 for any arbitrary pair of taxa in an  $N$ -taxon topology given an outgroup using characteristic functions.

#### Two Taxa

One coalescent height difference,  $\Delta = t_1 - t_2$ , exists for two taxa  $A$  and  $B$  related by the rooted topology  $(A : t_1, B : t_2)$ , where  $t_1$  and  $t_2$  are the branch lengths connecting  $A$  and  $B$  to their MRCA. We find the distribution of  $\Delta|\mu$  under the prior by letting  $\phi(s)$  be its characteristic function. Then,  $\phi(s) = E\{e^{is\Delta}\} = 1/(1 + \mu^2 s^2)$ , as  $t_i$  are iid Exponential( $\mu$ ) random variables. By the uniqueness theorem (Feller 1971),  $\Delta|\mu \sim \text{Double-exponential}(\mu)$ , where  $q_{\Delta|\mu}(x) = \exp(-|x|/\mu)/(2\mu)$  and  $q_{\Delta|\mu}(0) = 1/(2\mu)$ .

#### Three Taxa

A sufficient set of coalescent height differences for three taxa,  $A, B$ , and  $C$ , related by the rooted topology  $(A : t_4, (B : t_1, C : t_2) : t_3)$  are

$$\Delta_1 = t_1 - t_2 \quad \text{and} \quad \Delta_2 = t_4 - t_3 - t_1, \quad (15)$$

where  $t_i$  are branch lengths.

Let  $\phi(s_1, s_2)$  be the joint characteristic function of  $(\Delta_1, \Delta_2)|\mu$ ; then,

$$\begin{aligned} \phi(s_1, s_2) &= E\{e^{i(s_1\Delta_1 + s_2\Delta_2)}\} \\ &= \frac{1}{1 - i\mu(s_1 - s_2)} \frac{1}{1 + i\mu s_1} \frac{1}{1 + i\mu s_2} \frac{1}{1 - i\mu s_2}, \end{aligned} \quad (16)$$

as  $t_i$  are iid Exponential( $\mu$ ) random variables.

To recover  $q(\Delta_1 = 0, \Delta_2 = 0|\mu)$ , we take the inverse Fourier transform of equation (16) evaluated at  $(\Delta_1, \Delta_2) = (0, 0)$ ,

$$\begin{aligned} q(\Delta_1 = 0, \Delta_2 = 0|\mu) &= \left(\frac{1}{2\pi}\right)^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-i(s_1 \times 0 + s_2 \times 0)} \phi(s_1, s_2) ds_1 ds_2. \end{aligned} \quad (17)$$

We integrate equation (17) first with respect to  $s_1$ .

Let

$$\begin{aligned} g(s_2) &= \int_{-\infty}^{\infty} \frac{1}{1 - i\mu(s_1 - s_2)} \frac{1}{1 + i\mu s_1} ds_1 \\ &= \frac{\pi}{\mu} \frac{1}{1 + i\frac{\mu}{2}s_2}, \end{aligned} \quad (18)$$

calculated by expanding out the integrand and completing the square. Substituting the solution of equation (18) back into equation (17) results in

$$\begin{aligned} q(\Delta_1 = 0, \Delta_2 = 0|\mu) &= \frac{1}{4\pi\mu} \int_{-\infty}^{\infty} \frac{1}{1 + i\frac{\mu}{2}s_2} \frac{1}{1 + \mu^2 s_2^2} ds_2. \end{aligned} \quad (19)$$

Expanding equation (19) by partial fractions,

$$\begin{aligned} q(\Delta_1 = 0, \Delta_2 = 0|\mu) &= \frac{1}{4\pi\mu} \left( \frac{2}{3} \int_{-\infty}^{\infty} \frac{2 - i\mu s_2}{1 + \mu^2 s_2^2} ds_2 - \frac{1}{3} \int_{-\infty}^{\infty} \frac{1}{1 + i\frac{\mu}{2}s_2} ds_2 \right) \\ &= \frac{1}{6\mu^2}. \end{aligned} \quad (20)$$

#### Marginal Constraint for Any Pair of Taxa

Here, we determine the prior density of a single  $\Delta_{ij}$  at 0 for any arbitrary pair of taxa. Let  $n_1$  and  $n_2$  count the numbers of internal nodes traversed from taxon  $i$  and taxon  $j$ , respectively, to their MRCA, including the

