

# Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data

Alexei J. Drummond,<sup>\*,1</sup> Geoff K. Nicholls,<sup>†</sup> Allen G. Rodrigo<sup>\*</sup> and Wiremu Solomon<sup>‡</sup>

<sup>\*</sup>School of Biological Sciences, <sup>†</sup>Department of Mathematics and <sup>‡</sup>Department of Statistics, University of Auckland 1001, Auckland, New Zealand

Manuscript received December 15, 2001

Accepted for publication March 12, 2002

## ABSTRACT

Molecular sequences obtained at different sampling times from populations of rapidly evolving pathogens and from ancient subfossil and fossil sources are increasingly available with modern sequencing technology. Here, we present a Bayesian statistical inference approach to the joint estimation of mutation rate and population size that incorporates the uncertainty in the genealogy of such temporally spaced sequences by using Markov chain Monte Carlo (MCMC) integration. The Kingman coalescent model is used to describe the time structure of the ancestral tree. We recover information about the unknown true ancestral coalescent tree, population size, and the overall mutation rate from temporally spaced data, that is, from nucleotide sequences gathered at different times, from different individuals, in an evolving haploid population. We briefly discuss the methodological implications and show what can be inferred, in various practically relevant states of prior knowledge. We develop extensions for exponentially growing population size and joint estimation of substitution model parameters. We illustrate some of the important features of this approach on a genealogy of HIV-1 envelope (*env*) partial sequences.

ONE of the most significant developments in population genetics modeling in recent times was the introduction of *coalescent* or genealogical methods (KINGMAN 1982a,b). The coalescent is a stochastic process that provides good approximations to the distribution of ancestral histories that arise from classical forward-time models such as the Fisher-Wright (FISHER 1930; WRIGHT 1931) and Moran population models. The explicit use of genealogies to estimate population parameters allows the nonindependence of sampled sequences to be accounted for. ("Genealogy" and "tree" are used interchangeably throughout. In both cases we are referring to a collection of edges, nodes, and node times that together completely specify a rooted history.) Many coalescent-based estimation methods focus on a single genealogy (FU 1994; NEE *et al.* 1995; PYBUS *et al.* 2000) that is typically obtained using standard phylogenetic methods. However, there is often considerable uncertainty in the reconstructed genealogy. To allow for this uncertainty it is necessary to compute the average likelihood of the population parameters of interest. The calculation involves integrating over genealogies distributed according to the coalescent (GRIFFITHS and TAVARE 1994; KUHNER *et al.* 1995). We can carry out this integration for some models of interest, using Monte Carlo methods. Importance-sampling algorithms have been developed to estimate the population param-

eter  $\Theta = 2N_e\mu$  (GRIFFITHS and TAVARE 1994; STEPHENS and DONNELLY 2000), migration rates (BAHLO and GRIFFITHS 2000), and recombination (GRIFFITHS and MARJORAM 1996; FEARNHEAD and DONNELLY 2001). Metropolis-Hastings Markov chain Monte Carlo (MCMC; METROPOLIS *et al.* 1953; HASTINGS 1970) has been used to obtain sample-based estimates of  $\Theta$  (KUHNER *et al.* 1995), exponential growth rate (KUHNER *et al.* 1998), migration rates (BEERLI and FELSENSTEIN 1999, 2001), and recombination (KUHNER *et al.* 2000).

In addition to developments in coalescent-based population genetic inference, sequence data sampled at different times are now available from both rapidly evolving viruses, such as human immunodeficiency virus (HIV; HOLMES *et al.* 1992; WOLINSKY *et al.* 1996; RODRIGO *et al.* 1999; SHANKARAPPA *et al.* 1999), and from ancient DNA sources (HANNI *et al.* 1994; LEONARD *et al.* 2000; LOREILLE *et al.* 2001; BARNES *et al.* 2002; LAMBERT *et al.* 2002). This temporally spaced data provides the potential to observe the accumulation of mutations over time and thus estimate mutation rate (DRUMMOND and RODRIGO 2000; RAMBAUT 2000). In fact, it is even possible to estimate variation in the mutation rate over time (DRUMMOND *et al.* 2001). This leads naturally to the more general problem of simultaneous estimation of population parameters and mutation parameters from temporally spaced sequence data (RODRIGO and FELSENSTEIN 1999; RODRIGO *et al.* 1999; DRUMMOND and RODRIGO 2000; DRUMMOND *et al.* 2001).

In this article we estimate population and mutation parameters, dates of divergence, and tree topology from temporally spaced sequence data, using sample-based

<sup>1</sup>Corresponding author: Department of Statistics and Department of Zoology, University of Oxford, South Parks Rd., Oxford, OX1 3PS, United Kingdom. E-mail: alexei.drummond@zoology.oxford.ac.uk

Bayesian inference. The important novelties in the inference are the data type (*i.e.*, temporally sampled sequences), the relatively large number of unknown model parameters, and the MCMC sampling procedures, not the Bayesian framework itself. The coalescent gives the expected frequency with which any particular genealogy arises under the Fisher-Wright population model. The coalescent may then be treated either as part of the observation process defining the likelihood of population parameters or as the prior distribution for the unknown true genealogy. In either case we must integrate the likelihood over the state space of the coalescent. Bayesian and purely likelihood-based population genetic inference use the same reasoning and software up to the point where prior distributions are given for the parameters of the coalescent and mutation processes.

Are there then any important difficulties or advantages in a Bayesian approach over a purely likelihood-based approach? The principal advantage is the possibility of quantifying the impact of prior information on parameter estimates and their uncertainties. The new difficulty is to represent different states of prior knowledge of the parameters of the coalescent and mutation processes as probability densities. However, such prior elicitation is often instructive. In the absence of prior information, researchers frequently choose to use non-informative/improper priors for the parameters of interest. Such an approach may be problematic and can result in improper posterior distributions. There exist a number of important cases in the literature in which knowledgeable authors inadvertently analyze a meaningless, improper posterior distribution. Why then do we choose to treat improper priors in this article? We are developing and testing inferential and sampling methods. These methods become more difficult as the amount of information in the prior is reduced. The sampling problem becomes significantly more difficult. We therefore treat the “worst case” prior that might naturally arise. Since this prior is improper, we are obliged to check that the posterior is proper. However, when confronted with a specific analysis, detailed biological knowledge should be encoded in the prior distributions wherever possible.

Although Bayesian reasoning has frequently been applied to phylogenetic inference (YANG and RANNALA 1997; THORNE *et al.* 1998; MAU *et al.* 1999; HUELSENBECK *et al.* 2000), it has thus far been the exception in population genetic inference (WILSON and BALDING 1998).

In this article, we begin with a description of the models we use. We then give the overall structure of the inferential framework, followed by an overview of how MCMC is carried out. We mention extensions of the basic inference that allow for (1) deterministically varying populations and (2) estimation of substitution parameters. Finally, we illustrate our methods with a group of studies of a sample of HIV-1 envelope (*env*) sequences

and a second group of studies of synthetic sequence data.

**Kingman coalescent with temporally offset leaves:** In this section we define the coalescent density for the constant-sized Fisher-Wright population model. In EXTENSIONS we give the corresponding density for the case of a population with deterministic exponential growth. It is assumed genealogies are realized by the Kingman coalescent process. Our time units in this article are “calendar units before the present” [*e.g.*, days before present (BP)], where the present is the time of the most recent leaf and set to zero. Let  $\rho$  denote the number of calendar units per generation and  $\theta = N_e\rho$ . The scale factor  $\theta$  converts “coalescent time” to calendar time and is one of two key objects of our inference. Note that we do not estimate  $\rho$  and  $N_e$  separately, only their product.

Consider a rooted binary tree  $g$  with  $n$  leaf nodes and  $n - 1$  ancestral nodes. For node  $i$ , let  $t_i$  denote the age of that node in calendar units. Node labels are numerically increasing with age so  $i > j$  implies  $t_i \geq t_j$ . Let  $I$  denote the set of leaf node labels and let  $Y$  denote the set of ancestral node labels. There is one leaf node  $i \in I$  associated with each individual in the data. These individuals are selected, possibly at different times, from a large background population. An edge  $\langle i, j \rangle$ ,  $i > j$  of  $g$  represents an ancestral lineage. Going back in time, an ancestral node  $i \in Y$  corresponds to a *coalescence* of two ancestral lineages. The root node, with label  $i = 2n - 1$ , represents the most recent common ancestor (MRCA) of all leaves. Let  $t_l$  be the times of the leaves and  $t_Y$  be the divergence times of the ancestral nodes. Let  $E_g$  denote the edge set of  $g$ , so that  $g = (E_g, t_Y)$  specifies a realization of the coalescent process. For given  $n$  and  $t_l$ , let  $\Gamma$  denote the class of all coalescent trees  $(E_g, t_Y)$  with  $n$  leaf nodes having fixed ages  $t_l$ . The ages  $t_Y$  are subject to the obvious parent-child age order constraint. The element of measure in  $\Gamma$  is  $dg = dt_{n+1} \dots dt_{2n-1}$  with counting measure over distinct topologies associated with the distinguishable leaves.

