Likelihood and Bayes Estimation of Ancestral Population Sizes in Hominoids Using Data From Multiple Loci

Ziheng Yang[1]

Galton Laboratory, Department of Biology, University College London, London WC1E 6BT, England

Manuscript received April 18, 2002
Accepted for publication September 6, 2002

ABSTRACT

Polymorphisms in an ancestral population can cause conflicts between gene trees and the species tree. Such conflicts can be used to estimate ancestral population sizes when data from multiple loci are available. In this article I extend previous work for estimating ancestral population sizes to analyze sequence data from three species under a finite-site nucleotide substitution model. Both maximum-likelihood (ML) and Bayes methods are implemented for joint estimation of the two speciation dates and the two population size parameters. Both methods account for uncertainties in the gene tree due to few informative sites at each locus and make an efficient use of information in the data. The Bayes algorithm using Markov chain Monte Carlo (MCMC) enjoys a computational advantage over ML and also provides a framework for incorporating prior information about the parameters. The methods are applied to a data set of 53 nuclear noncoding contigs from human, chimpanzee, and gorilla published by Chen and Li. Estimates of the effective population size for the common ancestor of humans and chimpanzees by both ML and Bayes methods are ~12,000–21,000, comparable to estimates for modern humans, and do not support the notion of a dramatic size reduction in early human populations. Estimates published previously from the same data are several times larger and appear to be biased due to methodological deficiency. The divergence between humans and chimpanzees is dated at ~5.2 million years ago and the gorilla divergence 1.1–1.7 million years earlier. The analysis suggests that typical data sets contain useful information about the ancestral population sizes and that it is advantageous to analyze data of several species simultaneously.

THE amount of sequence polymorphism in a population is mainly determined by the parameter $\theta = 4N\mu$, where N is the (effective) population size and μ is the mutation rate per nucleotide site per generation. In addition to its effect on the amount of neutral variation maintained in a population, the population size is also important in affecting the efficiency of natural selection and is a critical parameter in population genetics models of mutation and selection. It is important to competing theories of origin and evolution of modern humans. When an estimate of the mutation rate is available, as is assumed here, the population size can be obtained from estimates of θ. The population size of a present-day species can be estimated by using observed DNA sequence variation in the population (*e.g.*, Kreitman 1983; Fu 1994; Takahata *et al.* 1995; Yang 1997a). The population size of modern humans has been estimated to be ~10,000 (*e.g.*, Takahata *et al.* 1995; Zhao *et al.* 2000; Yu *et al.* 2001).

The population size of an extinct ancestral species, such as the common ancestor of humans and chimpanzees, is harder to estimate, but a few methods have been developed (see Edwards and Beerli 2000 and Satta *et al.* 2000 for extensive reviews). They require data from multiple loci and make use of the information in the conflict of gene trees among loci. For example, Takahata (1986) suggested a method for estimating the population size of the common ancestor of two closely related species when a pair of homologous DNA sequences are available from the two species at each locus. The average divergence at many loci provides information on the species divergence time, while the variation of sequence divergence among loci reflects the ancestral population size. The method of Takahata (1986) uses summary statistics from the data and was later extended to a full-likelihood analysis and to the case of three species, where the population sizes of the two extinct ancestors as well as the two speciation dates were estimated (Takahata *et al.* 1995).

In the case of three species, another simpler method has also been used (Nei 1987; Wu 1991; Hudson 1992). This approach uses the proportion of loci at which the gene tree does not match the species tree to estimate the ancestral population size and exploits the fact that ancestral polymorphism creates conflicts between the gene tree and the species tree. I refer to it as the *tree-mismatch* method. Its application to human and great ape sequence data has led to estimates of the population size for the ancestor of humans and chimpanzees that are much larger than estimated population sizes for modern humans (Ruvolo 1997; Chen and Li 2001).

[1] *Address for correspondence:* Department of Biology, University College London, Darwin Bldg., Gower St., London WC1E 6BT, England. E-mail: z.yang@ucl.ac.uk

For example, Chen and Li (2001) sequenced 53 noncoding contigs from human, common chimpanzee, gorilla, and orangutan, and estimated the population size of the common ancestor of humans and chimpanzees to range from 52,000 to 150,000, depending on the generation time used (15 or 20 years) and on whether parsimony or neighbor joining was used to infer the gene trees.

The tree-mismatch approach has room for improvement. First, the conflicts in the estimated gene trees among loci can be due to both ancestral polymorphism and sampling errors in tree reconstruction. As the sequences are highly similar, with few variable or informative sites at each locus, the reconstructed gene tree, no matter what method was used to infer it, is unreliable. Ignoring the sampling error in the estimated gene tree leads to overestimation of the conflict among loci and overestimation of the ancestral population size (see below). Second, the branch lengths in the gene tree provide, in a probabilistic sense, information about the ancestral population parameters, but are ignored by the method. The method clearly makes poor use of the information in the data. Third, the method assumes that one knows the species divergence times, while they might not be known. Indeed, some authors argue for the importance of accounting for ancestral polymorphism when estimating speciation dates (Takahata *et al.* 1995).

It seems advantageous to use information in both the conflicting gene genealogies as well as the branch lengths while accounting for their uncertainties. This can be achieved by using the likelihood method under a coalescent model developed by Takahata *et al.* (1995). However, the infinite-sites model assumed in the method is sometimes violated by the data. In this article, I extend the method of Takahata *et al.* for estimating ancestral population sizes, using data from three species. I use the nucleotide substitution model of Jukes and Cantor (1969) to correct for multiple substitutions at the same site. The model is also implemented in a Bayes framework, with Markov chain Monte Carlo (MCMC) used to achieve efficient calculation. The methods are applied to the data of Chen and Li (2001).

MAXIMUM-LIKELIHOOD ESTIMATION

The species phylogeny is represented ((12)3) and is assumed known. Species 1 and 2 diverged τ_1 generations ago while species 3 diverged τ_0 generations earlier (Figure 1a). In this article, species 1, 2, and 3 represent human (H), chimpanzee (C), and gorilla (G), respectively. The effective population size of the ancestor of species 1 and 2 is N_1, and that of the common ancestor of all three species is N_0. The mutation rate μ is measured as the number of nucleotide substitutions per site per generation. As rate and time are confounded in such analysis, the parameters of the model are $\theta_0 = 4N_0\mu$, $\theta_1 = 4N_1\mu$, $\gamma_0 = \tau_0\mu$, and $\gamma_1 = \tau_1\mu$ (Figure 1a). Collectively they are denoted $\Theta = \{\theta_0, \theta_1, \gamma_0, \gamma_1\}$. The data contain DNA sequences from multiple loci, with one individual sequenced from each of the three species at each locus. It is assumed that there is no recombination within a locus and free recombination between loci. The population is assumed to be random mating, with no gene flow after species divergence.

The likelihood function: Consider the probability distribution of the gene tree and its branch lengths at any locus i. Two cases are possible and are dealt with separately. Case I is represented by Figure 1b, where the gene tree is $T_0 = ((12)3)$, identical to the species tree, and sequences 1 and 2 coalesce between the two speciation events C and D. The coalescent times t_0 and t_1 are defined in Figure 1b. Case I occurs when $t_1 < \tau_0$, with probability

$$\psi = \int_0^{\tau_0} \frac{1}{2N_1} e^{-t_1/(2N_1)} dt_1 = 1 - e^{-2\gamma_0/\theta_1} \qquad (1)$$

(*e.g.*, Takahata *et al.* 1995). In case II, both coalescent events occur in the population ancestral to all three species (Figure 1, c–e). The three gene trees $T_1 = ((12)3)$, $T_2 = ((23)1)$, and $T_3 = ((31)2)$ occur with equal probability $(1 - \psi)/3$.