MRCA itself. Then,  $h_1$  and  $h_2$ , the sums of branch lengths from the taxa to the MRCA, are independent Gamma( $n_1, 1/\mu$ ) and Gamma( $n_2, 1/\mu$ ) random variables. Let  $\Delta_{ij} = h_1 - h_2$ . Using characteristic functions, we determine the distribution of  $\Delta_{ij}|\mu$ . Let  $\phi(s)$  be a characteristic function of  $\Delta_{ij}|\mu$ ; then,

$$\begin{aligned}\phi(s) &= E(e^{-is\Delta_{ij}}) = E(e^{-ish_1+ish_2}) \\ &= (1 - i\mu s)^{-n_1}(1 + i\mu s)^{-n_2} \\ &= (1 + \mu^2 s^2)^{-n_2}(1 - i\mu s)^{-(n_1-n_2)},\end{aligned}\quad (21)$$

where, without loss of generality, we assume  $n_1 \geq n_2$ .

By the uniqueness theorem (Feller 1971),

$$\Delta_{ij}|\mu \sim \sum_{k=1}^{n_2} D_k + G,$$

where

$$D_k \sim \text{Double-exponential}(\mu),$$

$$f(x) = \frac{1}{2\mu} e^{-|x|/\mu}, \quad \text{and}$$

$$G \sim \text{Gamma}\left(n_1 - n_2, \frac{1}{\mu}\right).\quad (22)$$

We take the inverse Fourier transform of equation (21) evaluated at  $\Delta_{ij} = 0$  to get the density  $q(\Delta_{ij}|\mu)$  at 0. For  $n_1 = n_2$ , we find by direct integration

$$q(\Delta_{ij} = 0|\mu) = \frac{1}{\mu} \frac{\Gamma(n_2 - 1/2)}{2\sqrt{\pi}(n_2 - 1)!}.\quad (23)$$

For  $n_1 > n_2$ , we use the recursion relationship

$$\begin{aligned}&\int_{-\infty}^{\infty} (1 - i\mu s)^{-n_1}(1 + i\mu s)^{-n_2} ds \\ &= \frac{n_1 + n_2 - 2}{2(n_1 - 1)} \int_{-\infty}^{\infty} (1 - i\mu s)^{-(n_1-1)}(1 + i\mu s)^{-n_2} ds\end{aligned}\quad (24)$$

$n_1 - n_2$  times to reduce cases down to integrals of the form solved in equation (23), where  $n_1 = n_2$ . This results in

$$\begin{aligned}q(\Delta_{ij} = 0|\mu) &= \frac{1}{\mu} \frac{(n_1 + n_2 - 2)! \Gamma(n_2 - 1/2)}{2^{(n_1-n_2+1)} \sqrt{\pi} (n_1 - 1)! (2n_2 - 2)!}.\end{aligned}\quad (25)$$

### Integrating Out $\mu$ to Recover Marginal Densities

In two-taxon cases, we recover the exact distribution of  $\Delta$  ( $q_\Delta(x)$ ) and its density at 0 by forming the joint distribution  $q(x, \mu) = q(x|\mu)q(\mu)$  and integrating out  $\mu$ ,

$$\begin{aligned}q_\Delta(x) &= \frac{1}{2} \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \mu^{-(\alpha+2)} e^{-(1/\mu)(\beta+|x|)} d\mu \\ &= \frac{1}{2} \alpha \beta^\alpha (\beta + |x|)^{-(\alpha+1)}, \quad \text{and}\end{aligned}$$

$$q_\Delta(0) = \frac{\alpha}{2\beta}.\quad (26)$$

This argument can be generalized for the remaining cases where we obtain the marginal densities evaluated at 0 by one final integration of  $q(\Delta_{ij} = 0|\mu)$  with respect to  $\mu$ . Recalling that  $\mu \sim \text{Inv-gamma}(2.1, 1.1)$ ,  $q(\Delta = 0) = 0.955$  and  $q(\Delta_1 = 0, \Delta_2 = 0) = 0.897$ .