The probability density for a tree,  $f_G(g|\theta)$ ,  $g \in \Gamma$  is computed as follows. Let  $k_i$  denote the number of lineages present in the interval of time between the node  $i - 1$  and the node  $i$ . The coalescent process generates  $g = (E_g, t_Y)$  with probability density

$$f_G(g|\theta) = \frac{1}{\theta^{n-1}} \cdot \prod_{i=2}^{2n-1} e^{(-k_i(k_i-1)/2\theta)(t_i-t_{i-1})}. \quad (1)$$

The interpretation is as follows. Fix a time  $t$  and suppose  $k$  lineages are present at that time. A coalescence event between any of the  $k(k - 1)/2$  pairs of distinguished lineages occurs at instantaneous rate  $1/\theta$ . Given that two lineages coalesce at time  $t$ , the probability it was some particular pair is  $2/k(k - 1)$ . It follows that, in the time interval of length  $t_i - t_{i-1}$  preceding the time of a leaf node  $i \in I$ , “nothing” happens with probability  $e^{(-k_i(k_i-1)/2\theta)(t_i-t_{i-1})}$  and that the length of time,  $t - t_{i-1}$ ,

preceding coalescent node  $i \in Y$  is a random variable with density  $(k_i(k_i - 1)/2\theta) \cdot e^{(-k_i(k_i-1)/2\theta)(t_i - t_{i-1})}$ . Taking the product of these factors over all intervals  $[t_{i-1}, t_i]$ ,  $i = 2, 3, \dots, 2n - 1$ , we obtain Equation 1 (RODRIGO and FELSENSTEIN 1999).

**Mutation:** We use the standard finite-sites selection-neutral likelihood framework (FELSENSTEIN 1981) with a general time-reversible (GTR) substitution model (RODRIGUEZ *et al.* 1990). However, as we are considering genealogies in calendar units (or generations) as opposed to mutations we take some space to develop notation.

Associated with each leaf node  $i \in I$  there is a nucleotide sequence  $D_i = (D_{i,1}, D_{i,2}, \dots, D_{i,s}, \dots, D_{i,L})$  of some fixed length  $L$ , say. Nucleotide base characters  $D_{i,s}$ ,  $i \in I$ ,  $s = 1, 2, \dots, L$  take values in the set  $\mathbf{C} = \{A, C, G, T\}$ . An additional gap character,  $\phi$ , indicates missing data. Let  $D = (D_1, D_2, \dots, D_n)^T$  denote the  $n \times L$  matrix of sequences associated with the tree leaves, and let  $D_A$  denote the  $(n - 1) \times L$  matrix of unknown sequences associated with the ancestral nodes. The data are  $D$  together with  $t_b$ , that is, the  $n$  sequences observed in the leaf individuals and the  $n$  ages at which those individual sequences were taken. Let  $\mathbf{D} = \mathbf{C}^{(n-1)L}$  denote the set of all possible ancestral sequences. Consider a site  $s = 1, 2, \dots, L$  in the nucleotide sequence of an individual. The character at site  $s$  mutates in forward time according to a Poisson jump process with  $4 \times 4$  rate matrix  $Q$ . Here,  $Q_{ij}$  is the instantaneous rate for the transition from character  $i$  to character  $j$ , and  $A \leftarrow 1$ ,  $C \leftarrow 2$ ,  $G \leftarrow 3$ ,  $T \leftarrow 4$ . We assume mutations are independent between sites. Let  $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$  be a  $1 \times 4$  vector of base frequencies, corresponding to the stationary distribution of the mutation process,  $\pi Q = (0, 0, 0, 0)$ .

The matrix  $Q$  is parameterized in terms of a symmetric “relative rate” matrix  $R$ ,

$$R = \begin{bmatrix} & R_{A \leftarrow C} & R_{A \leftarrow G} & R_{A \leftarrow T} \\ R_{A \leftarrow C} & & R_{C \leftarrow G} & R_{C \leftarrow T} \\ R_{A \leftarrow G} & R_{C \leftarrow G} & & 1 \\ R_{A \leftarrow T} & R_{C \leftarrow T} & 1 & \end{bmatrix} \quad (2)$$

as

$$Q_{ij} = \frac{\pi_i R_{ij}}{\sum_k \pi_k \sum_{l \neq k} \pi_l R_{kl}}, \quad i \neq j$$

$$Q_{ij} = -\sum_{j \neq i} Q_{ij}. \quad (3)$$

The time units of the rate  $Q_{ij}$  have been chosen so that the mean number of mutations per unit time occurring at a site is equal to one. Let  $\mu$  give the mean number of mutations per calendar unit (*e.g.*, mutations per year) at a site.

The conversion factor  $\mu$  is the second of the two principal objects of our inference. In addition to  $\mu$ , the relative rates,  $R$ , may be estimated. We have found that wherever it is feasible to estimate the scale parameters  $\mu$  and  $\theta$ , our data are informative about the elements

of  $R$ . We return to inference for relative rates in EXTENSIONS.

We now write the likelihood for  $\mu$ . Consider an edge  $\langle i, j \rangle \in E_g$  of tree  $g$ . The individual associated with node  $j$  is a direct descendant of the individual associated with node  $i$ . However, the sequences  $D_i$  and  $D_j$  may differ if mutations have occurred in the interval. Let  $e^Q$  denote the  $4 \times 4$  matrix exponential of  $Q$ . In the standard finite-sites selection-neutral likelihood framework  $\Pr\{D_{j,s} = c' | D_{i,s} = c\} = [e^{-Q\mu(t_i - t_j)}]_{c,c'}$  for  $c \in \mathbf{C}$ . The probability for any particular set of sequences  $D$ ,  $D_A$  to be realized at the nodes of a given tree is

$$\Pr\{D, D_A | g, \mu\} = \prod_{\langle i,j \rangle \in E_g} \prod_{\substack{s=1 \\ D_{i,s} \neq \phi}}^L [e^{Q\mu(t_i - t_j)}]_{D_{i,s} D_{j,s}} \quad (4)$$

(in the above formula, compact notation is obtained by including in the product over edges an edge terminating at the root from an ancestor of infinite age). We may eliminate the unknown ancestral sequences  $D_A$  from the above expression by simply summing all  $D_A \in \mathbf{D}$ ,

$$\Pr\{D | g, \mu\} = \sum_{D_A \in \mathbf{D}} \Pr\{D, D_A | g, \mu\}. \quad (5)$$

It is feasible to evaluate this sum, using a pruning algorithm (FELSENSTEIN 1981).

**Bayesian inference for scale parameters:** We now consider Bayesian inference for scale parameters  $\mu$  and  $\theta$ . Both of these quantities take a real positive value. The joint posterior density,  $h_{M\Theta G}(\mu, \theta, g | D)$ , for the scale parameters and genealogy, is given in terms of the likelihood and coalescent densities above and two additional densities,  $f_M(\mu)$  and  $f_\Theta(\theta)$ . These functions quantify prior information about the scale parameters. Let  $Z$  be an unknown normalizing constant. The posterior is then

$$h_{M\Theta G}(\mu, \theta, g | D) = \frac{1}{Z} \Pr\{D | g, \mu\} f_G(g | \theta) f_M(\mu) f_\Theta(\theta). \quad (6)$$

We are interested in the marginal density,  $h_{M\Theta}(\mu, \theta | D)$ . We summarize this density using samples  $(\mu, \theta, g) \sim h_{M\Theta G}$ . The sampled genealogies can be thought of as uninteresting “missing data.”

Consider now the densities  $f_M(\mu)$  and  $f_\Theta(\theta)$ . In any particular application these functions will be chosen to summarize available prior knowledge of scale parameters. It is common practice to avoid the problem of prior elicitation and attempt to construct a “noninformative” prior. This notion is poorly defined, since a prior may be noninformative with respect to some hypotheses, but informative with respect to others. Nevertheless, we illustrate sample-based Bayesian inference under a prior that contains little information. We do this for two reasons. First, we wish to give our sampling instruments a thorough workout. From this point of view an improper prior is the best choice. Second, when carrying out Bayesian inference, it is necessary to test the sensitivity of conclusions to changes in the state of prior knowledge.

What conclusions would a person in a state close to ignorance reach from these data? The improper prior we consider represents ignorance of a rather natural kind. People using our methods will very likely want to consider this particular state of knowledge, along with others that are more representative of their own.

In our case  $\mu$  and  $\theta$  are both scale parameters (for time). The Jeffreys prior,  $f(z) \propto 1/z$ ,  $z > 0$ , invariant under scale transformations  $z \rightarrow az$ , and the uniform prior on  $z > 0$  are candidates for  $f_{\mu}(\mu)$  and  $f_{\theta}(\theta)$ . If  $f_{\mu} \propto 1/\mu$ ,  $f_{\theta} \propto 1/\theta$ , and  $f_{\bar{c}}(g|\theta)$  and  $\Pr\{D|g, \mu\}$  are as given in Equations 1 and 5 then it may be shown that the posterior density in Equation 6 is not finitely normalizable. We may nevertheless consider ratios of posterior densities. But that means the only feasible Bayesian inference, at least under the uniform, improper prior, is exactly frequentist inference. We cannot treat the parameters of interest as random variables. Suppose fixed upper limits  $\mu \leq \mu^*$  and  $t_{\text{root}} \leq t_{\text{root}}^*$  may be set, along with a lower limit  $\theta \geq \theta^*$ . For the problems we use to illustrate our methods in EXAMPLES, conservative limits of this kind determine a state of knowledge that arises quite naturally. Moreover it may be shown that the posterior density is finitely normalizable under uniform priors on the restricted state space, even though the prior on  $\theta$  remains improper.

#### MARKOV CHAIN MONTE CARLO FOR EVOLUTIONARY PARAMETERS

The posterior density  $h_{\text{MCG}}$  is a complicated function defined on a space of high dimension (between 30 and 40 in the examples that follow). We summarize the information it contains by computing the expectations, over  $h_{\text{MCG}}$ , of various statistics of interest. These expectations are estimated using samples distributed according to  $h_{\text{MCG}}$ . We use MCMC to gather the samples we need. MCMC and importance sampling are part of a family of Monte Carlo methods that may be used individually or in concert to solve the difficult integration problems that arise in population genetic inference. Earlier work on this subject is cited in the Introduction. Figure 1 shows a cartoon of two proposal mechanisms used. See the APPENDIX for details of the proposal mechanisms and MCMC integration performed.

