The Effects of Rate Variation on Ancestral Inference in the Coalescent

Lada Markovtsova,* Paul Marjoram† and Simon Tavaré*,†,‡

**Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113, †Biostatistics Division, Department of Preventive Medicine, University of Southern California, Los Angeles, California 90033 and ‡Program in Molecular Biology, Department of Biological Sciences, University of Southern California, Los Angeles, California 90089-1340*

Manuscript received February 22, 2000
Accepted for publication July 17, 2000

ABSTRACT

We describe a Markov chain Monte Carlo approach for assessing the role of site-to-site rate variation in the analysis of within-population samples of DNA sequences using the coalescent. Our framework is a Bayesian one. We discuss methods for assessing the goodness-of-fit of these models, as well as problems concerning the separate estimation of effective population size and mutation rate. Using a mitochondrial data set for illustration, we show that ancestral inference concerning coalescence times can be dramatically affected if rate variation is ignored.

IT is widely recognized in the phylogenetic community that variation in mutation rates across sites can have significant impact on phylogenetic analyses based on sequence data. The review article of Yang (1996a) provides an historical overview with particular reference to gamma models for rate variation. A number of authors have studied the effects of rate variation for within-species samples of sequences (Aris-Brosou and Excoffier 1996; Deng and Fu 1996; Tajima 1996; Yang 1996b; Misawa and Tajima 1997). This work has concentrated largely on the number of segregating sites observed in the sample. The focus on this simple summary statistic was predicated in part on the lack of theoretical and computational approaches for studying complete sequence data. Studies of rate variation in hypervariable region I of human mtDNA (*cf.* Lundstrom *et al.* 1992b; Wakeley 1993; Excoffier and Yang 1999; Meyer *et al.* 1999) have emphasized the likely drawbacks of using simple mutation processes such as the infinitely many sites model to analyze data where site-to-site variation is known to occur. Other studies, such as Sigurdardóttir *et al.* (2000), have attempted to take advantage of large, detailed pedigree information to infer mutation rates from observed mutation events.

In this article we develop a Markov chain Monte Carlo (MCMC) approach for studying rate variation for within-population samples of DNA sequences evolving according to a coalescent model (Kingman 1982). Our method generates observations from the posterior distribution of the coalescent tree topology, the coalescence times, and the mutation parameters, conditional on the sequence data observed in the sample. MCMC methods for maximum-likelihood estimation in the coalescent were introduced by Kuhner *et al.* (1995, 1998), and further developed by Wilson and Balding (1998) in the Bayesian setting. Bayesian MCMC approaches are now available in the phylogenetic setting as well (*cf.* Yang and Rannala 1997; Larget and Simon 1999; Mau *et al.* 1999). The Bayesian approach provides a useful computational device for maximum-likelihood estimation, as we illustrate in results. We also suggest a method for assessing the adequacy of the fit of such models to the data.

We assume we have a random sample of n chromosomes, for each of which we have the DNA sequence of a region of interest. We denote the collection of sequence data by **D**. We illustrate our approach with a sample of mitochondrial sequences from the Nuu Chah Nulth obtained by Ward *et al.* (1991). The data **D** are 360-bp sequences from region I of the control region obtained from a sample of $n = 63$ individuals. The observed base frequencies are $(\pi_A, \pi_G, \pi_C, \pi_T) = (0.3297, 0.1120, 0.3371, 0.2212)$. The data have 26 segregating sites, a mean heterozygosity of 0.0145 per site and 28 distinct haplotypes with a haplotype homozygosity of 0.0562. We show that allowance for rate variation provides a better fit to these data. Furthermore, rate variation is seen to give much smaller values for the conditional coalescence times than are predicted using an infinitely many sites model. This is true in general, since the infinitely many sites model fails to allow for recurrent mutations and thus tends to underestimate the mutation rate, with a consequent overestimation of coalescence times. For the data here, the expected time to the most recent common ancestor of the sample is reduced by a factor of about two. These results suggest that allowance should be made for rate variation when using coalescent-based models for ancestral inference. Our approach provides a way to do this. We conclude

Corresponding author: Simon Tavaré, Program in Molecular Biology, Department of Biological Sciences, SHS 172, 835 W. 37th St., University of Southern California, Los Angeles, CA 90089-1340.
E-mail: stavare@gnome.usc.edu

by discussing whether it is reasonable to use such methods to make inferences about population size and mutation rate separately, rather than, as is more common, estimating a compound parameter that is proportional to their product.

The coalescent: The simplest version of the coalescent assumes that the population is random mating and of a large, constant, effective size N. The coalescent describes the ancestry of a random sample of n chromosomes taken from the present-day population. For convenience we refer to the time at which the sample was taken as time 0 and let time increase as we look back into the past. For a wide range of models, the probability that two individuals share a common ancestor in the previous generation is $\sigma^2/(N-1)$, where σ^2 is the variance of the distribution of the number of offspring produced by a parent in a single generation (Cannings 1974). For convenience we scale time in units of N/σ^2 generations so that in a large population a given pair of individuals coalesces at rate 1. As is common, we shall assume $\sigma^2 = 1$ so that a coalescent time of 1 corresponds to N generations. Note that this assumption affects the interpretation of estimates of ages and mutation rates on the basis of the data. If the true σ^2 is >1 then estimates for times to the MRCA and ages of mutations will most likely be overestimates, with a consequent underestimation of mutation rates. Let T_j denote the time period during which the sample has j distinct ancestors. The T_j have independent exponential distributions with parameter $j(j-1)/2$. When a coalescence event occurs, two lines of ancestry are picked, uniformly at random, and are coalesced to form one resulting line. The process of coalescences terminates when a single line of ancestry remains. The genealogy can be viewed as consisting of two components: the topology of the tree structure and the times between coalescent events. We use Λ to denote the topology of the tree and $T = \{T_n, T_{n-1}, \ldots, T_2\}$ to denote the set of coalescence times in the sample. Accessible reviews of the coalescent are given by Hudson (1991) and Donnelly and Tavaré (1995), for example.