In case I (tree T_0 in Figure 1b), the coalescent time t_1 is an exponential random variable truncated at $t_1 < \tau_0$, and t_0 is an independent random variable with an exponential distribution

$$\begin{aligned} f_1(t_1|T_0) &= \frac{1}{2N_1} e^{-t_1/(2N_1)}/\psi, \quad 0 < t_1 < \tau_0, \\ f_1(t_0|T_0) &= \frac{1}{2N_0} e^{-t_0/(2N_0)}, \qquad 0 < t_0 < \infty. \end{aligned} \qquad (2)$$

Let branch lengths b_0 and b_1 in the gene tree of Figure 1b be defined as follows:

$$\begin{aligned} b_0 &= (\tau_0 + t_0 - t_1)\mu = \gamma_0 + t_0\mu - t_1\mu, \\ b_1 &= (\tau_1 + t_1)\mu = \gamma_1 + t_1\mu. \end{aligned} \qquad (3)$$

Note that b_0 is the length of the internal branch AB and b_1 is the distance from sequence 1 to ancestor B (Figure 1b). Given that the gene tree at locus i is $T_i = T_0$ and its branch lengths are b_{i0} and b_{i1}, the probability of observing sequence data D_i at locus i, $P(D_i|T_0, b_{i0}, b_{i1})$, is the likelihood in traditional molecular phylogenetics (Felsenstein 1981). Calculation of this probability is discussed in the next section. By averaging over the distribution of the random variables t_0 and t_1 (or b_0 and b_1), the probability of observing the data at locus i for case I is

$$\begin{aligned} f(D_i|T_0) &= \int_0^\infty \int_0^{\tau_0} P(D_i|T_0, \gamma_0 + t_0\mu - t_1\mu, \gamma_1 + t_1\mu) \\ &\quad \times f_1(t_1|T_0) f_1(t_0|T_0)\, dt_1 dt_0 \\ &= \frac{1}{\psi} \int_0^\infty \int_0^{2\gamma_0/\theta_1} P(D_i|T_0, \gamma_0 + \frac{1}{2}\theta_0 x_0 - \frac{1}{2}\theta_1 x_1, \gamma_1 + \frac{1}{2}\theta_1 x_1) \\ &\quad \times e^{-(x_0 + x_1)} dx_1 dx_0, \end{aligned} \qquad (4)$$

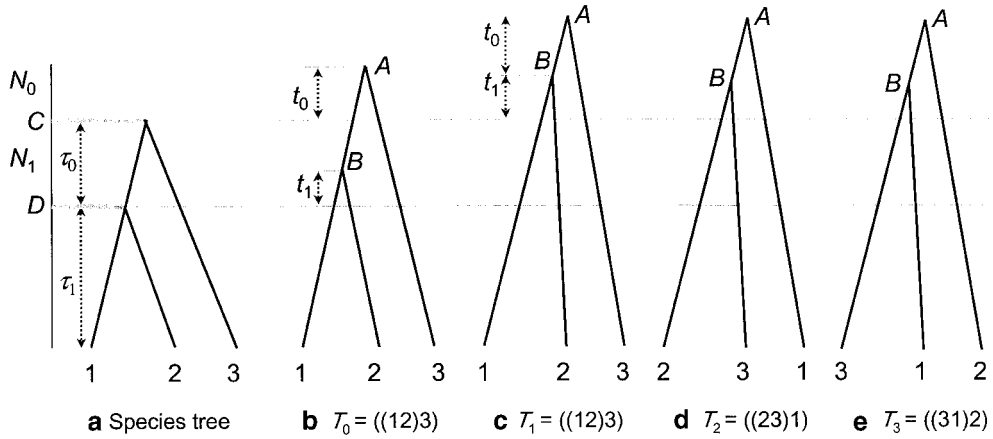

FIGURE 1.—(a) The species tree ((12)3) for three species: 1 (human), 2 (chimpanzee), and 3 (gorilla). Species 1 and 2 diverged τ_1 generations ago and species 3 diverged τ_0 generations earlier. The population sizes are N_0 in the common ancestor of all three species and N_1 in the common ancestor of species 1 and 2. The four parameters in the model are $\theta_0 = 4N_0\mu$, $\theta_1 = 4N_1\mu$, $\gamma_0 = \tau_0\mu$, and $\gamma_1 = \tau_1\mu$, where μ is the mutation rate per site per generation. There are four possible gene trees, represented by b–e. In case I (b), sequences 1 and 2 coalescence between the two speciation events and the gene tree T_0 = ((12)3) is consistent with the species tree. This is referred to as case I in the text. In c–e, both coalescence events occur in the common ancestor of all three species. In this case (case II), the gene tree can be T_1 = ((12)3), T_2 = ((23)1), or T_3 = ((31)2), with equal probability.

after a change of variables.

In case II (Figure 1, c–e), both coalescence events occur in the population ancestral to all three species. The three gene trees T_1, T_2, and T_3 have equal probabilities. The coalescent times t_0 and t_1, defined in Figure 1, c–e, are independent random variables with exponential distributions

$$f_{II}(t_1|T_k) = \frac{3}{2N_0}e^{-3t_1/(2N_0)}, \quad 0 < t_1 < \infty,$$
$$f_{II}(t_0|T_k) = \frac{1}{2N_0}e^{-t_0/(2N_0)}, \quad 0 < t_0 < \infty, \tag{5}$$

for $k = 1$, 2, or 3. Similarly, the branch lengths in the gene tree are defined as

$$b_0 = t_0\mu,$$
$$b_1 = \gamma_0 + \gamma_1 + t_0\mu. \tag{6}$$

Calculation of the probability, $P(D_i|T_k, b_{i0}, b_{i1})$, of observing data D_i at locus i, given the gene tree $T_i = T_k$ ($k = 1, 2, 3$) and its branch lengths b_{i0} and b_{i1}, will be described in the next section. The probability of observing the data at locus i for case II is

$$\begin{aligned} f(D_i|T_k) &= \int_0^\infty\int_0^\infty P(D_i|T_k, t_0\mu, \gamma_0 + \gamma_1 + t_1\mu) \\ &\quad \times f_{II}(t_1|T_k)f_{II}(t_0|T_k)\,dt_1dt_0 \\ &= \int_0^\infty\int_0^\infty P(D_i|T_k, \tfrac{1}{2}\theta_0x_0, \gamma_0 + \gamma_1 + \tfrac{1}{2}\theta_0x_1) \\ &\quad \times 3e^{-(x_0+3x_1)}dx_1dx_0, \end{aligned} \tag{7}$$

for $k = 1, 2, 3$.

Averaging over cases I and II or over the four gene trees T_0, T_1, T_2, T_3 in Figure 1, we obtain the marginal probability of observing data D_i at locus i as

$$\begin{aligned} f(D_i|\theta_0, \theta_1, \gamma_0, \gamma_1) &= \psi f(D_i|T_0) + \frac{1-\psi}{3}\sum_{k=1}^{3} f(D_i|T_k) \\ &= \int_0^\infty\int_0^{2\gamma_0/\theta_1} P(D_i|T_0, \gamma_0 + \tfrac{1}{2}\theta_0x_0 - \tfrac{1}{2}\theta_1x_1, \gamma_1 + \tfrac{1}{2}\theta_1x_1) \\ &\quad \times e^{-(x_0+x_1)}dx_1dx_0 \\ &\quad + e^{-2\gamma_0/\theta_1}\int_0^\infty\int_0^\infty \sum_{k=1}^{3} P(D_i|T_k, \tfrac{1}{2}\theta_0x_0, \gamma_0 + \gamma_1 + \tfrac{1}{2}\theta_0x_1) \\ &\quad \times e^{-(x_0+3x_1)}dx_1dx_0. \end{aligned} \tag{8}$$

The log likelihood is a sum over all the L loci,

$$\ell(\theta_0, \theta_1, \gamma_0, \gamma_1|D) = \sum_{i=1}^{L}\log\{f(D_i|\theta_0, \theta_1, \gamma_0, \gamma_1)\}, \tag{9}$$

where $D = \{D_i\}$ represents data at all L loci. Parameters θ_0, θ_1, γ_0, and γ_1 are estimated by numerical maximization of the log-likelihood function. A C-optimization routine from the PAML package (YANG 1997b) was used. Each likelihood calculation requires evaluation of $2L$ two-dimensional integrals (Equation 8), which are calculated numerically using Mathematica. The Mathlink library was used to establish communication between the C routine and Mathematica. For the data of CHEN and LI (2001) with $L = 53$ loci, the ML iteration takes ~1 hr on a Pentium III PC at 1.2 GHz.