As always in MCMC, it is not feasible to test for convergence to equilibrium. MCMC users are obliged to test for stationarity as a proxy. We make three basic tests. First, we check that results are independent of the starting state using 10 independent runs with very widely dispersed initializations. Second, we visually inspect output traces. These should contain no obvious trend. Third, we check that the MCMC output contains a large number of segments that are effectively independent of one another, independent, at least, in the distribution determined empirically by the MCMC output. Let  $\rho_f(k)$  give the autocorrelation at lag  $k$  for some function  $f$  of the MCMC output. Let  $\gamma_f$  denote the asymptotic stan-

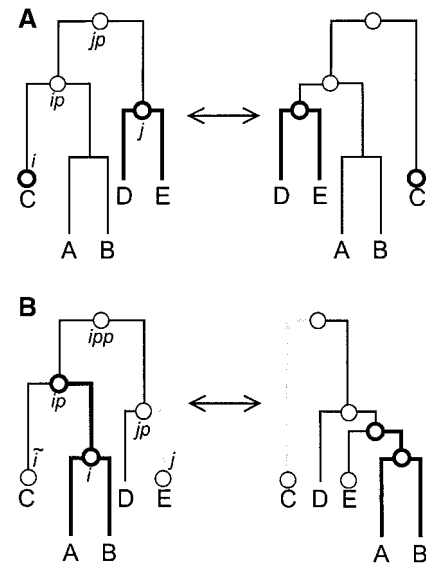


FIGURE 1.—Diagrams of two proposal mechanisms used to modify tree topology during an MCMC analysis. (A) This move is called the “narrow exchange” and is similar to a nearest neighbor interchange. This move picks two subtrees at random under the constraint that they have an aunt-niece relationship; *i.e.*, the parent of one is the grandparent of the other, but neither is parent of the other. Once picked these two subtrees are swapped so long as doing so does not require any modifications in node heights to maintain parent-child order constraints. (B) This move is similar to one proposed by WILSON and BALDING (1998) and involves removing a subtree and reattaching it on a new parent branch.

dard deviation of some estimate of  $\rho_f(k)$ , formed from the MCMC output. Large lag autocorrelations should fall off to zero and remain within  $O(\gamma_f)$  of zero, as discussed by GEYER (1992). Note that in the EXAMPLES section, these standards are not uniformly applied. The first two analyses pass all three checks. The last two analyses pass the first test. Here we are displaying the limitations of our MCMC algorithm. However, we believe the convergence is adequate for the points we make. In the APPENDIX, *Convergence and standard errors* describes the integrated autocorrelation time (IACT) and effective sample size (ESS) measures used to test the efficiency of our sampler.

The MCMC algorithm we used was implemented twice, more or less independently, by A. Drummond, in JAVA and by G. K. Nicholls in MatLab. This allowed us to compare results and proved very useful in debugging some of the more complex proposal mechanism combinations. To minimize programming burden, one of our implementations (G. K. Nicholls in MatLab) was partial, allowing only fixed population size and fixed  $R$  to be compared. This is discussed more extensively in *Implementation issues* in the APPENDIX.

#### EXTENSIONS

Extending the framework of the Introduction and MCMC FOR EVOLUTIONARY PARAMETERS to include de-



terministically varying models of population history and estimation of relative rate parameters is straightforward. Let  $\Phi = (0, \infty)^5$  be the state space for the relative rates of  $R$  above the diagonal and excluding  $R_{G \leftrightarrow T}$ . Let  $s = (\mu, \theta, g, r, R)$ , and let  $h_s(s|D)$  denote the posterior density for  $S \in \Omega_s^*$ , where  $\Omega_s^* = \Omega_{\text{M0G}}^* \times \mathfrak{R} \times \Phi$  (see the APPENDIX). The posterior probability density has the form

$$h_s(s|D) = \frac{1}{Z} \Pr\{D|g, \mu, R\} f_G(g|\theta, r) f_M(\mu) f_\theta(\theta) f_r(r) f_R(R). \quad (7)$$

Let  $T$  denote the age of the most recent leaf, *i.e.*,  $T = \min_{i \in I} t_i$ . In this article  $T = 0$ . Let  $t \geq T$  be a generic age. In this model  $N_e = N_e(t)$ . Recall that  $\rho$ , the number of calendar units per generation, is an unknown constant. Define a constant  $\theta = N_e(T)\rho$  and a growth rate parameter  $r$ . The density  $f_G(g|\theta, r)$  is the density determined by the coalescent process with a population growing as  $N_e(t) = (\theta/\rho) e^{-r(t-T)}$  (SLATKIN and HUDSON 1991). In terms of the notation defined in *Kingman coalescent with temporally offset leaves* in connection with Equation 1, for genealogies with temporally spaced tips the density is

$$f_G(g|\theta, r) = \frac{1}{\theta^{n-1}} \cdot \prod_{i=2}^{2n-1} e^{r t_i} e^{(-k_i(k_i-1)/2\theta r)(e^{r t_i} - e^{r t_{i-1}})}. \quad (8)$$

If all of the relative rates in  $R$ , except  $R_{G \leftrightarrow T}$ , are estimated we are fitting a general time-reversible model of substitution. However, it is sometimes useful to consider simpler nested models. One such model is the Hasegawa-Kishino-Yano (HKY) model (HASEGAWA *et al.* 1985). In the HKY model transitions occur at rate  $\kappa$  relative to transversions. Thus  $R_{A \leftrightarrow G} = R_{C \leftrightarrow T} = \kappa$  and  $R_{A \leftrightarrow C} = R_{A \leftrightarrow T} = R_{C \leftrightarrow G} = R_{G \leftrightarrow T} = 1$ . Either a Jeffreys prior or a uniform prior can be used for the relative rates. However, as a result of our parameterization, the Jeffreys prior provides more accurate estimates. In the examples that follow, a uniform prior is used for  $R$  and  $\kappa$  as this represents the most ignorant state of knowledge and is more than adequate for the purpose of illustrating the methodology. In the same spirit  $f_r(r)$  is set uniform on  $r$ , and this also proves acceptable.

## EXAMPLES

In this section, we illustrate our methods on two HIV-1 *env* data sets and a series of synthetic data sets of comparable size.

**HIV-1 *env* data:** The method was first tested on HIV-1 partial envelope sequences obtained from a single patient over five sampling occasions spanning  $\sim 3$  years: an initial sample (day 0) followed by additional samples after 214, 671, 699, and 1005 days. Details of this dataset have been published previously (RODRIGO *et al.* 1999). An important feature of these data is that monotherapy with Zidovudine was initiated on day 409 (DRUMMOND *et al.* 2001) and continued during the remainder of the

study. The total dataset consists of 60 sequences from these five time points. The length of the alignment is 660 nucleotides. Gapped columns were included in the analysis. The evidence for recombination seems to be negligible in this dataset (RODRIGO *et al.* 1999) and recombination is ignored for the purposes of illustrating our method. Rough estimates of  $N_e$  may be obtained by assuming a generation length of  $\rho = 1$  day per generation (RODRIGO *et al.* 1999). However, we emphasize that we estimate  $N_e\rho$  only in this work. The dataset was split into two subsets for separate analysis. One contained all pretreatment sequences (28 sequences), and the other contained all sequences after treatment commenced (32 sequences; henceforth called posttreatment). The rationale behind this split is that both (1) population size and (2) mutation rate per unit time may be affected by a replication inhibitor such as Zidovudine. In all of the analyses, base frequencies were fixed to empirically determined values; however, inference of these would have been trivial. Two analyses are undertaken on each dataset. The pretreatment data are strongly informative for all parameters estimated. The results are robust to the choice of priors and MCMC convergence is quick. In contrast, the posttreatment data are only weakly informative for  $\mu$ ,  $\theta$ , and  $t_{\text{root}}$  parameters; the results are sensitive to the choice of prior; and MCMC convergence is very slow.

**Pretreatment data, constant population size, HKY substitution:** In this first analysis of the pretreatment dataset, we fit the HKY substitution model and assume a constant population size. We are estimating  $\mu$ ,  $\theta$ ,  $g$ , and  $\kappa$ . We illustrate our methods using uniform prior distributions on  $\mu$  and  $\theta$ , an upper limit on mutation rate of  $\mu^* = 1$ , a lower limit on  $N_e\rho$  of  $\theta^* = 1$ , and a very conservative upper limit on  $t_{\text{root}}$  of  $t^* = 10^7$  days. Ten MCMC runs were made, with starting values for mutation rate distributed on a log scale from  $5 \times 10^{-3}$  down to  $10^{-7}$  mutations/site/day. This range greatly exceeds the range of values supported by the posterior. To test MCMC convergence on tree topologies, each of the 10 MCMC runs was started on a random tree drawn from a coalescent distribution with population size equal to 1000 (in exploratory work we initialize on a sUPGMA or neighbor-joining topology). The 10 Markov chain simulations were run for 2,000,000 steps and the first 100,000 steps were discarded as burn-in. Each run took  $\sim 4$  hr on a machine with a 700 MHz Pentium III processor. The mean IACT of the mutation rate parameter was 4190, giving an ESS of  $\sim 450$  per simulation. Table 1 presents parameter estimates for all 10 runs, illustrating close concordance between runs. Note also that the variability, between runs, of estimated means is in line with standard errors estimated within runs. This is a consistency check on our estimation of the IACT. Figures 2 and 3 show the marginal posterior density of  $\mu$  and  $\theta$  for each of the 10 runs. In all 10 runs the consensus tree computed from the MCMC output was the same, despite the fact that the starting trees were drawn ran-

TABLE 1  
Parameter estimates for 10 independent analyses of the pretreatment dataset assuming constant population size and HKY model of mutation