MATERIALS AND METHODS

Mutation model for sequences: We use a variety of finite-sites models, in which mutations are assumed to occur according to a model of Felsenstein, described in detail in Thorne *et al.* (1992). Each model has a transition-transversion parameter κ, assumed to be the same at each site, and a rate parameter g_i at the ith of the L sites in the sequence. Among the parameters of interest is the effective per-site average substitution rate θ defined in Equation 4; it may be calculated from the values of g_i, κ, and the base frequencies in the sequences. Further details of the mutation model are deferred to the appendix.

The parameters in the model are denoted by the vector M. In all cases, we treat the unknown rate parameters M as having a prior distribution, and we simulate observations from the posterior distribution of M given **D**. In the most general setting, there are $L+1$ rates, resulting in a highly overparameterized model. We therefore examine a number of special cases of this general model, as described below.

Model 1: All sites mutate at the same rate, so that $g_i \equiv g$ for all sites i. Here $M = (g, \kappa)$.

Model 2: The special case of model 1 in which κ is assumed known, so that $M = (g)$. This model serves as a simple description of mutation in hypervariable region I of mtDNA. It was used by Kuhner *et al.* (1995) in their analysis of the same data set.

In the remaining models, we assume κ is known. There are L rate parameters, and $M = (g_1, \ldots, g_L)$. To reduce the number of rate parameters we assign each site to one of a number of rate classes. If there are c classes, then the model has c rates and $M = (h_1, \ldots, h_c)$, where h_l now refers to the common rate for the lth class. We treat two subcases of this model, one based on a statistical description of rate variation, the second on a biological one.

Model 3: Meyer *et al.* (1999) developed a method for studying rate variation in the hypervariable region I of mtDNA using the Tamura-Nei mutation model (Tamura and Nei 1993) with gamma rate variation. They used a worldwide sample of sequences to estimate the relative rate at each site in the region. We used their results to classify the sites in hypervariable region I into five different rate classes, assigning all sites that they classified as unvarying to the class with the lowest rate. This classification resulted in 194 sites in the class with the smallest rate, 103 in the next class, 23 in the third and fourth classes, and 17 sites in the most rapidly evolving class. The classification of each site may be obtained from http://hto-e.usc.edu/datasets/mtrates.txt. The mutation rate for each of the five classes is allowed to vary. Thus $g_i = h_l$ if site i is assigned to class l. The mutation parameters are $M = (h_1, h_2, h_3, h_4, h_5)$.

Model 4: Here we use just two rate classes, one for purines and one for pyrimidines (*cf.* Lundstrom *et al.* 1992b). In this case $M = (h_1, h_2)$. It is reasonable to do this since we observe no transversions in the data.

Computational approach: Here we describe our approach for generating observations from the posterior distribution of the mutation parameters M and features of the tree topology Λ and times T, given the observed sequence data. We define the effective mutation parameter $\theta/2$ to be the expected number of substitutions per site per unit time that result in a change of base. Details of the derivation of θ are given in Equations 4 and 6 in the appendix. Among the issues we address are the sensitivity of the effective mutation parameter θ to different mutation models, the effect these differences might have on ancestral inference concerning (for example) the time to the MRCA of the sample, and the adequacy of the fit of the models to the data.

We assume a prior distribution for M and develop an MCMC method for generating observations from the conditional density $f(G|\mathbf{D})$ of $G = (\Lambda, T, M)$ given **D**. Since the topology of the coalescent is independent of the times of coalescent events, and both are independent of the mutation process, we can write

$$f(G|\mathbf{D}) = \mathbb{P}(\mathbf{D}|G)\,p_1(\Lambda)\,p_2(T)\,p_3(M)/f(\mathbf{D}). \qquad (1)$$

The first term on the right can be computed using a peeling algorithm (*cf.* Felsenstein 1981) and the mutation model we have described. The term $p_1(\Lambda)$ on the right of (1) is the coalescent tree topology distribution, $p_2(T)$ is the density of the coalescence times T, and $p_3(M)$ is the prior distribution for the mutation rates M. The normalizing constant $f(\mathbf{D})$ is

unknown and hard to compute. We therefore use a version of the Metropolis-Hastings method (Metropolis *et al.* 1953; Hastings 1970) to simulate from the required conditional distribution.

Markov chain Monte Carlo method: The algorithm produces correlated samples from a distribution π of interest, in our case $\pi(G) \equiv f(G|\mathbf{D})$. It starts with an arbitrary choice of Λ, T, and M. New realizations of G are then proposed and accepted or rejected, according to the basic Metropolis-Hastings method:

1. Denote the current state by $G = (\Lambda, T, M)$.
2. Output the current value of G.
3. Propose $G' = (\Lambda', T', M')$ according to a kernel $Q(G \rightarrow G')$.
4. Compute the acceptance probability

$$h = \min\left\{1, \frac{\pi(G')\,Q(G' \rightarrow G)}{\pi(G)\,Q(G \rightarrow G')}\right\}. \quad (2)$$

5. Accept the new state G' with probability h, otherwise stay at G.
6. Return to step 1.

Let $X(t)$ denote the state of this chain after t iterations. Once $X(t)$ has reached stationarity its values represent samples from the distribution $\pi(G) = \pi(\Lambda, T, M)$. In many cases it is desirable to simulate approximately independent samples from the posterior distribution of interest, in which case we use output from every mth iteration for a suitable choice of m.

There are many possible choices for the updating mechanism $Q(\cdot, \cdot)$. The particular $Q(\cdot, \cdot)$ we use will impact the efficiency of the scheme, both in terms of speed of computation and rate of approach to stationarity. Furthermore the chain $X(t)$ must be irreducible and positive recurrent to ensure that the limiting distribution is indeed $\pi(\Lambda, T, M)$. Informally, the chain $X(t)$ must be able to reach any feasible state, from any feasible starting point, in a finite period of time. Any of the other updating schemes referenced in the Introduction might be adapted to the problems explored here, but we have chosen to use a version of the scheme given in Markovtsova *et al.* (2000). For full details we refer the reader to that article, but we now give a brief informal description. The algorithm makes local changes to the tree by picking a random coalescence event and considering this event and the next coalescence event. So if the first event involves the transition from k ancestors to $k - 1$, then we also consider the coalescence involving the transition from $k - 1$ to $k - 2$ ancestors. We then make a local rearrangement of the lines involved in these two coalescences. We simultaneously propose new times for the periods during which there are k and $k - 1$ ancestors to the sample. These new times may be generated according to the predata coalescent distribution, or according to Normal distributions with mean given by the current values of the times. We update the mutation parameters M every 10 iterations in the analyses presented in the next section. New values are proposed according to a Normal distribution centered around the currently accepted value.