Probability of data at a locus given the gene tree and branch lengths: Given the gene tree T_i, which is one of T_0, T_1, T_2, or T_3 in Figure 1b, and its branch lengths (b_{i0} and b_{i1}), the probability of observing the data, $P(D_i|T_i, b_{i0}, b_{i1})$, in Equation 8 can be calculated under any model of nucleotide substitution (LIO and GOLDMAN 1998). The model of JUKES and CANTOR (1969) is used in this article, which seems sufficient for such highly similar sequences. Under this model, the sequence data can be summarized as counts of five possible site patterns

TABLE 1

Number of site patterns from the data of Chen and Li (2001)

Locus (i)	n_i	n_{i0}	n_{i1}	n_{i2}	n_{i3}	n_{i4}	$d_{\text{HCG-O}}$	Rate
1-2609	472	462	3	3	4	0	0.0299	1.010
2-1251	531	509	13	4	5	0	0.0662	2.236
3-2659	560	551	4	1	4	0	0.0206	0.696
7-2012	528	511	10	2	5	0	0.0284	0.959
8-1364	475	465	6	3	1	0	0.0295	0.996
9-1386	484	471	3	3	7	0	0.0437	1.476
10-1412N	474	462	8	3	1	0	0.0254	0.858
10-2207	480	475	3	2	0	0	0.0337	1.138
10-2215-3	515	505	4	3	3	0	0.0278	0.939
10-2891-2	545	538	1	3	3	0	0.0227	0.767
11-1419-2	474	464	4	1	5	0	0.0317	1.071
11-2224	371	369	0	0	2	0	0.0200	0.675
11-73646	463	451	5	1	6	0	0.0294	0.993
12-1482-2	368	359	4	2	3	0	0.0299	1.010
12-2906	396	390	5	0	1	0	0.0259	0.875
12-2924	492	479	6	3	4	0	0.0229	0.773
12-2927-2	471	461	3	5	2	0	0.0298	1.006
14-2960	301	297	3	1	0	0	0.0248	0.838
14-2963	366	360	6	0	0	0	0.0253	0.854
15-2265-2	459	450	5	2	2	0	0.0259	0.875
15-2266	510	497	8	1	4	0	0.0243	0.821
16-598D4	518	511	3	0	4	0	0.0164	0.554
17-0787	450	443	2	3	2	0	0.0264	0.892
17-0801	491	485	3	2	1	0	0.0321	1.084
17-0812-2	374	361	8	2	3	0	0.0348	1.175
17-0813	444	436	4	2	2	0	0.0291	0.983
17-1574	359	352	4	1	2	0	0.0226	0.763
17-2294	514	500	9	4	1	0	0.0277	0.936
17-2987	497	486	4	1	5	1	0.0285	0.963
17-2988	494	481	7	4	2	0	0.0306	1.034
17-0784	462	454	3	1	4	0	0.0484	1.635
17-276O15	433	419	7	3	4	0	0.0324	1.094
17-2984	502	483	8	7	4	0	0.0245	0.827
17-2986	419	412	2	2	3	0	0.0185	0.625
18-0864	373	360	7	5	1	0	0.0415	1.402
18-0866	443	434	2	4	3	0	0.0175	0.591
18-1506	461	456	3	0	2	0	0.0257	0.868
18-1584	481	470	1	5	5	0	0.0400	1.351
18-2558	443	434	5	4	0	0	0.0372	1.256
19-0946	320	310	6	4	0	0	0.0169	0.571
19-0953	479	474	2	1	1	1	0.0240	0.811
20-1636	535	517	4	10	4	0	0.0447	1.510
20-2012	511	502	5	2	2	0	0.0391	1.321
20-2018	535	522	5	4	4	0	0.0319	1.077
20-2019	410	401	5	2	2	0	0.0299	1.010
20-2020	450	439	3	3	5	0	0.0334	1.128
20-2064	512	504	6	1	1	0	0.0260	0.878
20-2085	542	534	1	2	5	0	0.0219	0.740
20-2352	511	502	2	4	3	0	0.0279	0.942
20-2472	452	444	4	2	2	0	0.0520	1.756
20-2560	517	510	5	1	0	1	0.0286	0.966
20-2563	545	535	1	4	5	0	0.0217	0.733
20-2568	487	480	2	4	1	0	0.0195	0.659

n_{i0}, n_{i1}, n_{i2}, n_{i3}, n_{i4} are counts of sites with patterns *xxx*, *xxy*, *yxx*, *xyx*, and *xyz* in human (H), chimpanzee (C), and gorilla (G), while $n_i = n_{i0} + n_{i1} + n_{i2} + n_{i3} + n_{i4}$ is the total number of sites at locus i.

(configurations): *xxx*, *xxy*, *yxx*, *xyx*, and *xyz*, where *x*, *y*, and *z* are any three different nucleotides. The data at locus i can thus be represented by the number of sites with those five site patterns: $D_i = \{n_{i0}, n_{i1}, n_{i2}, n_{i3}, n_{i4}\}$. The observed number of counts from the data of Chen and Li (2001) is listed in Table 1.

For gene trees T_0 or T_1, the probabilities of observing the five site patterns are

$$\begin{aligned}
p_0(b_0, b_1) &= \text{prob}(xxx)\\
&= (1 + 3e^{-8b_1/3} + 6e^{-8(b_0+b_1)/3} + 6e^{-(8b_0+12b_1)/3})/16,\\
p_1(b_0, b_1) &= \text{prob}(xxy)\\
&= (3 + 9e^{-8b_1/3} - 6e^{-8(b_0+b_1)/3} - 6e^{-(8b_0+12b_1)/3})/16,\\
p_2(b_0, b_1) &= \text{prob}(yxx)\\
&= (3 - 3e^{-8b_1/3} + 6e^{-8(b_0+b_1)/3} - 6e^{-(8b_0+12b_1)/3})/16,\\
p_3(b_0, b_1) &= \text{prob}(xyx)\\
&= (3 - 3e^{-8b_1/3} + 6e^{-8(b_0+b_1)/3} - 6e^{-(8b_0+12b_1)/3})/16 = p_2,\\
p_4(b_0, b_1) &= \text{prob}(xyz)\\
&= (6 - 6e^{-8b_1/3} - 12e^{-8(b_0+b_1)/3} + 12e^{-(8b_0+12b_1)/3})/16
\end{aligned} \tag{10}$$

(Yang 1994). For gene trees T_2 and T_3 (Figure 1, d and e), the probabilities can be obtained by considering the symmetry of the problem. Thus with functions p_0–p_4 defined above, the probability of data at locus i, conditional on the gene tree T_k, $k = 1$ (or 0), 2, 3, and its branch lengths b_0 and b_1, is given by the multinomial distribution

$$\begin{aligned}
P(D_i|T_1, b_0, b_1) &= C \times p_0^{n_{i0}} p_1^{n_{i1}} p_2^{n_{i2}+n_{i3}} p_4^{n_{i4}},\\
P(D_i|T_2, b_0, b_1) &= C \times p_0^{n_{i0}} p_1^{n_{i2}} p_2^{n_{i3}+n_{i1}} p_4^{n_{i4}},\\
P(D_i|T_3, b_0, b_1) &= C \times p_0^{n_{i0}} p_1^{n_{i3}} p_2^{n_{i1}+n_{i2}} p_4^{n_{i4}},
\end{aligned} \tag{11}$$

where $C = n_i!/\Pi_{j=0}^{4} n_{ij}!$, and $n_i = n_{i0} + n_{i1} + n_{i2} + n_{i3} + n_{i4}$.

Mutation rate variation among loci: An important factor that may influence the estimation of ancestral population sizes is the variation of mutation rates among loci (Yang 1997a; Chen and Li 2001). For example, estimation of ancestral population size from comparison between two species was found to be sensitive to even slight rate variation (Yang 1997a). If relative rates for the loci are available, it will be straightforward to incorporate them in the likelihood calculation (Yang 1997a). In this article, I use the average distance from the orangutan to the three African apes to calculate the relative rate for the locus (Table 1). This *ad hoc* approach appears sensible since the orangutan diverged from the African apes very early and ancestral polymorphism in their common ancestor does not seem important. The likelihood calculation proceeds as before except that the branch lengths for the gene tree at each locus are multiplied by the relative rate for that locus.

Application to the data of Chen and Li: Chen and Li (2001) sequenced one individual from each of the four

TABLE 2

Maximum-likelihood estimates of parameters

Parameter	One rate for all loci	Variable rates among loci
$\hat{\theta}_0$ ($\hat{N}_0$)	0.003057 (38,000)	0.002348 (29,000)
$\hat{\theta}_1$ ($\hat{N}_1$)	0.000990 (12,000)	0.001650 (21,000)
$\hat{\gamma}_0$ (time in MY)	0.001089 (1.1 MY)	0.001704 (1.7 MY)
$\hat{\gamma}_1$ (time in MY)	0.005194 (5.2 MY)	0.004936 (4.9 MY)
ℓ	−3,099.41	−3,100.01

In converting θ into N and γ into speciation time, the generation time is assumed to be $g = 20$ years in all species and the mutation rate to be 10^{-9} substitutions per site per year.

species, human, chimpanzee, gorilla, and orangutan, at 53 independent noncoding loci (contigs). An advantage of the data set is that the sequences are expected to be outside and far away from coding regions and not affected by selection at linked sites or loci. The model of this article assumes the molecular clock and uses only three species. The counts of site patterns are listed in Table 1. At some loci, the total number of sites used in this article is larger than that in Chen and Li (2001; Table 1), because some sites had alignment gaps in the orangutan and were removed by Chen and Li.