Run	Mutation rate (mutations/generation/site $\times 10^5$ )	Population size $\times$ generation length ( $\theta$ )	Age of root (days)	Transition/transversion bias parameter ( $\kappa$ )
1	6.238 (0.0517) <sup>a</sup>	1284 (13.0)	796 (6.03)	4.132 (0.00634)
2	6.173 (0.0498)	1304 (12.7)	799 (5.99)	4.141 (0.00599)
3	6.218 (0.0466)	1291 (12.7)	794 (5.45)	4.124 (0.00631)
4	6.168 (0.0434)	1303 (14.0)	797 (5.65)	4.138 (0.00629)
5	6.297 (0.0474)	1269 (12.8)	784 (5.45)	4.134 (0.00640)
6	6.159 (0.0458)	1309 (12.4)	802 (6.21)	4.135 (0.00630)
7	6.308 (0.0539)	1270 (13.9)	784 (5.90)	4.130 (0.00678)
8	6.256 (0.0463)	1279 (11.5)	790 (5.63)	4.133 (0.00674)
9	6.247 (0.0474)	1283 (13.1)	791 (5.75)	4.122 (0.00661)
10	6.201 (0.0578)	1291 (15.4)	801 (7.54)	4.123 (0.00736)
Overall	6.227	1288	794	4.131
95% HPD interval	[4.20, 8.28]	[660, 2050]	[580, 1040]	[3.07, 5.31]

<sup>a</sup> Numbers in parentheses are the standard errors of the means calculated using IACT statistic.

domly (data not shown). Combining the output of all 10 runs, the 95% highest posterior density (HPD) intervals for the mutation rate and  $t_{\text{root}}$  are, respectively,  $[4.20, 8.28] \times 10^{-5}$  mutations per site per day, and  $[580, 1040]$  days.

*Pretreatment data, exponential growth, general substitution model:* In this second analysis of the pretreatment dataset, we fit the general time-reversible substitution model, with exponential growth of population size. We are estimating  $\mu$ ,  $\theta$ ,  $g$ ,  $r$ ,  $R_{A \rightarrow C}$ ,  $R_{A \rightarrow G}$ ,  $R_{A \rightarrow T}$ ,  $R_{C \rightarrow G}$ , and  $R_{C \rightarrow T}$ . This is the most parameter-rich model we fit. To assess the convergence characteristics of this analysis we ran 10

independent runs of 3,000,000 cycles, each starting with an independent random tree topology (the mean IACT for  $\mu$  was 7955 giving an ESS of 358 per run). Figure 4 shows the 10 estimates of the marginal posterior density of mutation rate. Table 2 shows parameter estimates for each of the 10 runs. Convergence is still achieved with the extra parameters.

Compare the distribution of summary statistics under the two models described here and in *Pretreatment data, constant population size, HKY substitution*. Given the nature of infection of HIV-1, it seems likely that an exponential growth rate assumption is more accurate. Esti-

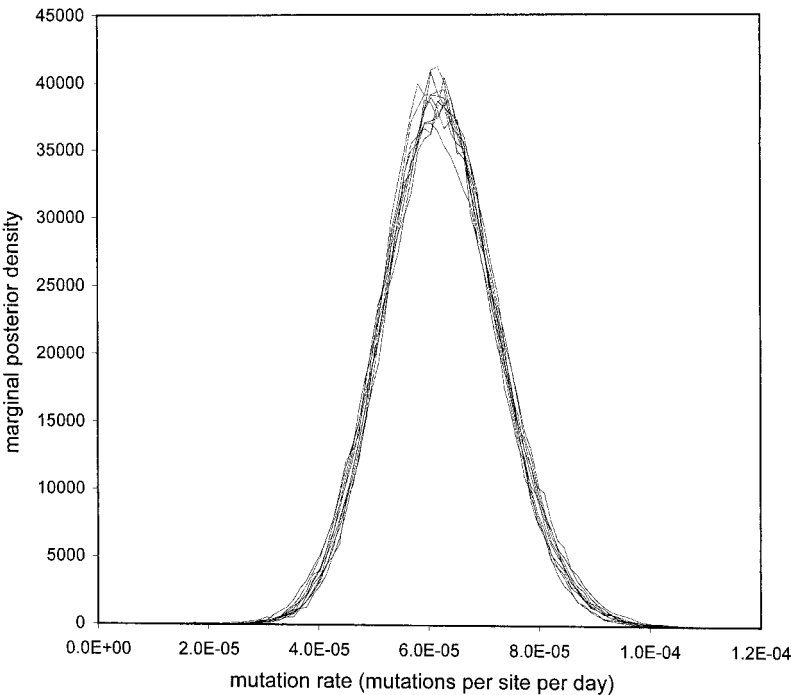


FIGURE 2.—The marginal posterior density of mutation rate for 10 independent MCMC runs on the pretreatment HIV-1 *env* dataset. Each run was started on a random topology. Initial mutation rates ranged from  $5e-3$  to  $1e-7$ .

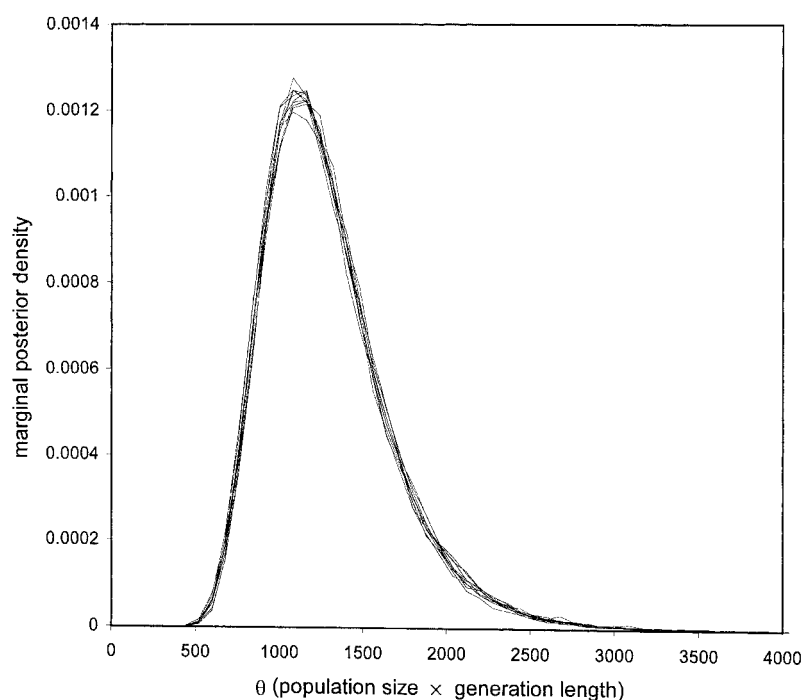


FIGURE 3.—The marginal posterior density of  $\theta$  for 10 independent MCMC runs on the pretreatment HIV-1 *env* dataset. Each run was started on a random tree topology. Initial mutation rates ranged from  $5e-3$  to  $1e-7$ .

mated 95% HPD intervals for the growth rate  $r$ ,  $[1.09 \times 10^{-3}, 6.65 \times 10^{-3}]$ , exclude small growth rates, corroborating this view. The 95% HPD intervals for the mutation rate and  $t_{\text{root}}$  are, respectively,  $[3.61, 8.11] \times 10^{-5}$  mutations per site per day and  $[570, 1090]$  days. Compare these with the model in *Pretreatment data, constant population, HKY substitution*. The change in model has minimal effect ( $<10\%$ ) on the posterior mean mutation rate.

*Posttreatment:* The posttreatment data are analyzed twice under the HKY substitution model with constant

population size. The first analysis uses the same priors as the first pretreatment analysis. In contrast to the pretreatment dataset, the mutation rate of the posttreatment dataset is difficult to estimate. This is illustrated in Figures 5 and 6, in which the marginal posterior densities of  $\mu$  and  $\theta$  estimated from 10 independent MCMC runs, each 5,000,000 cycles long, are compared. We were unable to compute an IACT for each run, so we are unable to compare within- and between-run variability. However, the between-run concordance visi-

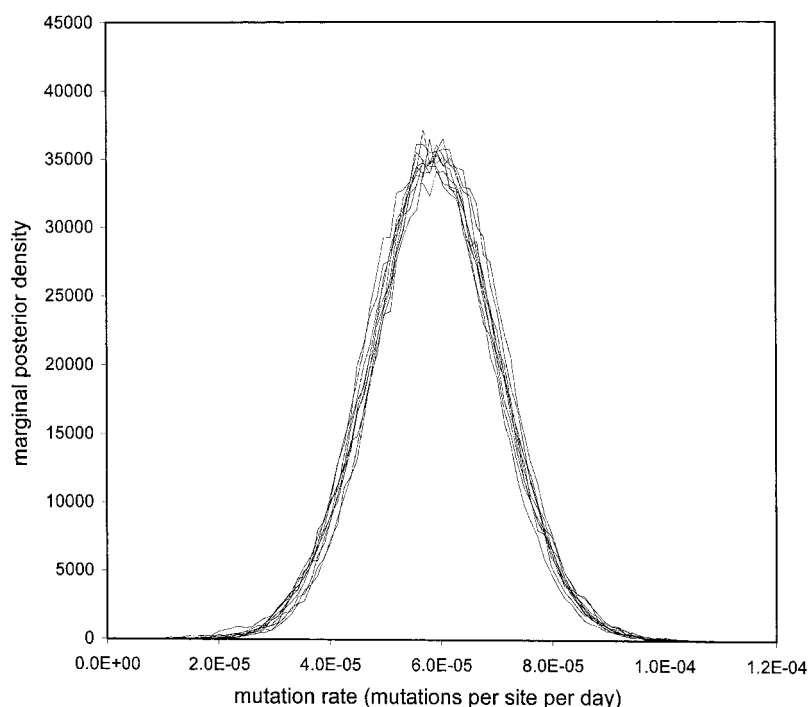


FIGURE 4.—The marginal posterior density of mutation rate for 10 independent MCMC runs on the pretreatment HIV-1 *env* dataset. An exponential growth rate mode of demography and a general time-reversible (GTR) model of substitution were assumed. Each run was started on a random tree topology. Initial mutation rates ranged from  $5e-3$  to  $1e-7$ .