RESULTS

We ran the updating scheme described in the previous section on the Nuu Chah Nulth data using mutation models 1–4. The output typically appeared to be nonstationary for up to 200,000 iterations of the algorithm. We sampled every 10,000th iteration to approximate a random sample from the stationary distribution. In a bid to be very conservative, and since the algorithms run rapidly, we generally discarded the first 2500 samples and based our analysis on the next 5000 samples. The acceptance rate was typically ~80%. For runs in which we needed to tune a variance parameter the burn-in length varied, but the estimated parameter values were unchanged for the different variances we tried.

One should begin to sample from the process $X(\cdot)$ once it has "reached stationarity." There are many heuristic tests for this, none of which is infallible. For a critique see Gilks *et al.* (1996). Some simple diagnostics are functions of the statistics of interest such as autocorrelations and moving averages. It is also valuable to run the chain from several different, widely spaced, starting points and compare the long-term behavior. We also used the tests contained in the software package CODA (Best *et al.* 1995). All tests were passed with the exception of the test that indicates the presence of correlation in the log-likelihood series; this could be removed if necessary by sampling less frequently.

Some time might be saved by starting the process from a genealogy (Λ, T) for which $\mathbb{P}(\Lambda, T|\mathbf{D})$ is relatively high. The rationale for this is that it is sensible to start from a region of the state-space that is well supported by the data. As an example of this one might use the UPGMA tree for the data set, as described in Kuhner *et al.* (1995). However, we usually started from random tree topologies since convergence from different starting points is potentially a useful diagnostic for stationarity.

The transition-transversion parameter κ: The analysis of model 1 produced a median value of κ of 65.1, with lower and upper quartiles of 32.7 and 162.7, respectively. Note that since the data are consistent with no transversions having occurred during the evolution of the sample, the posterior distribution for κ has a very long right tail and statistics for the mean, which are strongly influenced by outliers, are potentially misleading and are therefore not presented. The median value of g was 6.87×10^{-4} and the median value for w was 4.47×10^{-2}. These results show that the data are consistent with a value of $\kappa = 100$, as has previously been used by Kuhner *et al.* (1995) in their analysis of the same data. For the analysis of subsequent models we treated $\kappa = 100$ as fixed.

The mutation parameter θ: Posterior statistics for the per-site average mutation rate θ for each of the models are presented in Table 1, and Figure 1 shows the estimated posterior density for each model. We note that the posterior distribution supports higher values for θ in model 3. This is largely due to the high mutation rate that is assigned to class 5, which contains 17 sites believed to mutate at the highest rate. The posterior distributions of the per-site mutation rate for each of the five rate classes in model 3 are shown in Figure 2.

Figure 3 shows the posterior distribution of θ for the purine and pyrimidine sites in model 4, under which

TABLE 1

Summary statistics for θ

θ	Model 1	Model 2	Model 3	Model 4
Mean	0.041	0.039	0.055	0.041
Median	0.040	0.038	0.054	0.040
25th percentile	0.035	0.033	0.046	0.035
75th percentile	0.047	0.045	0.062	0.047

both mutation rates had uniform prior distributions. The mean rate for purine sites is 0.015 compared to a mean rate of 0.062 for pyrimidine sites. We note the considerable range of plausible values for the pyrimidine rate. We may use the Bayesian approach to estimate the maximum-likelihood estimate (MLE) of the two rates. We obtained estimates of 0.012 for purines and 0.059 for pyrimidines (data not shown). These estimates can be compared to those of Lundstrom *et al.* (1992a) and Kuhner *et al.* (1995), who analyze purine and pyrimidine sites separately. The first authors obtained rates of 0.008 for purines and 0.024 for pyrimidines; the second, 0.005 and 0.052, respectively. The joint analysis of the two types of site tells a somewhat different story; the purine rate is much higher than the separate analyses might have predicted. Presumably this arises because, for this particular data set, including information about pyrimidines when analyzing purines happens to encourage trees to be shorter than they would be if purines were analyzed alone.

We conclude by returning to the results from the simplest mutation mechanism, given by model 2. We have taken $\kappa = 100$ and a uniform prior on (0, 100) for θ. Since the posterior density of θ is proportional to the likelihood in this case, we may use an estimate of the posterior density to find the maximum-likelihood estimate of θ. From the density shown in Figure 1, we obtained an MLE of $\hat{\theta} = 0.038$. Kuhner *et al.* (1995) obtained the value $\hat{\theta} = 0.040$ for these data, using the same value of κ. The difference in estimates arises from a combination of the parameters chosen for the density estimation, the different approaches to the optimization, and noise. From an estimate of the curvature of the log density we get an estimate of the standard error of $\hat{\theta}$ of 0.010, resulting in an ~95% confidence interval of (0.018, 0.058). The Watterson (1975) estimator of θ, based on 26 segregating sites in the data, is 0.015 with an estimated standard error of 0.005; the 95% confidence interval for θ is then (0.005, 0.025). The lower value obtained using the Watterson estimator is expected, because multiple mutations at the same site are ignored.

Goodness-of-fit: The adequacy of the fit of these models can be assessed using a variant of the parametric bootstrap. We simulated observations from the *posterior* distribution of (Λ, T, M), and for each of the trees (Λ, T) we then simulated the mutation process with parameters specified by M. The distribution of certain summary statistics observed in the simulated data is found, and the values of the statistics actually observed in the data are compared to these distributions. We chose to use the number of haplotypes, the maximal haplotype frequency, the haplotype homozygosity, the number of segregating sites, and a measure of nucleotide diversity. In practice, we use the output from the MCMC runs to generate the observations on (Λ, T, M).