The three-species data are analyzed by the maximum-likelihood (ML) method of this article. The estimates of parameters are given in Table 2. If we assume a generation time of 20 years and a mutation rate of 10^{-9} substitutions per site per year (*e.g.*, Nachman and Crowell 2000), the estimates suggest a population size for the ancestor of humans and chimpanzees of ~12,000. This is several times smaller than the estimates of Chen and Li (2001) from the same data, at a minimum of 52,000. The estimate is also similar to estimates of the population size of modern humans, for example, 12,000 by Yu *et al.* (2001). The population size for the common ancestor of all three species is estimated to be ~38,000. The same analysis estimated the human-chimpanzee divergence time at 5.1 million years ago (MYA) and the gorilla divergence at ~1.1 million years (MY) earlier. Those estimates are largely consistent with those of previous studies (*e.g.*, Hasegawa *et al.* 1985; Ruvolo 1997; Kumar and Hedges 1998; Yoder and Yang 2000). Figure 2a shows the likelihood surface as a function of θ_0 and θ_1 when γ_0 and γ_1 are fixed at their maximum-likelihood estimates (MLEs). The ~95% confidence region is given by the likelihood contour at 3.32 units below the optimum, that is, at −3102.73 (Figure 2a). The sampling errors are quite large. Analysis of the human and chimpanzee sequences at the 53 loci using the ML method of Takahata *et al.* (1995) under the infinite-sites model leads to $\hat{\theta}_1 = 0.0017$ and $\hat{\gamma}_1 = 0.0055$ (N. Takahata, personal communication). Those estimates are similar to the MLEs of Table 2.

To examine the effect of mutation rate variation among loci, I calculated the average sequence distance under the model of Jukes and Cantor (1969) from the human, chimpanzee, and gorilla to the orangutan, that is, $d_{HCG\text{-}O} = (d_{H\text{-}O} + d_{C\text{-}O} + d_{G\text{-}O})/3$. This is divided by the average across all loci to give the relative rate for that locus (Table 1). The average distance from the orangutan to the African apes is found to be 0.0296; this is consistent with a mutation rate of 10^{-9} substitutions per site per year and an orangutan divergence date of ~13 MYA (*e.g.*, Hasegawa *et al.* 1987) since $0.0296/(2 \times 13 \times 10^6) = 1.1387 \times 10^{-9}$. The relative rates for loci calculated this way are used as fixed constants in the likelihood calculation for the data of three species. The MLEs of parameters are shown in Table 2. Using the same generation time and mutation rate as above, we get the estimate of the population size for the ancestor of humans and chimpanzees to be 21,000, larger than the estimate under the assumption of a constant rate for all loci. The population size for the ancestor of all three species is estimated to be 29,000, smaller than under the constant-rate model. The differences between the two analyses are somewhat surprising, as one might expect the population sizes to be smaller when rate variation among loci is accounted for. However, it is noted that the distance from orangutan to the African apes has only a weak correlation (0.44) with the average distance within the H-C-G group, which appears to suggest that the mutation rates are rather homogeneous among loci and that the conflict among loci in sequence divergence is mainly caused by ancestral polymorphism.

If the average distances within the H-C-G group, $(d_{HC} + d_{CG} + d_{GH})/3$, are used as relative rates for the loci, parameter estimates become $\hat{\theta}_0 = 0.0000014$, $\hat{\theta}_1 = 0.002902$, $\hat{\gamma}_0 = 0.003229$, and $\hat{\gamma}_1 = 0.004555$, with $\ell = -3069.73$. Those correspond to a population size of 36,000 for the ancestor of humans and chimpanzees, a population size of only 18 for the ancestor of all three species, 4.5 MY for the H-C divergence, and 7.7 MY for the gorilla divergence. This calculation effectively attributes all variation in sequence divergence among loci to mutation rate variation and causes underestimation of θ_0 and γ_1 and overestimation of θ_1 and γ_0 (see also Table 2).

Comparison with the tree-mismatch method: When the gene tree is T_2 or T_3 (Figure 1), there is a mismatch between the species tree and the gene tree. This occurs with probability $P_{SG} = f(T_2) + f(T_3) = 2(1 - \psi)/3 = \frac{2}{3}e^{-2\gamma_0/\theta_1}$ (*e.g.*, Nei 1987, pp. 288–289). The tree-mismatch method estimates θ_1 by equating this probability to the proportion ($\hat{p}$) of loci at which the estimated gene tree differs from the species tree, with γ_0 being assumed known; that is, $\hat{\theta}_1 = -2\gamma_0/\log\{3\hat{p}/2\}$. Chen and Li (2001) used the orangutan to root the H-C-G tree and were able to resolve the gene tree at 33 loci, out of which 9 mismatches were found, at a proportion of 27.3%. Several coding loci were included as well, so that

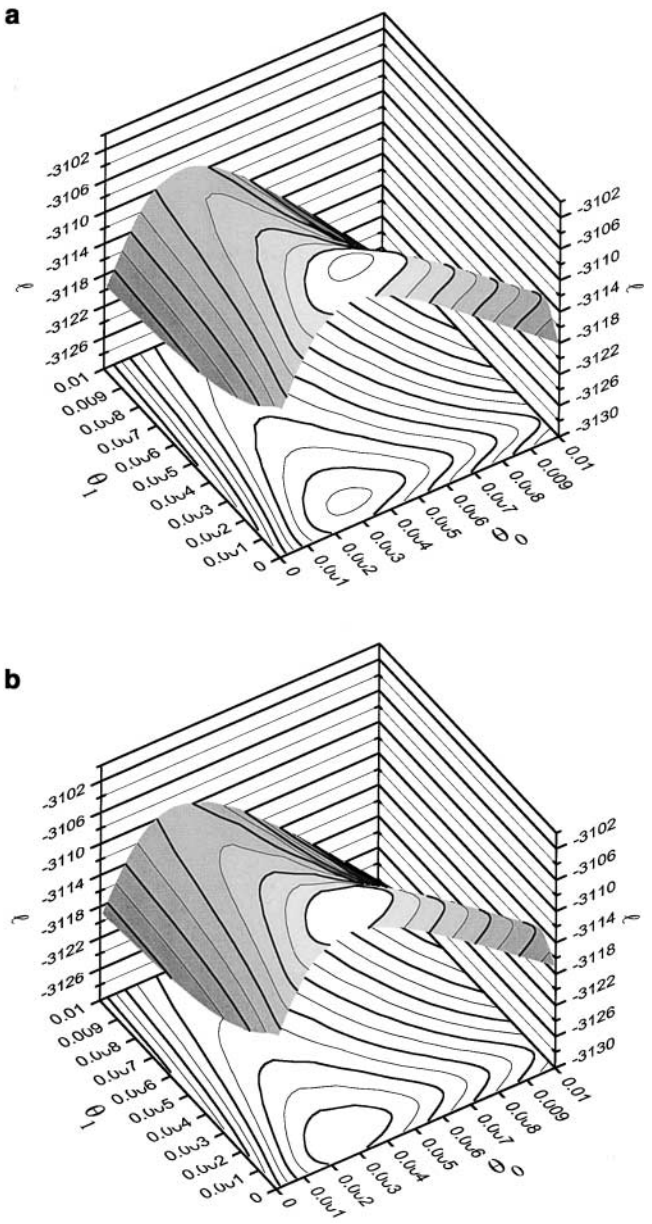

FIGURE 2.—Log-likelihood surface (contour) as a function of θ_0 and θ_1 when γ_0 and γ_1 are fixed at their MLEs. (a) The same substitution (mutation) rate is assumed for all loci. (b) Fixed relative rates obtained from comparison between the orangutan and the African apes (human, chimpanzee, and gorilla) are used to account for possible evolutionary rate variation among loci. Maximum-likelihood estimates of parameters are listed in Table 2.

16 mismatches were found at a total of 52 resolved loci, at the proportion 30.8%. The authors assumed an H-C divergence at $\gamma_1 = 1.6$ MYA and arrived at $\hat{\theta}_1 = 0.00414$, which, if the generation time is 20 years, corresponds to a minimum population size of $\hat{N}_1 = 52{,}000$ for the ancestor of humans and chimpanzees. In this article, the molecular clock has been assumed, which can also be used to root the H-C-G tree. Under the clock, T_1, T_2, or T_3 is the ML tree if n_{i1}, n_{i2}, or n_{i3} is the greatest among the three, respectively (SAITOU 1988; YANG 1994). The clock-rooting approach uses more "informative" sites than out-group rooting and resolves the gene tree at 49 of the 53 loci, out of which 18 are mismatches, at the proportion 36.7%. This proportion is even higher than those of Chen and Li and produces even larger estimates of θ_1 and N_1.