TABLE 2  
Parameter estimates for 10 independent analyses of the pretreatment dataset assuming exponential growth and GTR model of mutation

Run	Mutation rate (mutations/generation/site $\times 10^5$ )	Population size $\times$ generation length ( $\theta$ )	Age of root (days)	Growth rate ( $r \times 10^3$ )
1	5.910 (0.0623) <sup>a</sup>	5404 (127)	800 (7.43)	3.815 (0.0407)
2	5.761 (0.0526)	5321 (125)	821 (7.05)	3.719 (0.0436)
3	6.045 (0.0550)	5089 (123)	786 (6.85)	3.832 (0.0418)
4	5.891 (0.0708)	5443 (172)	806 (8.56)	3.839 (0.0377)
5	5.849 (0.0609)	5338 (113)	812 (8.05)	3.815 (0.0423)
6	5.930 (0.0615)	5242 (170)	804 (8.66)	3.748 (0.0409)
7	5.857 (0.0589)	5318 (148)	806 (7.33)	3.780 (0.0388)
8	5.809 (0.0605)	5236 (123)	817 (7.51)	3.696 (0.0382)
9	5.982 (0.0542)	5064 (127)	795 (5.63)	3.786 (0.0382)
10	5.859 (0.0692)	5306 (188)	813 (10.2)	3.708 (0.0400)
Overall	5.889	5276	806	3.774
95% HPD interval	[3.61, 8.11]	[920, 12450]	[570, 1090]	[1.09, 6.65]

<sup>a</sup> Numbers in parentheses are the standard errors of the means calculated using IACT statistic.

ble in Figure 5 justifies the following statement. The posttreatment mutation rate shows one mode at  $\sim 2.8 \times 10^{-5}$  mutations/site/day with a second mode on the lower boundary. The data determine a diffuse, and bi-modal, marginal posterior on  $\mu$ . One of the modes is associated with states  $(\mu, \theta, g)$  with physically unrealistic root times (greater than the age of the patient). These are allowed, if we are not prepared to assert some restriction on  $t_{\text{root}}$ . This behavior also occurs when we use a Jeffreys prior on the mutation rate (data not shown). It reflects a real property of the data, namely that states

of low  $\mu$  and large  $t_{\text{root}}$  are not well distinguished from otherwise identical states of larger  $\mu$  and smaller  $t_{\text{root}}$ . In the second posttreatment analysis, we revise the upper limit on  $t_{\text{root}}$  downwards, from  $10^7$  to  $t^* = 3650$ , a value more representative of actual prior knowledge for this dataset. The new limit, set 3 years before sero-conversion occurred in the infected patient, is still conservative. Here we explored the prior belief that HIV infection most often originates from a small, homogeneous population and then subsequently accumulates variation. This prior effectively assumes that all viruses

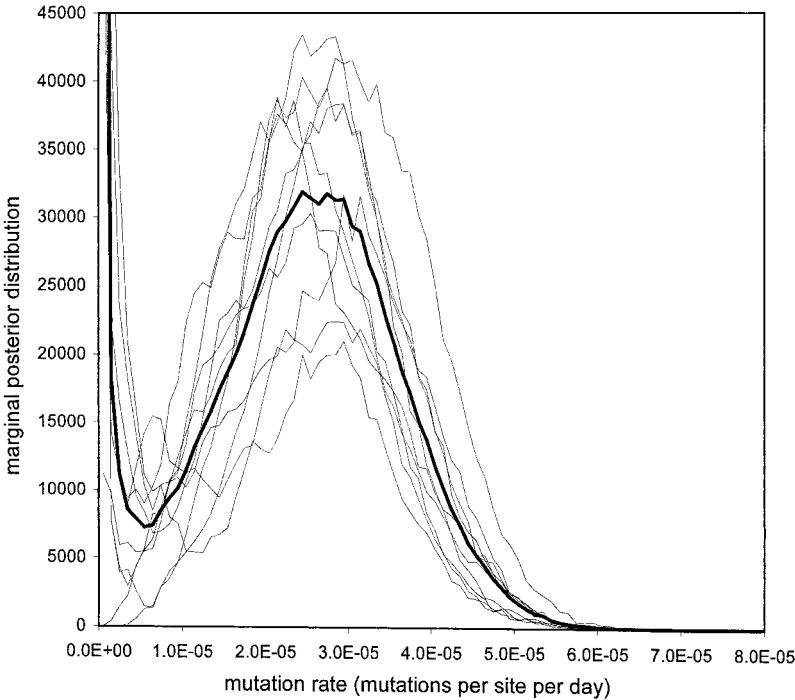


FIGURE 5.—The marginal posterior density of mutation rate for 10 independent MCMC runs on the posttreatment HIV-1 *env* dataset. The thick line represents the density of all 10 runs combined. Each run was started on a random tree topology. Initial mutation rates ranged from  $5e-3$  to  $1e-7$ .



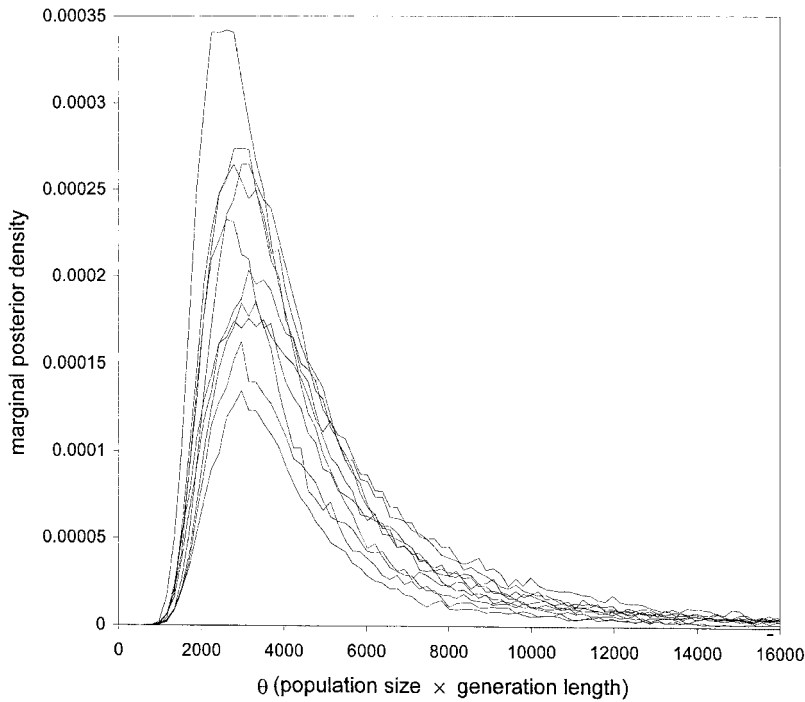


FIGURE 6.—The marginal posterior density of  $\theta$  for 10 independent MCMC runs on the post-treatment HIV-1 *env* dataset. Each run was started on a random tree topology. Initial mutation rates ranged from  $5e-3$  to  $1e-7$ .

in an infected individual share a common ancestor at most as old as the time of infection of the host. The estimated 95% HPD interval for the mutation rate was  $[1.16, 4.27] \times 10^{-5}$  mutations/site/day, markedly down from the pretreatment mutation rate. Figure 7 depicts the resulting unimodal marginal posterior density for mutation rate, showing that the spurious mode has been eliminated. Again, no IACT was computed. However, between-run variability was much improved over Figures

5 and 6. Information about  $t_{\text{root}}$  has been converted into information about mutation rates and population size.

**Simulated sequence data:** To test the ability of our inference procedure to recover accurate estimates of parameters from the above HIV-1 dataset we undertook four simulation studies. In each experiment we generated 100 synthetic datasets. For experiment 1, the posterior estimates of  $\theta$ ,  $\mu$ , and  $\kappa$  obtained from the pretreatment dataset in *Pretreatment data, constant population size*,

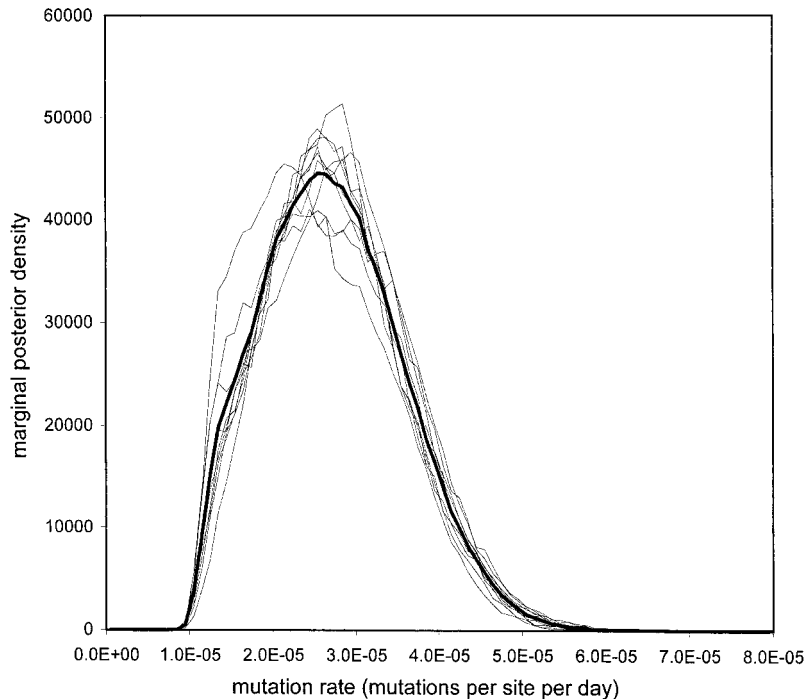


FIGURE 7.—The marginal posterior density of mutation rate for 10 independent MCMC runs on the posttreatment HIV-1 *env* dataset where the age of the root had an upper limit of 10 years (3650 days). The thick line represents the density of all 10 runs combined. Each run was started on a random tree topology. Initial mutation rates ranged from  $5e-3$  to  $1e-7$ .