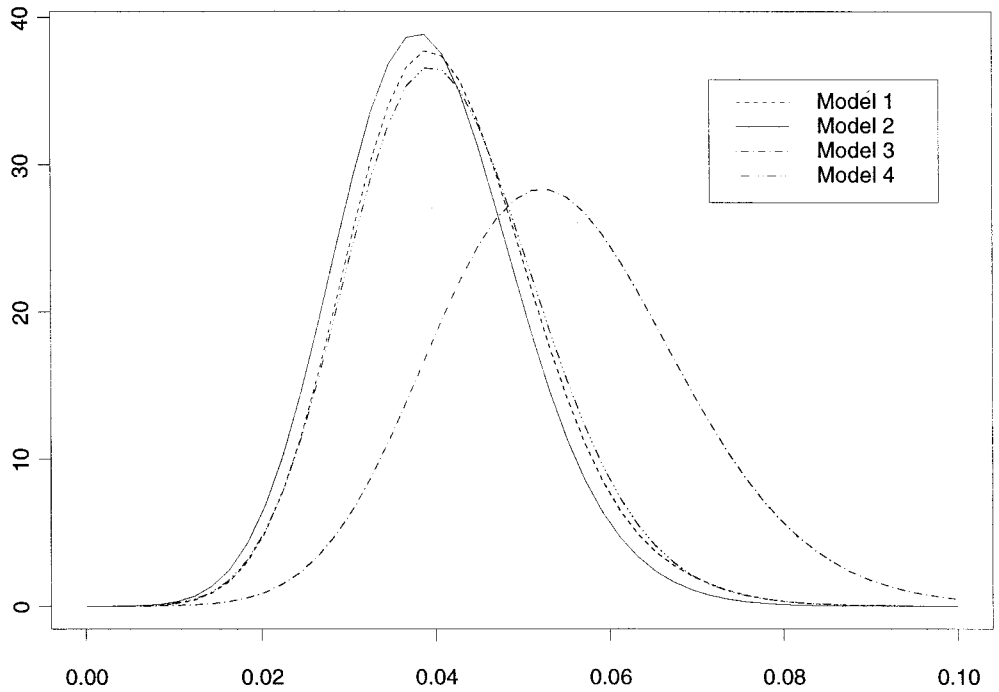

Figure 1.—Posterior density of per site mutation rate θ for each model.

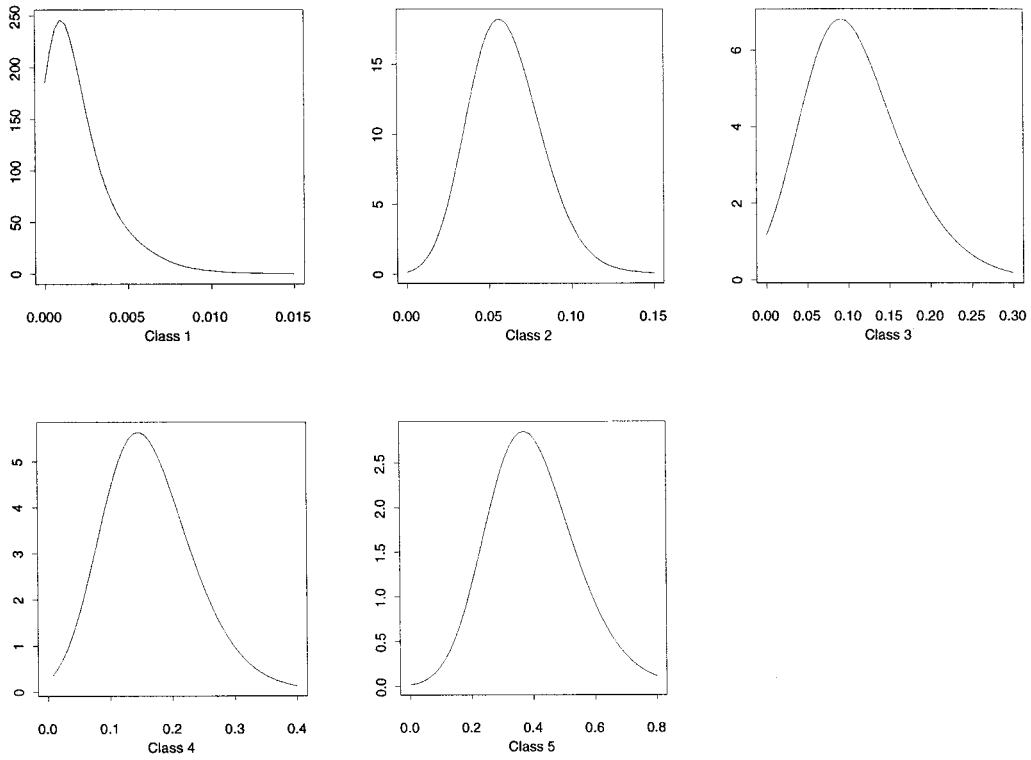

FIGURE 2.—Posterior density of per site θ's for model 3.

In Table 2 we give the results of this comparison for models 2 and 3 using 4000 values from each posterior distribution. In principle, for a perfect fit, we might expect to see ~50% of the simulations with values less than or equal to the observed value in the data. The criteria for accepting our model would be that the observed values fall with the central 95% of the distribution generated by the simulations. There is some evidence that the constant rate model fits less well than the variable rate case, particularly regarding the haplotype distribution. The number of segregating sites seen in each of the classes in model 3 also seems to fit reasonably well. The total number of segregating sites observed in the bootstrap samples gives some evidence of lack-of-fit; both models predict more segregating sites than are seen in the data. One explanation for this apparent