To understand the difference between the tree-mismatch method and ML, note that three aspects of the data are ignored by the tree-mismatch method and accounted for by ML: (i) uncertainty in the estimated gene tree due to the finite number of nucleotide sites at the locus, (ii) unresolved loci (ties), and (iii) branch lengths in the gene tree reflected in the sequence divergences. While all three probably contribute to the large differences discussed above, uncertainty in the estimated gene tree seems to be the predominant reason. A more "proper" tree-mismatch method should equate the observed proportion of mismatches ($\hat{p}$) not to P_{SG} but to P_{SE}, the probability of a mismatch between the species tree and the *estimated* gene tree. This probability is given by

$$P_{SE} = \sum_{\max\{n_{i2}, n_{i3}\} > n_{i1}} f(D_i|\theta_0, \theta_1, \gamma_0, \gamma_1), \qquad (12)$$

where f is the probability of data $D_i = \{n_{i0}, n_{i1}, n_{i2}, n_{i3}, n_{i4}\}$ in Equation 8, and the summation is over all data configurations in which the ML tree for the locus is either T_2 or T_3 (Figure 1). Unlike P_{SG}, P_{SE} is dependent on all four parameters, θ_0, θ_1, γ_0, and γ_1, as well as the sequence length n_i and appears no easier to calculate than the full likelihood (Equation 9). Instead I use Monte Carlo simulation to calculate those probabilities to assess the impact of errors in gene tree reconstruction on the difference between P_{SG} and P_{SE}. The MLEs of parameters in Table 2 ("constant rate") are used to generate gene trees, which are used to "evolve" sequences. The sites are counted to obtain the data $D_i = \{n_{i0}, n_{i1}, n_{i2}, n_{i3}, n_{i4}\}$, which are used to estimate the gene tree by ML.

Figure 3 shows the probabilities that the species tree (S), the gene tree (G), and the estimated gene tree (E) differ from each other. The probability of a mismatch between the species tree and the gene tree is $P_{SG} = 2(1 - \psi)/3 = 0.0739$, much lower than the observed mismatch proportion $\hat{p} = 0.367$. The probability of a mismatch between the species tree and the estimated gene tree is higher. With 466 sites (the average across the 53 loci, Table 1), $P_{SE} = 0.2028$, which is 2.7 times as high as P_{SG} (Figure 3). Now consider the four gene trees T_0, T_1, T_2, and T_3 (Figure 1), which occur with probabilities $f(T_0) = \psi = 0.8892$ and $f(T_1) = f(T_2) = f(T_3) = (1 - \psi)/3 = 0.0369$ (Equation 1). According to the simulation, the probability that the topology of the gene tree is incorrectly reconstructed when the true gene tree is T_0, T_1, T_2, or T_3 is $P_0 = 0.1453$ and $P_1 = P_2 = P_3 = 0.1981$. Note that for the estimated gene tree, we consider only the topology and disregard its divergence times relative to the speciation times. The

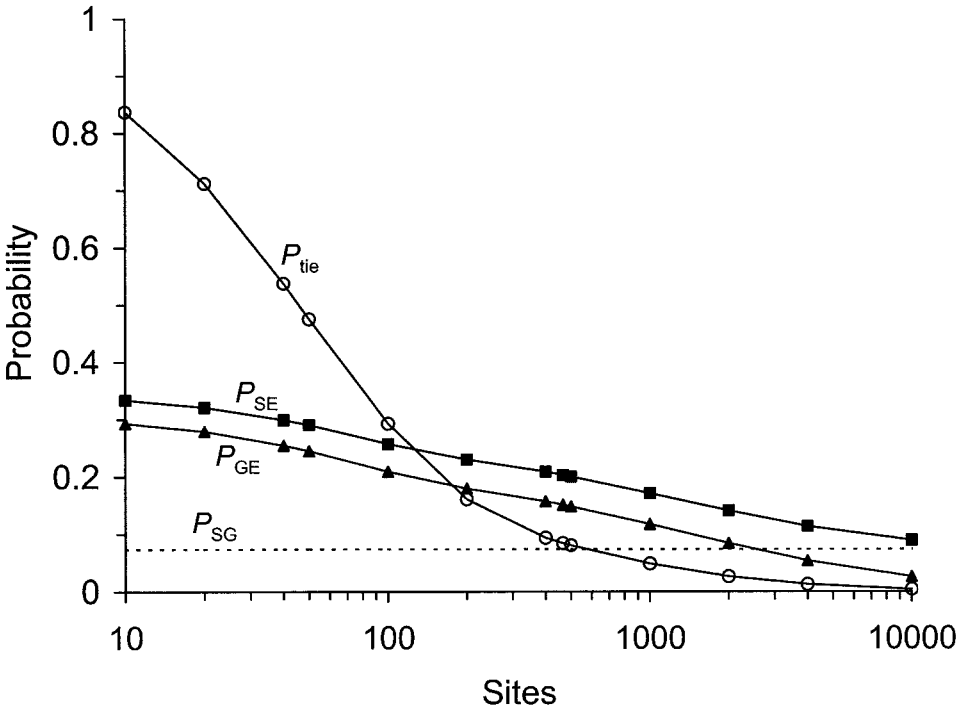

FIGURE 3.—Tree-mismatch probabilities calculated using Monte Carlo simulation plotted as functions of the sequence length. MLEs of parameters in Table 2 (one rate for all loci) are used in the simulation. Note that S, G, and E in the subscripts stand for the species tree, the gene tree, and the estimated gene tree (the ML tree), respectively. Thus P_{SG} is the probability that the species tree and the gene tree differ; this is 0.0739 for the parameter values used. P_{SE} is the probability of a mismatch between the species tree and the estimated gene tree, and P_{GE} is the probability of a mismatch between the gene tree and the estimated gene tree. P_{tie} is the proportion of replicates in which a tie occurs, that is, the two best trees are equally good. Ties are excluded in calculation of P_{SE} and P_{GE}. Ten million replicates were simulated for each sequence length.

average error probability of gene tree reconstruction is thus $P_{GE} = \sum_{i=0}^{3} f(T_i)P_i = 0.8892 \times 0.1453 + 3 \times 0.0369 \times 0.1981 = 0.1512$ (see Figure 3). When gene trees T_0 or T_1 are incorrectly reconstructed, the estimated gene tree will always be a mismatch with the species tree; such errors will cause an overcount of $f(T_0)P_0 + f(T_1)P_1 = 0.1366$. Conversely, when gene trees T_2 or T_3 are incorrectly reconstructed, the estimated tree will not be counted as a mismatch one-half of the time, so the undercount is $f(T_2)P_2/2 + f(T_3)P_3/2 = f(T_2)P_2 = 0.0074$. The difference between those two error rates gives rise to the net error due to gene tree reconstruction: $P_{SE} - P_{SG} = f(T_0)P_0 = \psi P_0 = 0.1290$. The above argument suggests that ignoring errors in gene tree reconstruction always causes overestimation of the mismatch between the species tree and the gene tree and leads to overestimation of the ancestral population size N_1. It is interesting that the bias in the tree-mismatch method is caused by reconstruction errors for gene tree T_0 alone and thus can be reduced if ψ is reduced, for example, if the two speciation events are very close or if the ancestral population size N_1 is large. Obviously factors that reduce the reconstruction error P_0, such as longer sequences (Figure 3) and higher mutation rates, will reduce the bias as well.

THE BAYES APPROACH USING MCMC

A Bayes approach is implemented under the same model, using MCMC. As parameters $\Theta = \{\theta_0, \theta_1, \gamma_0, \gamma_1\}$ are all positive, I use independent gamma distributions as the prior. The gamma density is

$$g(x; \alpha, \beta) = \beta^{\alpha} e^{-\beta x} x^{\alpha-1}/\Gamma(\alpha), \quad (13)$$

with mean α/β and variance α/β^2. The hyperparameters α and β are chosen to reflect the range and likely values of the parameters.

Instead of the coalescent times t_0 and t_1, which have different definitions in different gene trees (Figure 1), branch lengths b_0 and b_1 in the gene tree are used in the MCMC. The joint prior distributions of the gene tree T_0 and its branch lengths b_0 and b_1 given Θ can be easily derived from the distributions of the coalescent times t_0 and t_1 (Equation 2),

$$\begin{aligned} f(T_0, b_0, b_1|\Theta) &= f(T_0|\Theta) f(b_1|T_0, \Theta) f(b_0|T_0, b_1, \Theta) \\ &= \psi \times \frac{2}{\psi\theta_1} e^{-2(b_1-\gamma_1)/\theta_1} \times \frac{2}{\theta_0} e^{-2(b_0+b_1-\gamma_0-\gamma_1)/\theta_0} \\ &= 4/(\theta_0\theta_1) \times e^{-2(b_1-\gamma_1)/\theta_1 - 2(b_0+b_1-\gamma_0-\gamma_1)/\theta_0}, \end{aligned} \quad (14)$$

for $\gamma_1 < b_1 < \gamma_1 + \gamma_0$ and $\gamma_1 + \gamma_0 - b_1 < b_0 < \infty$.