TABLE 3

Percentage of times that the true parameter was found in the 95% HPD region of the marginal posterior density

Parameter	Experiment			
	1	2	3	4
Mutation rate	92	96	96	97
$\theta$	98	99	96	97
Growth rate	91	92	94	92
RA $\rightarrow$ C	87*	93	96	92
RA $\rightarrow$ G	79*	90	96	94
RA $\rightarrow$ T	83*	90	94	96
RC $\rightarrow$ G	88*	96	98	91
RC $\rightarrow$ T	88*	92	98	94

\*Success rate was significantly <95%.

*HKY substitution* were used to generate 100 coalescent trees and then simulate sequences on each of the resulting trees. The synthetic data were generated under a constant-size population model with the HKY mutation model but analyzed under an exponentially growing population model and a GTR mutation model. In the second experiment, 100 synthetic datasets were generated using the pretreatment parameter estimates in *Pretreatment data, exponential growth, general substitution model* as the true values. In this case the models for simulation and inference are matched. Synthetic data were generated under an exponentially growing population model and a GTR mutation model. In both experiments 1 and 2 uniform bounded priors were used for all parameters. Experiments 3 and 4 differed from experiments 1 and 2 only in that we used JEFFREYS' (1946) prior for scale parameters (mutation rate, population size, and relative rates).

All datasets had the same number of sequences (28), the same sampling times (0 and 214 days), and the same sequence length (660) as the pretreatment dataset. Table 3 shows that the true values are successfully recovered (*i.e.*, fall within the 95% HPD interval)  $\geq 90\%$  of the time in all cases except for the relative rate parameters in experiment 1. In the most complex model we fit, we recover true parameter values. The overparameterization present in experiments 1 and 3 does not seem problematic for estimating mutation rate,  $\theta$ , or growth rate. These results suggest that inference of biologically realistic growth rates is quite feasible. The relative rates performed most poorly in the parameters of interest. This is caused predominantly because the uniform prior on relative rates introduces metric factors that inflate the densities. In experiment 1, when the true value of a relative rate parameter was not within the 95% HPD interval (which occurred 75 times out of 500), it was almost always overestimated (74 out of 75 times). Furthermore, conditioning on a transversion ( $R_{G \leftrightarrow T} = 1$ ), a rare event, may also have an impact. However, experi-

ments 3 and 4 demonstrate that the use of a Jeffreys prior for these and other scale parameters results in  $>90\%$  recovery in all parameters. We are not aiming to prescribe any particular noninformative prior. Our choice of uniform prior in earlier experiments is deliberately crude. However, it allows us to lay out the methodology with as little emphasis as possible on prior elicitation. The reader should undertake this process for a specific problem.

## DISCUSSION

We have described Bayesian coalescent-based methods to estimate and assess the uncertainty in mutation parameters, population parameters, tree topology, and dates of divergence from aligned temporally spaced sequence data. The sample-based Bayesian framework allows us to bring together information of different kinds to reduce uncertainty in the objects of the inference. Much of the hard work is in designing, implementing, and testing a suitable Monte Carlo algorithm. We found a suite of MCMC updates that do the job.

We have analyzed two contrasting HIV-1 datasets and 400 synthetic datasets to illustrate the main features of our methods. The results of the three *HIV-1 env data* subsections show that a robust summary of parameter-rich models, including the joint estimation of mutation rate and population size, is possible for some moderate-sized datasets. The pretreatment data restrict the set of plausible parameter values to a comparatively small range. For this dataset, useful results can be obtained from a state of ignorance about physically plausible outcomes. This situation is in contrast to the situation illustrated in the *Posttreatment* section. For this dataset, prior ignorance implies posterior ambiguity, in the form of a bimodal posterior distribution for the mutation rate. One of these modes is supported by genealogies conflicting with very basic current ideas about HIV population dynamics. We modify the coalescent prior on genealogies to account for this prior knowledge, restricting the most recent common ancestor to physically realistic values. The ambiguity in mutation rate is removed. Similar results could be obtained in a likelihood-based analysis of the posttreatment data, since the prior information amounts to an additional hard constraint on the root time of the coalescent genealogy.

There is some redundancy in the set of MCMC updates we used, in the sense that the limiting distribution of the MCMC is unaltered if we remove the scaling update (move 1) or the Wilson-Balding update (move 2; see APPENDIX for details of these moves). However, these two updates types are needed in practice. There are two timescales in MCMC, time to equilibrium and mixing time in equilibrium. The scaling move sharply reduces mixing time in equilibrium. The Wilson-Balding update is needed to bring the equilibrium time to acceptable values. We have seen MCMC simulations,

minus the Wilson-Balding move, in which an apparently stationary Monte Carlo process undergoes a sudden and unheralded mean shift at  $\sim 2,000,000$  updates. This problem was picked up at the debugging stage, in comparisons between our two MCMC implementations. Subsequent simulation has shown that the genealogies explored in the first 2,000,000 updates of that simulation were just one of the tree clusters supported by the target distribution.

The methods presented here reduce to those of Felsenstein and co-workers (KUHNER *et al.* 1995) in the case of a uniform prior on  $\Theta = 2N_c\mu$ , a fixed  $R$ , a fixed  $\mu$ , and contemporaneous data, if instead of summarizing results using 95% HPD interval estimates, we use the mode and curvature of the posterior density for  $\Theta$  to recover the maximum-likelihood estimate (MLE) and its associated confidence interval.

A distinction can be made between a dataset, like the pretreatment dataset, for which there is strong statistical information about mutation rates (we refer to populations from which such datasets may be obtained as “measurably evolving”), and a dataset, like the posttreatment data, in which the statistical signal is weak. In both of these datasets the familiar parameter  $\Theta = 2N_c\mu$  is in fact well determined by the data (not shown above), so that MCMC convergence in  $\Theta$  is quick. However, it is only in the pretreatment data that this parameter can be separated easily into its two factors. This is related to the well-known problem of identifiability for population size and mutation rate. We can see that temporally spaced data may or may not contain information that allows us to separate these two factors. In this particular example, lineages of the posttreatment viruses branch from those of the pretreatment viral population. Consequently a more appropriate analysis for this dataset would allow for a change of mutation rate and/or population size over the genealogy of the entire set of sequences. In the case of mutation rate this has already been demonstrated within a likelihood framework (DRUMMOND *et al.* 2001). In a Bayesian analysis, coalescence of posttreatment lineages with pretreatment lineages will tend to limit the age of the most recent common ancestor of the posttreatment data, so that the pretreatment lineages will play the role of the reduced upper-bound  $t_{\text{root}}^*$  in the *Posttreatment* section.

A software package called molecular evolutionary population inference (MEPI), developed using the phylogenetic analysis library (PAL; DRUMMOND and STRIMMER 2001), implementing the described method and further extensions (codon position rate heterogeneity, etc.), is available from <http://www.cebl.auckland.ac.nz/mepi/index.html>.

We gratefully acknowledge two anonymous reviewers for helpful comments that much improved the manuscript. In addition, A.D. thanks A. Ferreira. A.D. was supported by a New Zealand Foundation for Research, Science and Technology Bright Futures scholarship. Research by A.G.R. and A.D. was also supported by National Institutes of Health grant GM59174.

## LITERATURE CITED

- BAHLO, M., and R. C. GRIFFITHS, 2000 Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* **57**: 79–95.
- BARNES, I., P. MATHEUS, B. SHAPIRO, D. JENSEN and A. COOPER, 2002 Dynamics of Pleistocene population extinctions in Beringian brown bears. *Science* **295**: 2267–2270.
- BEERLI, P., and J. FELSENSTEIN, 1999 Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**: 763–773.
- BEERLI, P., and J. FELSENSTEIN, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci. USA* **98**: 4563–4568.
- DRUMMOND, A., and A. G. RODRIGO, 2000 Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. *Mol. Biol. Evol.* **17**: 1807–1815.
- DRUMMOND, A., and K. STRIMMER, 2001 PAL: an object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics* **17**: 662–663.
- DRUMMOND, A., R. FORSBERG and A. G. RODRIGO, 2001 The inference of stepwise changes in substitution rates using serial sequence samples. *Mol. Biol. Evol.* **18**: 1365–1371.
- FEARNHEAD, P., and P. DONNELLY, 2001 Estimating recombination rates from population genetic data. *Genetics* **159**: 1299–1318.
- FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**: 368–376.
- FISHER, R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.
- FU, Y. X., 1994 A phylogenetic estimator of effective population size or mutation rate. *Genetics* **136**: 685–692.
- GEYER, C. J., 1992 Practical Markov chain Monte Carlo. *Stat. Sci.* **7**: 473–511.
- GREEN, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**: 711–732.
- GRIFFITHS, R. C., and P. MARJORAM, 1996 Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3**: 479–502.
- GRIFFITHS, R. C., and S. TAVARE, 1994 Ancestral inference in population genetics. *Stat. Sci.* **9**: 307–319.
- HANNI, C., V. LAUDET, D. STEHELIN and P. TABERLET, 1994 Tracking the origins of the cave bear (*Ursus spelaeus*) by mitochondrial DNA sequencing. *Proc. Natl. Acad. Sci. USA* **91**: 12336–12340.
- HASEGAWA, M., H. KISHINO and T. YANO, 1985 Dating of 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.
- HOLMES, E. C., L. Q. ZHANG, P. SIMMONDS, C. A. LUDLAM and A. J. LEIGH BROWN, 1992 Convergent and divergent sequence evolution in the surface envelope glycoprotein of HIV-1 within a single infected patient. *Proc. Natl. Acad. Sci. USA* **89**: 4835–4839.
- HUELSENBECK, J. P., B. LARGET and D. SWOFFORD, 2000 A compound poisson process for relaxing the molecular clock. *Genetics* **154**: 1879–1892.
- JEFFREYS, H., 1946 An invariant form for the prior probability in estimation problems. *Proc. R. Soc. A* **186**: 453–461.
- KINGMAN, J. F. C., 1982a The coalescent. *Stoch. Proc. Appl.* **13**: 235–248.
- KINGMAN, J. F. C., 1982b On the genealogy of large populations. *J. Appl. Probab.* **19A**: 27–43.
- KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140**: 1421–1430.
- KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**: 429–434.
- KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 2000 Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**: 1393–1401.
- LAMBERT, D. M., P. A. RITCHIE, C. D. MILLAR, B. HOLLAND, A. J. DRUMMOND *et al.*, 2002 Rates of evolution in ancient DNA from Adelie penguins. *Science* **295**: 2270–2273.
- LARGET, B., and D. SIMON, 1999 Markov chain Monte Carlo algo-

- gorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* **16**: 750–759.
- LEONARD, J. A., R. K. WAYNE and A. COOPER, 2000 From the cover: population genetics of ice age brown bears. *Proc. Natl. Acad. Sci. USA* **97**: 1651–1654.
- LOREILLE, O., L. ORLANDO, M. PATOU-MATHIS, M. PHILIPPE, P. TABERLET *et al.*, 2001 Ancient DNA analysis reveals divergence of the cave bear, *Ursus spelaeus*, and brown bear, *Ursus arctos*, lineages. *Curr. Biol.* **11**: 200–203.
- MAU, B., M. A. NEWTON and B. LARGET, 1999 Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* **55**: 1–12.
- METROPOLIS, N., A. ROSENBLUTH, M. ROSENBLUTH, A. TELLER and E. TELLER, 1953 Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**: 1087–1091.
- NEE, S., E. C. HOLMES, A. RAMBAUT and P. H. HARVEY, 1995 Inferring population history from molecular phylogenies. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **349**: 25–31.
- PYBUS, O. G., A. RAMBAUT and P. H. HARVEY, 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155**: 1429–1437.
- RAMBAUT, A., 2000 Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**: 395–399.
- RODRIGO, A. G., and J. FELSENSTEIN, 1999 Coalescent approaches to HIV population genetics, pp. 233–272 in *Molecular Evolution of HIV*, edited by K. CRANDALL. Johns Hopkins University Press, Baltimore.
- RODRIGO, A. G., E. G. SHPAER, E. L. DELWART, A. K. IVERSEN, M. V. GALLO *et al.*, 1999 Coalescent estimates of HIV-1 generation time in vivo. *Proc. Natl. Acad. Sci. USA* **96**: 2187–2191.
- RODRIGUEZ, F., J. L. OLIVER, A. MARIN and J. R. MEDINA, 1990 The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**: 485–501.
- SHANKARAPPA, R., J. B. MARGOLICK, S. J. GANGE, A. G. RODRIGO, D. UPCHURCH *et al.*, 1999 Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**: 10489–10502.
- SLATKIN, M., and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**: 555–562.
- SOKAL, A., 1989 Monte Carlo methods in statistical mechanics: foundations and new algorithms. *Cours de Troisième Cycle de la Physique en Suisse Romande*.
- STEPHENS, M., and P. DONNELLY, 2000 Inference in molecular population genetics. *J. R. Stat. Soc. B* **62**: 605–655.
- Swofford, D. L., 1999 *PAUP\*. Phylogenetic Analysis Using Parsimony (\* and Other Methods)*. Sinauer Associates, Sunderland, MA.
- 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.
- WILSON, I. J., and D. J. BALDING, 1998 Genealogical inference from microsatellite data. *Genetics* **150**: 499–510.
- WOLINSKY, S. M., B. T. M. KORBER, A. U. NEUMANN, M. DANIELS, K. J. KUNTSMAN *et al.*, 1996 Adaptive evolution of HIV-1 during the natural course of infection. *Science* **272**: 537–542.
- WRIGHT, S., 1931 Evolution in Mendelian populations. *Genetics* **16**: 97–159.
- YANG, Z., and B. RANNALA, 1997 Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.* **14**: 717–724.

Communicating editor: J. HEIN

## APPENDIX: MCMC DETAILS AND MOVE TYPES

Markov chain Monte Carlo for temporally spaced sequence data including proposal mechanism used is described.

Denote by  $\Omega_{\text{MCG}}$  the space  $[0, \infty) \times [0, \infty) \times \Gamma$  of all possible  $(\mu, \theta, g)$  values. Let

$$\Omega_{\text{MCG}}^* = \{(\mu, \theta, (E_g, t_r)) \in \Omega_{\text{MCG}} : \mu \leq \mu^*, \theta \geq \theta^*, t_{\text{root}} \leq t_{\text{root}}^*\}.$$

We now describe a Monte Carlo algorithm realizing a Markov chain  $X_n$ ,  $n = 0, 1, 2, \dots$  with states  $x = (\mu, \theta, g)$ ,  $x \in \Omega_{\text{MCG}}^*$ , and equilibrium  $h_X = h_{\text{MCG}}$ .

Suppose  $X_n = x$ . A value for  $X_{n+1}$  is computed using a Metropolis-Hastings algorithm. Define a set of random operations on the state. A given move may alter one or more of  $\mu$ ,  $\theta$ , and  $g$ . Label the different move types  $m = 1, 2, \dots, \mathbf{M}$ . The random operation with label  $m$ , acting on state  $x$ , generates state  $x'$ , with probability density  $q_m(x'|x)$ , say. Let  $(a \wedge b)$  equal  $a$  if  $a < b$  and otherwise  $b$  and  $(a \vee b)$  equal  $a$  if  $a > b$  and otherwise  $b$ , let

$$P(x, x') = h_X(x'|D)/h_X(x|D)$$

stand for the ratio of posterior densities, and let

$$Q_m(x, x') = q_m(x|x')/q_m(x'|x)$$

give the ratio of the densities for proposals  $x' \rightarrow x$  and  $x \rightarrow x'$ . The algorithm determining  $X_{n+1}$  given  $X_n$  can be described as follows. First, a label  $m$  is chosen according to some arbitrary fixed probability distribution on the  $\mathbf{M}$  move types. A value for the candidate state  $x'$  is drawn according to the density  $q_m(x'|x)$ . Second, we accept the candidate, and set  $X_{n+1} = x'$  with probability

$$\alpha_m(x, x') = 1 \wedge (P(x, x')Q_m(x, x')). \quad (9)$$

Otherwise, with probability  $1 - \alpha_m(x, x')$ , the candidate is rejected and we set  $X_{n+1} = x$ .

**Proposal mechanisms:** In this section we describe the proposal mechanisms (moves) and their acceptance probabilities. In each move,  $X_n = x$ , with  $x = (\mu, \theta, (E_g, t_r))$ . For each node  $i$  let  $\text{parent}(i) \in Y$  denote the label of the node ancestral to  $i$  and connected to  $i$  by an edge. We get a compact notation if we treat  $Y$  and  $g$  as if  $Y$  contained a notional  $\text{parent}(\text{root})$  node with  $t_{\text{parent}(\text{root})} = \infty$ , as we did in Equation 4. Also, we now drop the convention that node labels increase with age.

Let  $dx = d\mu \, d\theta \, dg$  in  $\Omega_{\text{MCG}}^*$  and

$$H_X(dx|D) = h_X(x|D)dx.$$

The moves listed below determine an  $H_X$ -irreducible aperiodic Metropolis-Hastings kernel. The MCMC is Harris recurrent and ergodic, with  $H_X$  its unique equilibrium distribution.

**Scaling move:** Label this move  $m = 1$ . Let a real constant  $\beta > 1$  be given. For  $\beta^{-1} \leq \delta \leq \beta$ , let  $x \rightarrow \delta x$  denote the transformation

$$(\mu, \theta, (E_g, t_r)) \rightarrow (\mu/\delta, \delta\theta, (E_g, \delta t_r)).$$

If  $x' = \delta x$  then  $x = \delta' x'$  with  $\delta' = 1/\delta$ . The change of variables in the product measure is

$$H_X(dx'|D)d\delta' = \delta^{n-3} H_X(dx|D)d\delta.$$

Note that this transformation is not simply a change of units. The times  $t_i$  associated with ancestral nodes  $i \in$



$Y$  are scaled while leaf node times  $t_i$ ,  $i \in I$  (which are part of the data) are left unchanged.

The move is as follows. Choose a  $\delta \sim \text{Unif}(\beta^{-1}, \beta)$  and set  $x' = \delta x$ . If  $x \notin \Omega_{\text{MCG}}^*$  (if, for example,  $\mu/\delta > \mu^*$ , or the parent-child age order constraint is violated at the unscaled leaves in the scaled tree), then the move fails and we set  $X_{n+1} = x$ . In a slight abuse of notation we set  $Q_1(x, x') = 1/\delta^{n-3}$  in the formula for  $\alpha_1(x, x')$  in Equation 9 (GREEN 1995 explains how this scale factor arises in Metropolis-Hastings MCMC). The choice  $\beta = 1.2$  gave reasonable acceptance rates in our simulations.

**Wilson-Balding move:** Label this move  $m = 2$ . A random subtree is moved to a new branch. This move is based on the branch-swapping move of WILSON and BALDING (1998). The SPR move in PAUP\* (SWOFFORD 1999) is similar. However, the move below acts on a rooted tree and maintains all node ages except one.