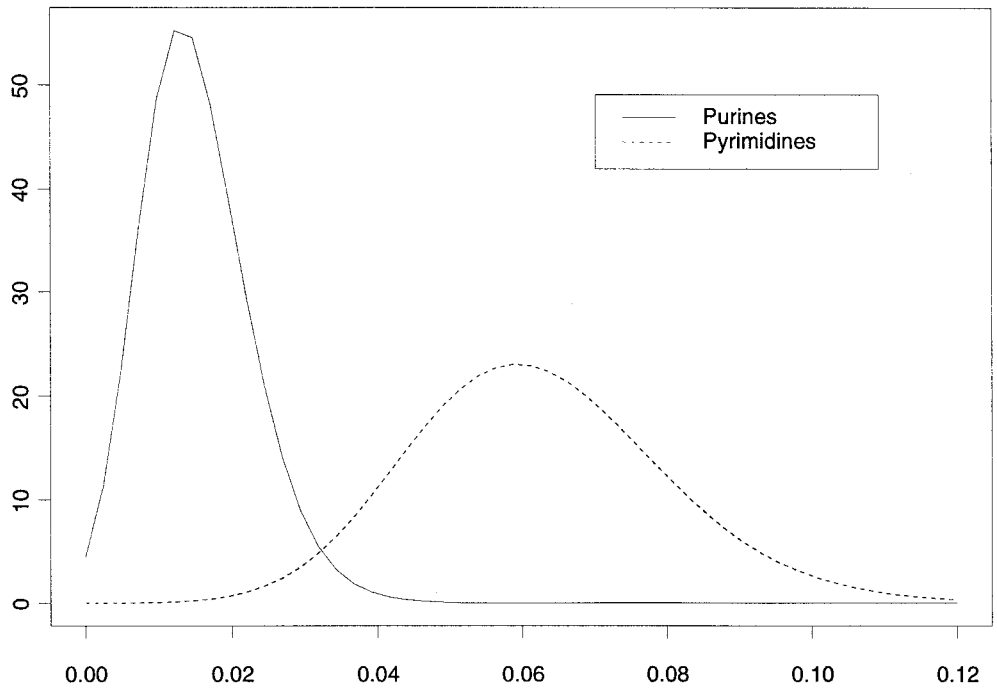

FIGURE 3.—Posterior density of per site θ for purines and pyrimidines under model 4.

TABLE 2

Assessing goodness-of-fit of the models

Statistic	Observed value	Fraction of simulations ≤ observed value	
		Model 3	Model 2
No. haplotypes	28	0.56	0.83
Max. haplotype frequency	9	0.54	0.36
Homozygosity	0.0562	0.32	0.12
Heterozygosity per site	0.0145	0.19	0.36
No. segregating sites	26	0.02	0.05
No. segregating in			
Class 1	0	0.51	
Class 2	9	0.12	
Class 3	3	0.24	
Class 4	6	0.35	
Class 5	8	0.20	

discrepancy might be that the models are not allowing for a high enough level of rate heterogeneity and therefore do not typically produce enough recurrent mutations. The mutations that do occur will then tend to be spread over a greater number of sites. A model that allows for more than five different mutation rate classes may well produce a better fit.

The distribution of time to the most recent common ancestor: Posterior statistics for the time to the most recent common ancestor (TMRCA) are given in Table 3, and the corresponding posterior densities appear in Figure 4. The prior distribution of the time to MRCA of a sample of $n = 63$ has a mean of $2(1 - 1/63) = 1.97$. With an effective size of $N = 600$, a 20-year generation time, and a value of $\sigma^2 = 1$ for the variance of the offspring distribution, this is ~23,600 years. The posterior distribution for each of the models presented here suggests that times considerably less than this are most likely. Note that the mean TMRCA for model 3 is slightly lower than that for the other models. This is a reflection of the higher estimate for θ under this model. Griffiths and Tavaré (1994) used the infinitely many sites model to obtain a mean posterior TMRCA of 14,400 years on the basis of an estimated per site θ of 0.014. This is substantially larger than the posterior mean of 7100 obtained from model 3, reflecting in part the smaller value of θ chosen. The joint posterior density of T and θ for model 3 is given in Figure 5.

TABLE 3

Summary statistics for time to MRCA

Time to MRCA	Model 1	Model 2	Model 3	Model 4
Mean	7800	8100	7100	7900
Median	7400	7700	6800	7500
25th percentile	6200	6500	5700	6300
75th percentile	9000	9300	8300	9100

Simultaneous estimation of *N* and *u*: Several authors (*e.g.*, Tavaré *et al.* 1997; Wilson and Balding 1998) have developed Bayesian methods that make inferences about the mutation rate u and the population size N separately. In the absence of other information, the statistical features of the sequence data are a consequence of the value of the compound mutation parameter $\nu = 2Nu$. Here we examine the influence of the assumed prior on the joint posterior distribution of N and u.

We illustrate this by considering the Nuu Chah Nulth data set once more. For simplicity we consider a case for which the prior distribution for N is uniform over the positive real line; this can be thought of as illustrating a case in which we wish to estimate an unknown effective population size. Using this framework the mode of the posterior distribution of N is the maximum-likelihood estimate of the population size. We use several different Normal prior distributions for u to indicate the influence the choice of prior distribution plays on the results. The mean of the Normal prior distribution for u, 7.6×10^{-5}, was chosen to be consistent with the results for θ given in this article. We vary σ, the standard deviation of the Normal prior, to reflect differing levels of uncertainty about u. We show results for three cases: case 1 is for $\sigma = 10^{-6}$, case 2 is for $\sigma = 2 \times 10^{-5}$, and case 3 is for $\sigma = 10^{-4}$. In all three cases the posterior distributions for the compound parameter $\nu = 2Nu$ are practically identical, *regardless of the choice of prior for u* (see Figure 6, in which the posterior densities of the per-site effective mutation rate θ are plotted).

However, the similarity of the posterior distributions for θ conceals a more complicated story. In Figure 6 we show the posterior distributions of N and u for each of cases 1, 2, and 3. In each case, the central plot is a scatter plot showing sampled values of N and u. Along the edges of each scatter plot are the corresponding marginal distributions of N and u.

As we can see, N is highly correlated with u. While the marginal distributions look well behaved, they hide the nonidentifiability that is revealed in the scatter plot. The correlation is not apparent in case 1, because the choice of prior for u has restricted the parameters to a small region. The data, in isolation, can be used only for inference about ν (or θ, a multiple of ν). The data support a particular posterior distribution for ν, the single evolutionary parameter. The choice of priors for N and u dictates the way in which the posterior distribution for ν is decomposed into separate distributions. For example, if one uses a tight prior for u, *i.e.*, a prior with low variance as in case 1, one attains a relatively low variance for the posterior distribution of N. However, if one assigns a prior with high variance to u, as in case 3, the posterior for N changes dramatically. In the limit, when we fix u, for example, the posterior distribution