### Acknowledgments

We thank James Lake for supplying the aligned sequences used in the TOL example and Karin Dorman and John Boscardin for their helpful criticism. M.A.S. was supported by a predoctoral fellowship from the Howard Hughes Medical Institute. J.S.S. was partially supported by USPHS grants AI28697 and CA16042.

### LITERATURE CITED

- AYALA, F. J., E. BARRIO, and J. KWIATOWSKI. 1996. Molecular clock or erratic evolution? A tale of two genes. *Proc. Natl. Acad. Sci. USA* **93**:11729–11734.
- BHATTACHARYA, D., and L. MEDLIN. 1995. The phylogeny of plastids: a review based on comparisons of small-subunit ribosomal RNA coding regions. *J. Phycol.* **31**:489–498.
- BROOKS, S. P., and P. GUIDICI. 1999. Convergence assessment for reversible jump MCMC simulations. Pp. 733–742 in J. BERNARDO, J. BERGER, A. P. DAWID, and A. F. M. SMITH, eds. *Bayesian Statistics 6*. Oxford University Press, Cambridge, Mass.
- BROWN, W. M., E. M. PRAGER, A. WANG, and A. C. WILSON. 1982. Mitochondrial DNA sequences of primates, tempo and mode of evolution. *J. Mol. Evol.* **18**:225–239.
- CRANDALL, K. A., ed. 1999. *The evolution of HIV*. Johns Hopkins University Press, Baltimore, Md.
- DURBIN, R., S. EDDY, A. KROGH, and G. MITCHINSON. 1998. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, England.
- FELLER, W. 1971. *An introduction to probability theory and its applications*. Vol. 2, 2nd edition. John Wiley and Sons, New York.
- FELSENSTEIN, J. 1978. The number of evolutionary trees. *Syst. Zool.* **27**:27–33.
- . 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**:368–376.
- GELMAN, A., G. O. ROBERTS, and W. R. GILKS. 1996. Efficient Metropolis jumping rules. Pp. 599–608 in J. M. BERNARDO, J. O. BERGER, A. P. DAWID, and A. F. M. SMITH, eds. *Bayesian Statistics 5*. Oxford University Press, Oxford, England.
- GILKS, W. R., S. RICHARDSON, and D. J. SPIEGELHALTER. 1996. *Markov chain Monte Carlo*. Chapman and Hall, New York.
- GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36**:182–198.
- GREEN, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**:711–732.
- 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. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**:97–109.
- HAYASAKA, K., K. T. GOJOBORI, and S. HORAI. 1988. Molecular phylogeny and evolution of primate mitochondrial DNA. *Mol. Biol. Evol.* **5**:626–644.