Two nodes,  $i, j \in I \cup Y$  are chosen uniformly at random without replacement. Let  $\text{jp} = \text{parent}(j)$  and  $\text{ip} = \text{parent}(i)$ . If  $t_{\text{jp}} \leq t_i$ , if  $\text{ip} = j$  or  $\text{ip} = \text{jp}$ , then the move fails and we set  $X_{n+1} = x$ . Given  $i$  and  $j$ , the candidate state  $x' = (\mu, \theta, g')$  is generated in the following way. Let  $\tilde{i}$  denote the child of  $\text{ip}$  that is not  $i$ , and let  $\text{ipp} = \text{parent}(\text{ip})$ , the grandparent of  $i$ . Reconnect node  $\text{ip}$  so that it is a child of  $\text{jp}$  and a parent of  $j$ ; that is, set

$$E'_g = \{\langle \text{jp}, j \rangle, \langle \text{ip}, \tilde{i} \rangle, \langle \text{ipp}, \text{ip} \rangle\} \cup E_g \setminus \{\langle \text{jp}, \text{ip} \rangle, \langle \text{ip}, j \rangle, \langle \text{ipp}, \tilde{i} \rangle\}.$$

If node  $j$  is not the root, assign to node  $\text{ip}$  a new time  $t'_{\text{ip}}$  chosen uniformly at random in the interval  $[(t_i \vee t_j), t_{\text{jp}}]$ . If node  $j$  is the root, choose  $\delta \sim \text{Exp}(\theta)$  and set  $t'_{\text{ip}} = t_j + \delta$ . Let  $t'_Y$  denote the set of node times with  $t_{\text{ip}}$  replaced by  $t'_{\text{ip}}$ . Let  $x' = (\mu, \theta, (E'_g, t'_Y))$ . If node  $j$  and node  $\text{ip}$  are not root, the ratio  $Q_2(x, x')$  in Equation 9 is

$$Q_2(x, x') = (t_{\text{jp}} - (t_i \vee t_j)) / (t_{\text{ipp}} - (t_i \vee t_j)).$$

If node  $j$  is the root,

$$Q_2(x, x') = \theta / (\exp(-\delta/\theta)(t_{\text{ipp}} - (t_i \vee t_j))),$$

and if  $\text{ip}$  is the root,

$$Q_2(x, x') = (t_{\text{jp}} - (t_i \vee t_j)) \exp(-(t_{\text{ip}} - t_i)/\theta) / \theta.$$

**Subtree exchange:** Label this move  $m = 3$ . Choose a node  $i \in I \cup Y$ . Let  $\text{ip} = \text{parent}(i)$ ,  $\text{jp} = \text{parent}(\text{ip})$ , and let  $j$  denote the child of  $\text{jp}$  that is not  $\text{ip}$ . If node  $i$  is the root or a direct child of the root, or  $t_{\text{ip}} < t_j$ , then the move fails and we set  $X_{n+1} = x$ . Given  $i$  and  $j$ , the candidate state  $x' = (\mu, \theta, g')$  is generated in the following way. Swap nodes  $i$  and  $j$ , setting

$$E'_g = \{\langle \text{ip}, j \rangle, \langle \text{jp}, i \rangle\} \cup E_g \setminus \{\langle \text{jp}, j \rangle, \langle \text{ip}, i \rangle\}.$$

Let  $x' = (\mu, \theta, (E'_g, t_Y))$ . The ratio  $Q_3(x, x') = 1$  in Equation 9.

The subtree exchange above is a local operation. In a second version of this move we chose node  $j$  uniformly at random over the whole tree.

**Node age move:** Label this move  $m = 4$ . Choose an internal node,  $i \in Y$ , uniformly at random. Let  $\text{ip} = \text{parent}(i)$  and let  $j$  and  $k$  be the two children of  $i$  [so  $i = \text{parent}(j)$  and  $i = \text{parent}(k)$ ,  $j \neq k$ ]. If  $i$  is not the root, choose a new time  $t'_i$  uniformly at random in  $[(t_j \vee t_k), t_{\text{ip}}]$ ; otherwise, if  $i$  is the root, choose  $\delta \sim \text{Unif}(\beta^{-1}, \beta)$  (see move  $m = 1$ ) and set  $t'_i = (t_j \vee t_k) + \delta(t_i - \delta(t_j \vee t_k))$ . Let  $t'_Y$  denote the set of ancestral node times,  $t_Y$ , with  $t_i$  replaced by  $t'_i$ . Let  $x' = (\mu, \theta, (E_g, t'_Y))$ . If  $i$  is not the root, then  $Q_4(x, x') = 1$  in Equation 9. If  $i$  is the root then  $Q_4(x, x') = 1/\delta$ .

**Random walk moves for  $\theta$  and  $\mu$ :** Label this move  $m = 5$ . The random walk update to  $\theta$  is as follows. Let a real constant  $w_\theta > 0$  be given. Choose  $\delta \sim \text{Unif}(-w_\theta, w_\theta)$  and set  $x' = (\mu, \theta + \delta, g)$ . If  $x \notin \Omega_{\text{MCG}}^*$ , then the move fails and we set  $X_{n+1} = x$ . Since the candidate generation process is symmetric,  $Q_5(x, x') = 1$ , in the formula for  $\alpha_5(x, x')$  in Equation 9. The random walk move for  $\mu$ , with random walk window parameter  $w_\mu$ , say, is similar to the move just described for  $\theta$ . The window sizes  $w_\theta$  and  $w_\mu$  must be adjusted to get reasonable sampling efficiency.

#### Implementation, convergence checking, and debugging:

**Convergence and standard errors:** The efficiency of our Markov sampler, as a tool for estimating the mean of a given function  $f$ , is measured by calculating from the output  $\tau_f = 1 + 2\Sigma p_j(k)$ , the IACT of  $f$ . Dividing the run length by  $\tau_f$ , we get the number of “effective independent” samples in the run (the number of independent samples required to get the same precision for estimation of the mean of  $f$ ). We call this the ESS. Better MCMC algorithms have smaller IACTs and thus larger ESSs, though it may be necessary to measure  $\tau$  in units of CPU time to make a really useful comparison. One will typically want to run the Markov chain at least a few hundred times the IACT, to test convergence and get reasonably stable marginal histograms. Note first that we do not know the IACT when we set the MCMC running. Exploratory runs are needed. Second, a statement like “We ran the MCMC for  $10^6$  updates discarding the first  $10^4$ ” is worthless without some accompanying measurement of an IACT or equivalent. This point is made in SOKAL (1989). The summation cutoff in the estimate for the IACT,  $\tau_f$ , is determined using a monotone sequence estimator (GEYER 1992). The IACTs we get for our MCMC algorithms suggest that analysis of large datasets (50–100 sequences and 500–1000 nucleotides) is feasible with current desktop computers. Examples may be found in EXAMPLES (Table 2). The inverse of the IACT of a given statistic is the “mixing rate.” Statistics with small mixing rates are called the “slow modes” of a MCMC algorithm. The mutation rate  $\mu$  was the slowest mode among those we checked, and we therefore present IACTs for that statistic in EXAMPLES.

**Implementation issues:** In this section we discuss debugging and MCMC efficiency of our two implementations.

We compare expectations computed in the coalescent with estimates obtained from MCMC output. Standard errors are obtained from estimates of the corresponding IACT. Consider a tree with four leaves, two at time zero and two offset  $\tau$  time units to greater age. Consider simulation in the coalescent, with no data. The expectation of  $t_{\text{root}}$  is

$$E_G\{t_{\text{root}}\} = (\tau + 4\theta/3)(1 - e^{-\tau/\theta}) + (\tau + 3\theta/2)e^{-\tau/\theta}.$$

A number of other expectations may be computed.

For problems involving data, expectations are not available. However, an MCMC algorithm with several different move types may be tested for consistency. The equilibrium is the posterior distribution of  $\mu$ ,  $\theta$ , and  $g$  and should not alter as we vary the proportions in which move types are used to generate candidate states. For example, move 2 (Wilson-Balding) is irreducible on its own, while moves 3 and 4 (subtree exchange and node-age move) form another irreducible group. We fix a small synthetic dataset and compare the output of two MCMC runs: one generated using move 2 alone and the other using moves 3 and 4 alone in tandem.

We now turn to questions of MCMC efficiency. Each update has a number of parameters. These are adjusted, by trial and error for each analysis, so that the MCMC is reasonably efficient. An *ad hoc* adaptive scheme, based on monitoring acceptance rates, and akin to that described in LARGET and SIMON (1999), was used. The samples used in output analysis are taken from the final portion of the run, in which these parameters are fixed.

The scaling and Wilson-Balding updates are particularly effective.

We have experimented with a range of other moves. However, while it is easy to think up computationally demanding updates with good mixing rates per MCMC update, we have focused on developing a set of primitive moves with good mixing rate per CPU second. In our experience simple moves may have low acceptance rates, but they are easy to implement accurately and are rapidly evaluated. They may give good mixing rates when we measure in CPU seconds. LARGET and SIMON (1999) have given an effective MCMC scheme for a similar problem. We did not use their scheme, as its natural data structure did not fit well with our other operators. A second update, which may be useful to us in the future, would use the importance-sampling process of STEPHENS and DONNELLY (2000) to determine an independence sampling update.

Because of the explicit nature of MCMC inference, the details of a particular analysis, including the proposal mechanisms, the chain length, the evolutionary model, and the prior distributions, can be quite difficult to keep track of. One of us (A. Drummond) developed an XML data format to describe phylogenetic/population genetic analyses. This enables the user to write down the details of an analysis in a human-readable format that can also be used as the input for the computer program. For the more visually inclined a graphical user interface (GUI) was developed that can generate the XML input files, given a NEXUS or PHYLIP alignment.