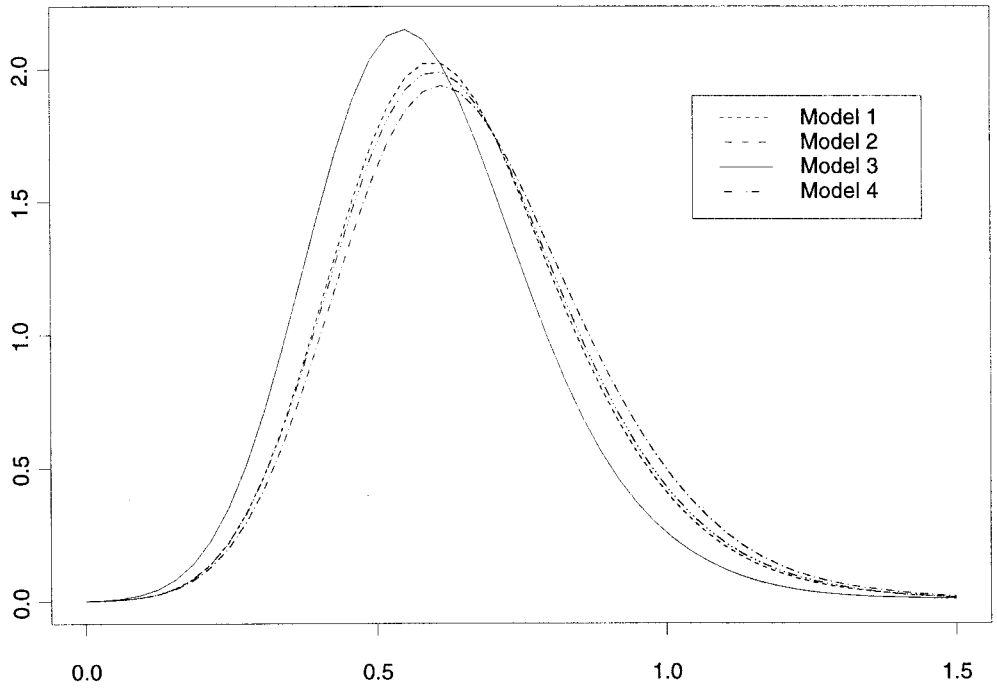

FIGURE 4.—Posterior density of time to MRCA for each model.

for N has a small variance. However, in this situation, the posterior for N is a rescaling of the posterior for v. In a Bayesian framework it is natural for the choice of prior to influence the shape of the posterior distribution, but it is important to be aware of the extent of this influence when interpreting the results. Consequently, we feel there is merit to such an analysis only if one has strong beliefs about one of the two parameters. The same observations apply to analyses using related approaches, as in TAVARÉ *et al.* (1997), and WILSON and BALDING (1998).

We have deliberately chosen widely differing priors for u to illustrate the points made here. However, the priors are not unreasonable and the results presented stress the vital role played by the choice of prior distributions. In the absence of any prior information the parameters u and N are confounded in the following sense: any combination of u, N and u', N' such that $uN = u'N'$ will have the same posterior likelihood, hence the behavior seen in Figure 6. The presence of a prior distribution for N and/or u lessens the degree of confounding, although, as can be seen, priors offering reasonable support to a wide range of parameter values will still leave a large degree of nonidentifiability.

DISCUSSION

We have described an MCMC approach that allows for the estimation of mutation parameters and statistics, such as the time to the MRCA, in a Bayesian setting. Such a setting allows reasonable prior information to be included in the analysis, possibly avoiding some of the confounding issues that might otherwise be present. This approach also provides an alternative method for finding more traditional maximum-likelihood estimates of mutation rates by using noninformative priors.

We noted that the adequacy of the model may be assessed using a parametric bootstrap based on simulating the mutation process using trees and mutation parameters simulated from the posterior distribution. We saw that the distribution of quantities such as the TMRCA of the sample is highly dependent on the mutation model used. For the variable rate model with five rate classes, which provides an adequate fit to the data, the estimated mean TMRCA is about one-half that estimated using an infinitely many sites model. This example emphasizes the need for caution in practical applications of theoretical approaches to ancestral inference.

It is straightforward to adapt the algorithm given in this article to allow for other demographic scenarios

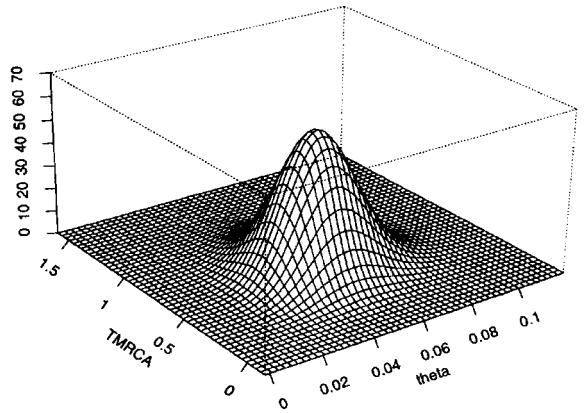

FIGURE 5.—Joint posterior density of TMRCA and θ under model 3.