- HILLIS, D. M., B. K. MABLE, and C. MORITZ. 1996. Applications of molecular systematics: the state of the field and a look to the future. Pp. 515–543 in D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. *Molecular systematics*. 2nd edition. Sinauer, Sunderland, Mass.
- HOLMES, E. C., O. G. PYBUS, and P. H. HARVEY. 1999. The molecular population dynamics of HIV-1. Pp. 177–207 in K. A. CRANDALL, ed. *The evolution of HIV*. Johns Hopkins University Press, Baltimore, Md.
- HUELSENBECK, J. P., and B. RANNALA. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**:227–232.
- HUELSENBECK, J. P., B. LARGET, and D. SWOFFORD. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* **154**:1879–1892.
- JEFFREYS, H. 1998. *Theory of probability*. Oxford classic texts in the physical sciences. 3rd edition. Oxford University Press, New York.
- JUKES, T., and C. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 in H. N. MUNRO, ed. *Mammalian protein metabolism*. Academic Press, New York.
- KASS, R. E., and A. E. RAFTERY. 1995. Bayes factors and model uncertainty. *J. Am. Stat. Assoc.* **90**:773–795.
- KIMURA, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**:111–120.
- KORBER, B., B. HAHN, B. FOLEY, J. W. MELLORS, T. LEITNER, G. MYERS, F. MCCUTCHAN, and C. L. KUIKEN, eds. 1997. *Human retroviruses and AIDS 1997: a compilation and analysis of nucleic acid and amino acid sequences*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. (<http://hiv-web.lanl.gov>).
- 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.
- . 1998. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**:429–434.
- LAKE, J. A. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* **331**:184–186.
- LANGE, K. 1997. *Mathematical and statistical methods for genetic analysis*. Springer, New York.
- LARGET, B., and D. L. SIMON. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* **16**:750–759.
- LEITNER, T., D. ESCANILLA, C. FRANZN, M. UHLN, and J. ALBERT. 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl. Acad. Sci. USA* **93**:10864–10869.
- LI, S., D. K. PEARL, and H. DOSS. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. *J. Am. Stat. Assoc.* **95**:493–508.
- LOFTSGAARDEN, D. O., and C. P. QUESENBERY. 1965. A non-parametric estimate of a multivariate density function. *Ann. Math. Stat.* **36**:1049–1051.
- MCCABE, K. M., G. KHAN, Y. H. ZHANG, E. O. MASON, and E. R. MCCABE. 1995. Amplification of bacterial DNA using highly conserved sequences: automated analysis and potential for molecular triage of sepsis. *Pediatrics* **95**:165–169.
- MARGULIS, L. 1981. *Symbiosis in cell evolution: life and its environment on the early earth*. W. H. Freeman, San Francisco.
- MAU, B., and M. A. NEWTON. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J. Comput. 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. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, and E. TELLER. 1953. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**:1087–1092.
- NAVIDI, W. C., G. A. CHURCHILL, and A. VON HAESLER. 1991. Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. *Mol. Biol. Evol.* **8**:128–143.
- . 1993. Phylogenetic inference: linear invariants and maximum likelihood. *Biometrics* **49**:543–55.
- NERURKAR, V. R., H. T. NGUYEN, W. M. DASHWOOD, P. R. HOFFMANN, C. YIN, D. M. MORENS, A. H. KAPLAN, R. DETELS, and R. YANAGIHARA. 1996. HIV type 1 subtype E in commercial sex workers and injection drug users in southern Vietnam. *AIDS Res. Hum. Retroviruses* **12**:841–843.
- RANNALA, B., and Z. YANG. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* **43**:304–311.
- RELMAN, D. A., T. M. SCHMIDT, A. GAJADHAR, M. SOGIN, J. CROSS, K. YODER, O. SETHABUTR, and P. ECHEVERRIA. 1996. Molecular phylogenetic analysis of *Cyclospora*, the human intestinal pathogen, suggests that it is closely related to *Eimeria* species. *J. Infect. Dis.* **173**:440–445.
- RICHMAN, A. D., and J. R. KOHN. 1999. Self-incompatibility alleles from *Physalis*: implications for historical inference from balanced genetic polymorphisms. *Proc. Natl. Acad. Sci. USA* **96**:168–172.
- RUDOLPH, K. M., A. J. PARKINSON, C. M. BLACK, and L. W. MAYER. 1993. Evaluation of polymerase chain reaction for diagnosis of pneumococcal pneumonia. *J. Clin. Microbiol.* **31**:2661–2666.
- RZHETSKY, A., and T. SITNIKOVA. 1996. When is it safe to use an oversimplified substitution model in tree-making? *Mol. Biol. Evol.* **13**:1255–1265.
- SIMON, C., L. NIGRO, J. SULLIVAN, K. HOLSINGER, A. MARTIN, A. GRAPPUTO, A. FRANKE, and C. MCINTOSH. 1996. Large differences in substitutional pattern and evolutionary rate of 12S ribosomal RNA genes. *Mol. Biol. Evol.* **13**:923–932.
- SINSHEIMER, J. S., J. A. LAKE, and R. J. LITTLE. 1996. Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics* **52**:193–210.
- SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inferences. Pp. 407–514 in D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. *Molecular systematics*. 2nd edition. Sinauer, Sunderland, Mass.
- TAMURA, K., and M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**:512–526.
- THORNE, J. L., H. KISHINO, and I. S. PAINTER. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**:1647–1657.
- TIERNEY, L. 1994. Markov chains for exploring posterior distributions (with discussion). *Ann. Stat.* **22**:1701–1762.
- VERDINELLI, I., and L. WASSERMAN. 1995. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Am. Stat. Assoc.* **90**:614–618.
- WHELAN, S., and N. GOLDMAN. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* **16**:1292–1299.
- YANG, Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* **139**:993–1005.
- 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 5, 2001