Similarly, the joint prior distribution of gene tree T_k, $k = 1, 2, 3$, and its branch lengths b_0 and b_1 given Θ is

$$\begin{aligned} f(T_k, b_0, b_1|\Theta) &= f(T_k|\Theta) f(b_0|T_k, \Theta) f(b_1|T_k, \Theta) \\ &= \frac{1-\psi}{3} \times \frac{2}{\theta_0} e^{-2b_0/\theta_0} \times \frac{6}{\theta_0} e^{-6(b_1-\gamma_0-\gamma_1)/\theta_0} \\ &= \frac{4}{\theta_0^2} e^{-2\gamma_0/\theta_1 - [2b_0+6(b_1-\gamma_0-\gamma_1)]/\theta_0}, \end{aligned} \quad (15)$$

with $0 < b_0 < \infty$ and $\gamma_1 + \gamma_0 < b_1 < \infty$.

The variables to be updated in the Markov chain include the parameters $\Theta = \{\theta_0, \theta_1, \gamma_0, \gamma_1\}$ and the gene

TABLE 3

Prior and posterior distributions of parameters in the Bayes analysis

Parameter	Prior		Posterior	
	$(\alpha, \beta)^a$	Mean (95% interval)[b]	Mean	Mean (95% interval)[b]
		Good priors		
θ_0 (N_0)	(2, 2,000)	12,500 (1,500, 34,800)	0.00158	19,700 (2,900, 41,600)
θ_1 (N_1)	As above		0.00100	12,400 (1,700, 32,100)
γ_0 (time in MY)	(4, 2,500)	1.6 MY (0.44 MY, 3.51 MY)	0.00164	1.6 MY (0.7 MY, 2.7 MY)
γ_1 (time in MY)	(20, 4,000)	5.0 MY (3.1 MY, 7.4 MY)	0.00530	5.3 MY (4.4 MY, 6.1 MY)
		Poor priors		
θ_0 (N_0)	(1.5, 300)	62,500 (4,500, 194,800)	0.00263	32,900 (5,300, 57,800)
θ_1 (N_1)	As above		0.00265	33,100 (5,300, 88,600)
γ_0 (time in MY)	(4, 2,000)	2.0 MY (0.55 MY, 4.4 MY)	0.00190	1.9 MY (0.8 MY, 3.2 MY)
γ_1 (time in MY)	(4, 800)	5.0 MY (1.3 MY, 11.0 MY)	0.00464	4.6 MY (3.4 MY, 5.7 MY)

[a] Parameters α and β are for the gamma priors; the prior mean is α/β (not shown).

[b] Mean and 2.5 and 97.5% percentiles of the prior or posterior distributions for population sizes or speciation times. In converting θ and γ into N and speciation time, the generation time is assumed to be $g = 20$ years and the mutation rate 10^{-9} substitutions per site per year.

trees and branch lengths at all L loci, $G = \{T_i, b_{i0}, b_{i1}\}$, $i = 1, 2, \ldots, L$. The Markov chain is constructed so that its steady-state distribution is the posterior distribution of those variables. Bayes inference is then based on the joint posterior distribution

$$f(\Theta, G|D) = \frac{f(D|G)f(G|\Theta)f(\Theta)}{f(D)} = \frac{\prod_{i=1}^{L} P(D_i|T_i, b_{i0}, b_{i1}) \times \prod_{i=1}^{L} f(T_i, b_{i0}, b_{i1}|\Theta) \times f(\theta_0)f(\theta_1)f(\gamma_0)f(\gamma_1)}{f(D)} \quad (16)$$

The denominator $f(D)$ is the marginal probability of the data

$$f(D) = \int_{\Theta}\int_{G} f(D|G)f(G|\Theta)f(\Theta)\,dG\,d\Theta, \quad (17)$$

where the integral over G represents summation over the four gene trees (T_0, T_1, T_2, T_3 in Figure 1) and integration over branch lengths in each tree. The posterior distribution of any parameter is then given by integrating over the joint posterior distribution. For example,

$$f(\Theta|D) = \int_{G} f(\Theta, G|D)\,dG. \quad (18)$$

A Metropolis-Hastings algorithm (Metropolis *et al.* 1953; Hasting 1970) is used to update variables in the MCMC. Given the current state of the chain (Θ, G), a new state (Θ^*, G^*) is proposed through a proposal distribution $q(\Theta^*, G^*|\Theta, G)$. The new state is then accepted with probability

$$R = \min\left\{1, \frac{f(\Theta^*, G^*|D)}{f(\Theta, G|D)} \times \frac{q(\Theta, G|\Theta^*, G^*)}{q(\Theta^*, G^*|\Theta, G)}\right\} = \min\left\{1, \frac{f(D|G^*)f(G^*|\Theta^*)f(\Theta^*)}{f(D|G)f(G|\Theta)f(\Theta)} \times \frac{q(\Theta, G|\Theta^*, G^*)}{q(\Theta^*, G^*|\Theta, G)}\right\}. \quad (19)$$

If the new state is accepted, the chain moves to the new state (Θ^*, G^*). Otherwise the chain stays in the old state (Θ, G). Note that $f(D)$ in Equation 16 cancels in calculation of the acceptance ratio R. Calculation of $f(\Theta^*, G^*|D)/f(\Theta, G|D)$ is straightforward due to the conditional independence in the model as described above. So the focus here is the proposal mechanism and the proposal ratio $q(\Theta, G|\Theta^*, G^*)/q(\Theta^*, G^*|\Theta, G)$.

The proposal density q can be rather flexible as long as it specifies an aperiodic and irreducible Markov chain. The algorithm I implemented cycles through several steps, with each step updating some but not all variables. In step 1, the gene tree and branch lengths at each locus i (T_i, b_{i0}, b_{i1}) are updated, while parameters Θ are fixed. Each locus is updated once in this step. Step 2 updates parameters Θ while the branch lengths $\{b_{i0}, b_{i1}\}$ are fixed. This step can cause changes to the gene trees at some loci. Step 3 is a mixing step, in which parameters θ_0, θ_1, γ_0, γ_1 and branch lengths at all loci are multiplied by a constant while the gene trees remain unchanged. The MCMC algorithm is tedious and the details are given in the appendix.

The Markov chain is started from a random initial state. Sampling starts after a certain number of generations, which are discarded as burn-in, and samples are taken every certain number of steps, thus "thinning" the chain. Convergence of the chain is checked by examining the output and also by running multiple chains. The algorithm is also run without sequence data, and the posterior distribution generated is found to be close to the prior.

Application to the data of Chen and Li: The Bayes MCMC algorithm is applied to the data of Chen and Li (2001; see Table 1). I used two sets of priors (Table

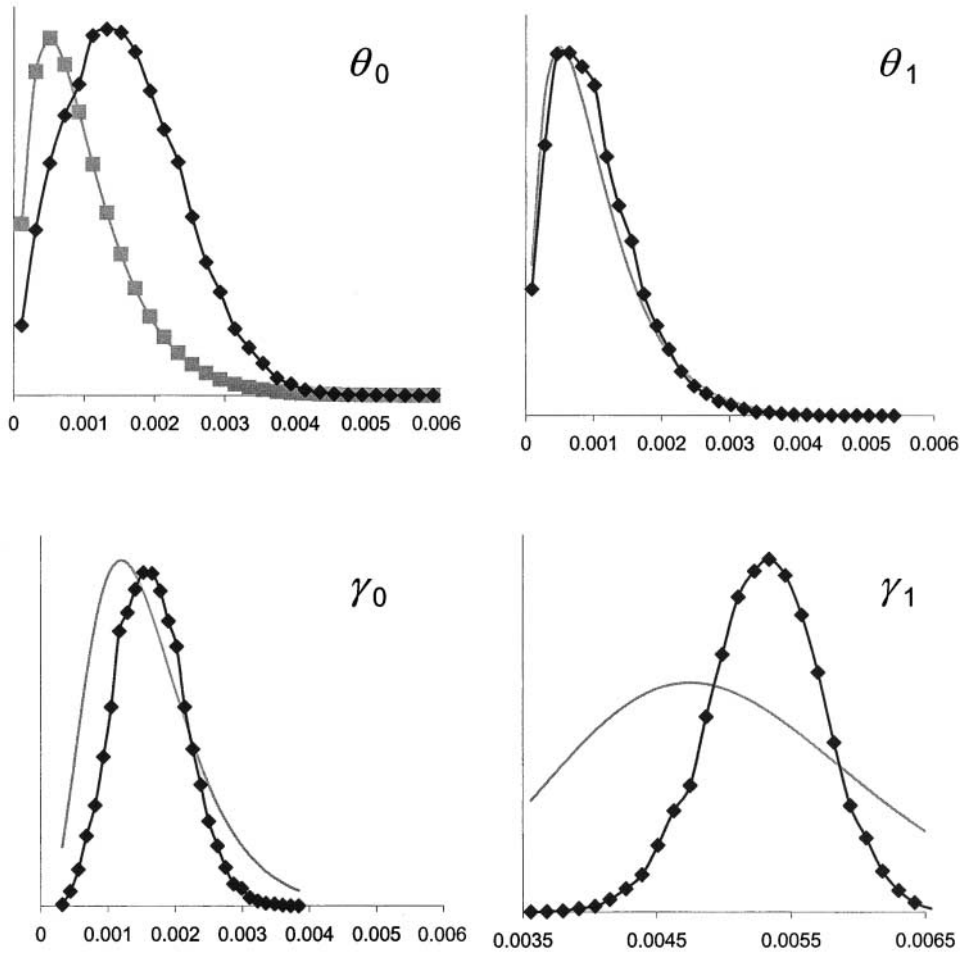

FIGURE 4.—Prior and posterior distributions for parameters θ_0, θ_1, γ_0, and γ_1. Parameter estimates are shown in Table 3 (good priors).