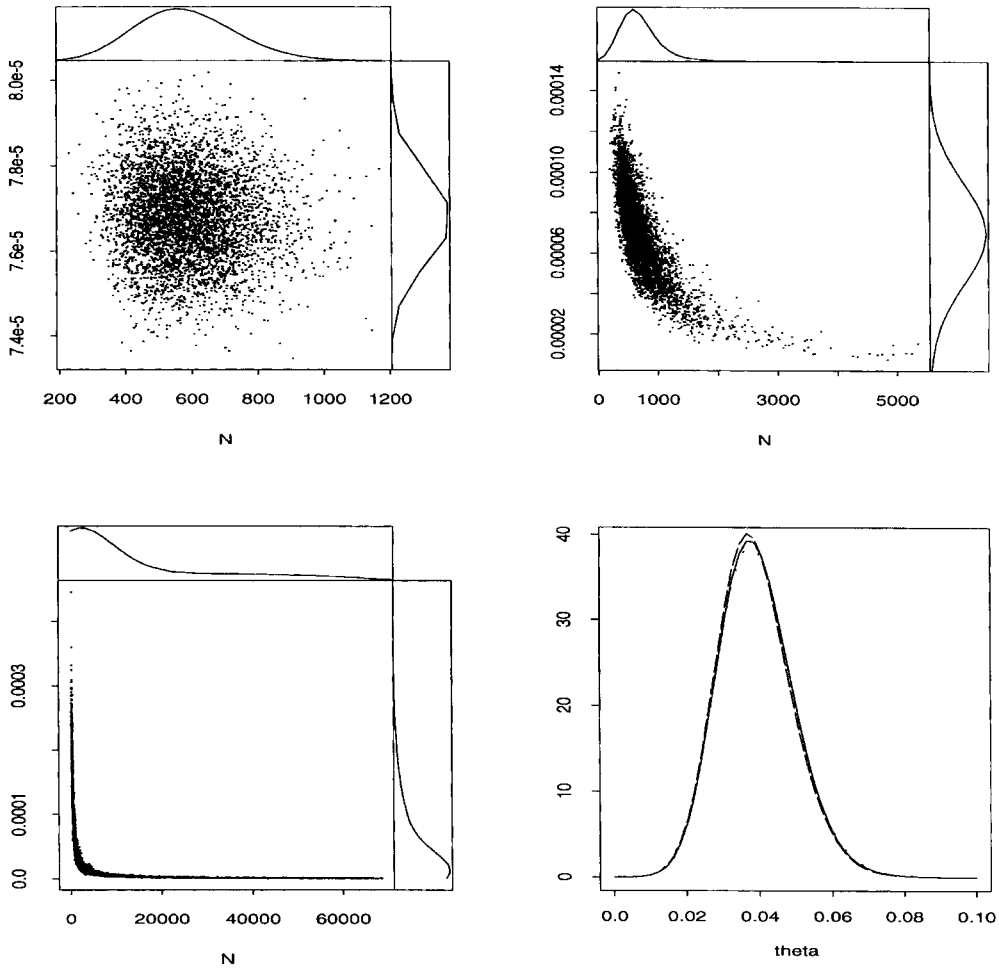

FIGURE 6.—Posterior distributions for N and u. Top left, case 1. Top right, case 2. Bottom left, case 3. Bottom right, posterior density of θ.

such as deterministic population size fluctuations. It is also straightforward to adapt the algorithm to include different mutation models. For an example concerning the age of a unique event polymorphism, see MARKOVTSOVA *et al.* (2000). We are currently developing an approach to rate variation that allows sites to change the rate class to which they belong, in contrast to the approach described in this article in which the class is fixed. Code that implements the methods described in this article is available in the form of C++ source code and executables from the authors and at http://hto-e.usc.edu.

We thank the referees for helpful comments on an earlier version of this article. The authors were supported in part by National Science Foundation grant BIR 95-04393 and National Institutes of Health grant GM 58897.

LITERATURE CITED

ARIS-BROSOU, S., and L. EXCOFFIER, 1996 The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol. Biol. Evol. **13:** 494–504.

BEST, N. G., M. K. COWLES and S. K. VINES, 1995 *CODA Manual Version 0.30.* MRC Biostatistics Unit, Cambridge, UK.

CANNINGS, C., 1974 The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. Adv. Appl. Prob. **6:** 260–290.

DENG, W-H., and Y-X. FU, 1996 The effects of variable mutation rates across sites on the phylogenetic estimation of effective population size of mutation rate of DNA sequences. Genetics **144:** 1271–1281.

DONNELLY, P., and S. TAVARÉ, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401–421.

EXCOFFIER, L., and Z. YANG, 1999 Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. Mol. Biol. Evol. **16:** 1357–1368.

FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequence data: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

GILKS, W. R., S. RICHARDSON and D. J. SPIEGELHALTER (Editors), 1996 *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London/New York.

GRIFFITHS, R. C., and S. TAVARÉ, 1994 Ancestral inference in population genetics. Stat. Sci. **9:** 307–319.

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

HUDSON, R. R., 1991 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. FUTUYMA and J. ANTONOVICS. Oxford University Press, Oxford/New York/London.

KINGMAN, J. F. C., 1982 On the genealogy of large populations. J. Appl. Prob. **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.

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.

Lundstrom, R. S., S. Tavaré and R. H. Ward, 1992a Estimating substitution rates from molecular data using the coalescent. Proc. Natl. Acad. Sci. USA **89:** 5961–5965.

Lundstrom, R. S., S. Tavaré and R. H. Ward, 1992b Modeling the evolution of the human mitochondrial genome. Math. Biosci. **112:** 319–336.

Markovtsova, L., P. Marjoram and S. Tavaré, 2000 The age of a unique event polymorphism. Genetics **156:** 401–409.

Mau, R., 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 machine. J. Chem. Phys. **21:** 1087–1091.

Meyer, S., G. Weiss and A. von Haeseler, 1999 Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. Genetics **152:** 1103–1110.

Misawa, K., and F. Tajima, 1997 Estimation of the amount of DNA polymorphism when the neutral mutation rate varies among sites. Genetics **147:** 1959–1964.

Sigurdardóttir, S., A. Helgason, J. R. Gulcher, K. Stefansson and P. Donnelly, 2000 The mutation rate in the human mtDNA control region. Am. J. Hum. Genet. **66:** 1599–1609.

Tajima, F., 1996 The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. Genetics **143:** 1457–1465.

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.

Tavaré, S., D. Balding, R. C. Griffiths and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145:** 505–518.

Thorne, J. L., H. Kishino and J. Felsenstein, 1992 Inching towards reality: an improved likelihood model of sequence evolution. J. Mol. Evol. **34:** 3–16.

Wakeley, J., 1993 Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. **37:** 613–623.

Ward, R. H., B. L. Frazier, K. Dew and S. Pääbo, 1991 Extensive mitochondrial diversity within a single Amerindian tribe. Proc. Natl. Acad. Sci. USA **88:** 8720–8724.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

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

Yang, Z., 1996a Among-site rate variation and its impact on phylogenetic analyses. TREE **9:** 367–372.

Yang, Z., 1996b Statistical properties of a DNA sample under the finite-sites model. Genetics **144:** 1941–1950.

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: G. A. Churchill

APPENDIX

Mutation model: We model only the effects of base substitutions, which we assume to occur according to a finitely many sites model as follows. In copying a sequence of L sites from one generation to the next, a potential substitution occurs with probability u. This potential substitution occurs at site i with probability p_i, where $p_i \geq 0$ and $p_1 + \cdots + p_L = 1$. Values of $p_i = 0$ correspond to invariable sites, and differences in the p_i reflect heterogeneities in the substitution rates at each site. A potential substitution at site i changes a base from j to k with probability $p_{jk}^{[i]}$. Here and in what follows we number the bases A, G, C, T as 1, 2, 3, and 4, respectively. The mutation matrices $P^{[i]}$ can in principle be different, but here we assume they are identical,

$$P^{[i]} := P, \quad i = 1, 2, \ldots, L,$$

so that we can focus on the effects of rate variation alone. We assume that P has stationary distribution $\pi = (\pi_A, \pi_G, \pi_C, \pi_T)$, so that at stationarity a randomly chosen sequence looks as though its bases are independent copies of π. In common with most studies of molecular variation we assume that π is known, usually from estimates of the proportion of each base in the set of sequences being analyzed.

In the coalescent setting, we define

$$\nu = 2Nu,$$

so that the locations of potential substitutions at each site on the branches of the coalescent tree evolve as independent Poisson processes, the process for site i having rate $\nu_i/2$, where $\nu_i = \nu p_i$. At any substitution point, the current base at site i is changed according to the transition matrix P. We note that not all potential substitutions have to result in changes to the existing base at a site, as $p_{jj} > 0$ is allowed. The effective substitution rate $\theta_i/2$ at site i is the expected number of substitutions per unit time that results in a change in base at site i:

$$\frac{\theta_i}{2} = \frac{\nu_i}{2}\sum_{j=1}^{4}\pi_j(1 - p_{jj}). \tag{3}$$

We write $\Theta = \theta_1 + \cdots + \theta_L$ for the overall effective mutation rate and

$$\theta = \Theta/L \tag{4}$$

for the average rate per site.

It remains to specify the form of P. Since our applications focus on mitochondrial DNA, in which transitions occur with much higher frequency than transversions, we use a model of Felsenstein. When a potential substitution occurs, it may be one of two types: *general*, in which case an existing base j is substituted by a base of type k with probability π_k, $1 \leq j, k \leq 4$; or *within-group*, in which case a pyrimidine is replaced by C or T with probability proportional to π_C and π_T, respectively, and a purine is replaced by A or G with probability proportional to π_A and π_G, respectively. The conditional probability of a general mutation is defined to be $1/(1 + \kappa)$, while the conditional probability of a within-group mutation is defined to be $\kappa/(1 + \kappa)$, where $\kappa \geq 0$ is the transition-transversion parameter. If we write

$$g_i = \frac{\nu_i}{2(1 + \kappa)}, \quad w_i = \kappa g_i, \tag{5}$$

then κ is the ratio of the within-class to general substitu-

tion rates. From (3), the effective mutation rate at site i is given by

$$\frac{\theta_i}{2} = g_i\left(1 - \sum_{j=1}^{4}\pi_j^2\right) + 2w_i\left(\frac{\pi_A\pi_G}{\pi_A + \pi_G} + \frac{\pi_C\pi_T}{\pi_C + \pi_T}\right). \tag{6}$$

We denote by $r_{jk}(t; g, w)$ the probability that a base of type j has changed to a base of type k a time t later, when the rate parameters in (5) are given by g and w, respectively. Thorne *et al.* (1992) show that

$$r_{jk}(t; g, w) = \begin{cases} e^{-(g+w)t} + e^{-gt}(1 - e^{-wt})\dfrac{\pi_k}{\pi_{H(k)}} + (1 - e^{-gt})\pi_k, \\ \quad j = k \\ e^{-gt}(1 - e^{-wt})\dfrac{\pi_k}{\pi_{H(k)}} + (1 - e^{-gt})\pi_k, \\ \quad j \neq k,\ H(j) = H(k) \\ (1 - e^{-gt})\pi_k, \\ \quad H(j) \neq H(k), \end{cases} \tag{7}$$

where $\pi_R = \pi_A + \pi_G$, $\pi_Y = \pi_C + \pi_T$, and $H(i)$ denotes whether base i is a purine or a pyrimidine, so that $H(A) = H(G) = R$ and $H(C) = H(T) = Y$.

The Hastings ratio: Writing $G = (\Lambda, T, M)$, the kernel Q can be expressed as the product of three terms:

$$Q(G \to G') = Q_1(\Lambda \to \Lambda')Q_2(T \to T'|\Lambda \to \Lambda')Q_3(M \to M').$$

Consequently the Hastings ratio appearing in (2) can be written in the form

$$\frac{\mathbb{P}(\mathbf{D}|G')}{\mathbb{P}(\mathbf{D}|G)}\frac{p_1(\Lambda')}{p_1(\Lambda)}\frac{p_2(T')}{p_2(T)}\frac{p_3(M')}{p_3(M)}\frac{Q_1(\Lambda' \to \Lambda)}{Q_1(\Lambda \to \Lambda')}\frac{Q_2(T' \to T|\Lambda' \to \Lambda)}{Q_2(T \to T'|\Lambda \to \Lambda')}\frac{Q_3(M' \to M)}{Q_3(M \to M')}, \tag{8}$$

the unknown term $f(\mathbf{D})$ canceling. We can further simplify (8) by noting that, since pairs of lines are chosen uniformly to coalesce, all topologies are, *a priori*, equally likely. Hence $p_1(\Lambda') = p_1(\Lambda)$. Furthermore, our transition kernel changes only two of the times on the tree, T_l and T_{l-1}, say. Finally, it is easy to show that $Q_1(\Lambda \to \Lambda') = Q_1(\Lambda' \to \Lambda)$, reducing (8) to

$$\frac{\mathbb{P}(\mathbf{D}|G')}{\mathbb{P}(\mathbf{D}|G)}\frac{p_2(T')p_3(M')}{p_2(T)p_3(M)}\frac{f_l(t_l)f_{l-1}(t_{l-1})}{f_l(t_l')f_{l-1}(t_{l-1}')}\frac{Q_3(M' \to M)}{Q_3(M \to M')}, \tag{9}$$

where $f_l(\cdot)$ *and* $f_{l-1}(\cdot)$ are the densities of the time updating mechanism at levels l and $l-1$. If one uses a transition kernel that proposes new times that are exponential with parameter $l(l-1)/2$ at level l (*i.e.*, the unconditional coalescent distribution for times), then further cross-cancellation reduces (9) to

$$\frac{\mathbb{P}(\mathbf{D}|G')}{\mathbb{P}(\mathbf{D}|G)}\frac{p_3(M')}{p_3(M)}\frac{Q_3(M' \to M)}{Q_3(M \to M')}. \tag{10}$$

A similar simplification also follows if one proposes new mutation rates independently of the currently accepted rate.