3). Parameters α and β in the gamma prior distributions are chosen by considering likely values and ranges of ancestral population sizes and species divergence times and converting them into parameters θ_0, θ_1, γ_0, and γ_1 using a generation time of 20 years and a mutation rate of 10^{-9} substitutions per site per year. The first set is considered more realistic and referred to as the "good priors" in Table 3. Ancestral population sizes N_0 and N_1 are centered ~12,500, close to estimates for modern humans, with the 2.5 and 97.5% percentiles at 1500 and 34,800, respectively. The divergence time for humans and chimpanzees is centered ~5 MY, while the divergence time for the gorilla is centered ~1.6 MY. Note that parameters θ_0, θ_1, γ_0, and γ_1 are all $\ll 1$ but are definitely >0; thus values of $\alpha > 1$ are used so that the gamma distribution has a mode >0.

The posterior distributions of parameters θ_0, θ_1, γ_0, and γ_1 are shown in Figure 4 together with their priors. They are also summarized in Table 3 (good priors). The means of the posterior distributions for θ_0, θ_1, γ_0, and γ_1 are listed, and then the means and the 95% credibility sets for the two population sizes (N_0 and N_1) and for the two speciation times are listed. The posterior means and medians are close. The population size for the ancestor of humans and chimpanzees is estimated to be 12,400, with the 95% credibility interval (CI) to be (1700, 32,100). The H-C divergence is dated at 5.3 MY, with the 95% CI to be (4.4, 6.1). The estimates of θ_1 and γ_1 are very similar to the MLEs. The posterior mean of θ_0 is smaller and that of γ_0 is larger than the MLEs (Tables 2 and 3). The correlation coefficients calculated from the posterior distributions of parameters θ_0, θ_1, γ_0, and γ_1 are shown in Table 4. There is strong negative correlation between θ_0 and γ_0 and between θ_1 and γ_1. Comparison of the prior and posterior distributions (Figure 4) suggests that the data contain much more information about the divergence times, especially the H-C divergence time (γ_1), than about the population sizes.

To see the effect of prior assumptions on the posterior distributions, I used a second set of priors, which are more spread out and also assume large population sizes (mean 62,500). The posterior distributions are summarized in Table 3 under the heading "Poor priors." The point estimates of both N_0 and N_1 are ~33,000, smaller than the prior means. The H-C divergence is dated at 4.6 MY, and the gorilla divergence is dated 1.9 MY earlier. Those estimates appear reasonable, although the

TABLE 4

Correlation coefficients in the posterior distribution

	θ_0	θ_1	γ_0
θ_1	0.05		
γ_0	−0.58	0.43	
γ_1	−0.16	−0.60	−0.41

H-C divergence date is too recent. Similar strong correlations between the parameters are discovered as in the analysis using the good priors. The negative correlation between γ_1 and θ_1 (calculated to be −0.76), combined with the assumed and estimated large population sizes, appears to have led to a H-C divergence date that is too recent.

DISCUSSION

ML and Bayes methods of this article estimated the population size for the common ancestor of humans and chimpanzees to be ~12,000, similar to estimates for modern humans. The estimates are several times smaller than those obtained by Chen and Li (2001) from the same data using the tree-mismatch method, which range from 52,000 to 150,000. Even the worst-case estimates—*e.g.*, 36,000 by ML under the assumption that all sequence divergence variation among loci is due to mutation rate variation and 33,000 from the Bayes analysis using the poor priors—are smaller than the minimum estimate of Chen and Li. The tree-mismatch method used by Chen and Li appears to have serious biases due to errors in gene tree reconstruction, and the likelihood and Bayes estimates reported here are probably more reliable. Thus it may be concluded that the sequence data of Chen and Li (2001) do not support much larger ancestral populations than the modern humans or the notion that early human populations experienced dramatic size reductions (Hacia 2001; Kaessmann *et al.* 2001).

While the ML and Bayes methods are expected to have better statistical properties than the simple tree-mismatch method, it is worthwhile to examine some of the assumptions made in this article. First, the evolutionary rate is assumed to be constant over lineages. This assumption seems reasonable as the species compared are very closely related; Chen and Li's (2001) relative-rate tests supported the molecular clock. The large differences between the tree-mismatch method and the likelihood and Bayes methods are clearly not due to the use of the clock assumption in this article; use of clock rooting in the tree-mismatch method produced even larger estimates of the population size for the ancestor of humans and chimpanzees. Second, the analysis assumes no recombination within a locus. The effect of recombination on estimation of parameters θ_0, θ_1, γ_0, and γ_1 is not well understood, although Satta *et al.* (2000) emphasized its possible significance. As the human, chimpanzee, and gorilla sequences are extremely similar, most of the recombination events will not be visible in the sequence data, and the few sites at which more than two nucleotides are observed in the data (see counts n_{i4} for site pattern *xyz* in Table 1) are probably due to multiple substitutions at the same site. Third, the substitution model of Jukes and Cantor (1969) is simplistic. More complex models, such as those that account for variable substitution rates among sites within the locus, can be easily implemented, but are expected to have little effect. The most serious issue seems to be mutation rate variation among loci. In the case of two species, the ancestral population size is overestimated when mutation rate variation is ignored and accounting for the bias leads to dramatic reduction in the estimated population size (Yang 1997a). In this article, the population size of the ancestor of humans and chimpanzees is not very large under the constant-rate model and becomes larger when variable rates for loci are assumed. The effect is much less important and also in the opposite direction compared with the two-species case. Lack of strong correlation among sequence distances with the orangutan seems to suggest that the rates are relatively homogeneous among those loci. It seems that simultaneous analysis of data from three species allows the parameters to constrain each other, leading to a better use of information in the data. It is quite likely that the estimation can be further improved by sampling multiple individuals from the same species.

The ML and Bayes methods produced similar results for the data analyzed in this article. However, the ML calculation is slower than the MCMC algorithm. The Bayes approach also provides a framework for incorporating prior information about the parameters. For example, a wealth of information is available about the divergence time between humans and chimpanzees. By forcing a very narrow prior distribution for γ_1, such information can be incorporated in the Bayes analysis. Using an informative prior will reduce the adverse effect of strong correlation among parameters when other parameters are estimated. Furthermore, the Bayes algorithm seems easier than ML to extend to data that contain more than three species and more than one individual from each species.

Program availability: C programs implementing the MCMC algorithm and calculating the mismatch probabilities (P_{SG}, P_{SE}, and P_{GE}, etc.) are available from the author upon request. The C and Mathematica programs for the likelihood method are available as well, but they make use of the Mathlink library and are awkward to use.

I am very grateful to Drs. W.-H. Li and F.-C. Chen for providing the data analyzed in this article. I thank M. Hasegawa and B. Larget for discussions and N. Takahata for comments. This study is supported by grant 31/G13580 from the Biotechnology and Biological Sciences Research Council (United Kingdom).

LITERATURE CITED

Chen, F.-C., and W.-H. Li, 2001 Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. **68:** 444–456.

Edwards, S. V., and P. Beerli, 2000 Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution **54:** 1839–1854.

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

Fu, Y.-X., 1994 A phylogenetic estimator of effective population size or mutation rate. Genetics **136:** 685–692.

Hacia, J. G., 2001 Genome of the apes. Trends Genet. **17:** 637–645.

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

Hasegawa, M., H. Kishino and T. Yano, 1987 Man's place in Hominoidea as inferred from molecular clocks of DNA. J. Mol. Evol. **26:** 132–147.

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

Hudson, R. R., 1992 Gene trees, species trees and the segregation of ancestral alleles. Genetics **131:** 509–513.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–123 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Kaessmann, H., V. Wiebe, G. Weiss and S. Paabo, 2001 Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat. Genet. **27:** 155–156.

Kreitman, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature **304:** 412–417.

Kumar, S., and S. B. Hedges, 1998 A molecular timescale for vertebrate evolution. Nature **392:** 917–920.

Lio, P., and N. Goldman, 1998 Models of molecular evolution and phylogeny. Genome Res. **8:** 1233–1244.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1092.

Nachman, M. W., and S. L. Crowell, 2000 Estimate of the mutation rate per nucleotide in humans. Genetics **156:** 297–304.

Nei, M., 1987 *Molecular Evolutionary Genetics.* Columbia University Press, New York.

Ruvolo, M., 1997 Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Mol. Biol. Evol. **14:** 248–265.

Saitou, N., 1988 Property and efficiency of the maximum likelihood method for molecular phylogeny. J. Mol. Evol. **27:** 261–273.

Satta, Y., J. Klein and N. Takahata, 2000 DNA archives and our nearest relative: the trichotomy problem revisited. Mol. Phylogenet. Evol. **14:** 259–275.

Takahata, N., 1986 An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. Genet. Res. **48:** 187–190.

Takahata, N., Y. Satta and J. Klein, 1995 Divergence time and population size in the lineage leading to modern humans. Theor. Popul. Biol. **48:** 198–221.

Wu, C.-I, 1991 Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics **127:** 429–435.

Yang, Z., 1994 Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst. Biol. **43:** 329–342.

Yang, Z., 1997a On the estimation of ancestral population sizes. Genet. Res. **69:** 111–116.

Yang, Z., 1997b PAML: a program package for phylogenetic analysis by maximum likelihood (http://abacus.gene.ucl.ac.uk/software/paml.html). Comput. Appl. Biosci. **13:** 555–556.

Yoder, A. D., and Z. Yang, 2000 Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. **17:** 1081–1090.

Yu, N., Z. Zhao, Y. X. Fu, N. Sambuughin, M. Ramsay *et al.*, 2001 Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. **18:** 214–222.

Zhao, Z., L. Jin, Y. X. Fu, M. Ramsay, T. Jenkins *et al.*, 2000 Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. USA **97:** 11354–11358.

Communicating editor: Y.-X. Fu

APPENDIX: IMPLEMENTATION OF THE MCMC ALGORITHM

Updating the gene tree and branch lengths at each locus: This step of the algorithm goes through the L loci to update the branch lengths and possibly the gene tree as well. The algorithm visits each locus once. The description in this section concerns one locus i. The subscript i in b_{i0} and b_{i1} may be suppressed. The proposal algorithm for gene trees T_2 and T_3 is simpler and is described first, before the algorithm for gene trees T_0 and T_1 (Figure 1).

If the current gene tree is either T_2 or T_3, only the branch lengths are updated and the gene tree is not changed. Note that b_0 and b_1 have to satisfy the requirements $0 < b_0 < \infty$ and $\gamma_0 + \gamma_1 < b_1 < \infty$. A sliding window is used to propose new branch lengths,

$$b_0^* \sim U(b_0 - w/2,\ b_0 + w/2),$$
$$b_1^* \sim U(b_1 - w/2,\ b_1 + w/2), \tag{A1}$$

where $U(a, b)$ is the uniform distribution in the interval (a, b). The window size is set at $w = H_1$, with H_1 a small fine-tuning parameter. If the proposed value is outside the range of the variables, the excess is reflected back into the range; that is, if $b_0^* < 0$, then b_0^* is reset to $-b_0^*$, and if $b_1^* < \gamma_0 + \gamma_1$, then b_1^* is reset to $2(\gamma_0 + \gamma_1) - b_1^*$. The proposal ratio is 1. The acceptance ratio is

$$R = \min\left\{1, \frac{P(D_i|T_i,\ b_{i0}^*,\ b_{i1}^*)}{P(D_i|T_i,\ b_{i0},\ b_{i1})} \times \frac{f(T_i,\ b_{i0}^*,\ b_{i1}^*|\theta_0,\ \theta_1,\ \gamma_0,\ \gamma_1)}{f(T_i,\ b_{i0},\ b_{i1}|\theta_0,\ \theta_1,\ \gamma_0,\ \gamma_1)}\right\}. \tag{A2}$$

If the current gene tree is either T_0 or T_1 (Figure 1, b and c), a change between those two trees is allowed when the branch lengths are updated. Both the current and the new branch lengths should satisfy the constraints $\gamma_1 < b_1 < \infty$ and $\gamma_0 + \gamma_1 < b_0 + b_1 < \infty$. To propose new branch lengths under those constraints, I apply the transformation

$$y_0 = b_0 + b_1 - (\gamma_0 + \gamma_1),$$
$$y_1 = (b_1 - \gamma_1)/(b_0 + b_1 - \gamma_1), \tag{A3}$$

with $y_0 > 0$ and $0 < y_1 < 1$. Note that y_0 is the distance between the root of the gene tree (node A in Figure 1) and the first speciation event (C in Figure 1), and changing y_0 will shrink or expand the height of the gene tree, while y_1 is the ratio of distance BD to AD, and changing y_1 will slide node B between A and D (Figure 1, b and c).

New states are proposed for y_0 and y_1 as

$$y_0^* = y_0 e^{H_1(r-0.5)},$$
$$y_1^* \sim U(y_1 - H_1/2,\ y_1 + H_1/2), \tag{A4}$$

where r is a random variable from $U(0, 1)$. If y_1^* is out of the range (0, 1), it is reflected back into the range. The new branch lengths b_0^* and b_1^* are calculated from the relationships

$$b_0 = (\gamma_0 + y_0)(1 - y_1),$$
$$b_1 = \gamma_1 + y_1(\gamma_0 + y_0). \quad \text{(A5)}$$

If $b_1^* > \gamma_0 + \gamma_1$, the new gene tree T_i^* is set to T_1; otherwise it is set to T_0. This change of gene tree does not change the proposal ratio. The proposal ratio for variables y_0 and y_1 is y_0^*/y_0. The proposal ratio for the original variables b_0 and b_1 can be derived by noting

$$\left|J\right| = \begin{vmatrix} \frac{\partial b_0}{\partial y_0} & \frac{\partial b_0}{\partial y_1} \\ \frac{\partial b_1}{\partial y_0} & \frac{\partial b_1}{\partial y_1} \end{vmatrix} = \begin{vmatrix} 1 - y_1 & -(\gamma_0 + y_0) \\ y_1 & \gamma_0 + y_0 \end{vmatrix} = \gamma_0 + y_0.$$

Thus the acceptance ratio is

$$R = \min\left\{1, \frac{P(D_i|T_i^*, b_{i0}^*, b_{i1}^*)}{P(D_i|T_i, b_{i0}, b_{i1})} \times \frac{f(T_i^*, b_{i0}^*, b_{i1}^*|\theta_0, \theta_1, \gamma_0, \gamma_1)}{f(T_i, b_{i0}, b_{i1}|\theta_0, \theta_1, \gamma_0, \gamma_1)} \times \frac{(\gamma_0 + y_0^*)y_0^*}{(\gamma_0 + y_0)y_0}\right\}. \quad \text{(A6)}$$

If the gene tree is T_1, T_2, or T_3, the algorithm may (with a small probability of, say, 0.2 or 0.3) attempt to swap the gene trees. The current gene tree is replaced by one of the other two trees chosen at random. The proposal ratio is 1, and the acceptance ratio is

$$R = \min\left\{1, \frac{P(D_i|T_i^*, b_{i0}, b_{i1})}{P(D_i|T_i, b_{i0}, b_{i1})} \times \frac{f(T_i^*, b_{i0}, b_{i1}|\theta_0, \theta_1, \gamma_0, \gamma_1)}{f(T_i, b_{i0}, b_{i1}|\theta_0, \theta_1, \gamma_0, \gamma_1)}\right\}. \quad \text{(A7)}$$

Updating population size and speciation date parameters: This step makes two proposals: the first to change θ_0 and θ_1 and the second to change γ_0 and γ_1. Parameters θ_0 and θ_1 are positive but are not constrained otherwise. They are updated simultaneously, with all other variables fixed. New values are proposed around the current values by

$$\theta_0^* = \theta_0 e^{H_2(r_0 - 0.5)},$$
$$\theta_1^* = \theta_1 e^{H_2(r_1 - 0.5)}, \quad \text{(A8)}$$

where r_0 and r_1 are uniform random numbers in the interval (0, 1) and H_2 is a small fine-tuning parameter. The proposal ratio is $\theta_0^*\theta_1^*/(\theta_0\theta_1)$ and the acceptance ratio is

$$R = \min\left\{1, \prod_{i=1}^{L} \frac{f(T_i, b_{i0}, b_{i1}|\theta_0^*, \theta_1^*, \gamma_0, \gamma_1)}{f(T_i, b_{i0}, b_{i1}|\theta_0, \theta_1, \gamma_0, \gamma_1)} \times \left(\frac{\theta_0^*\theta_1^*}{\theta_0\theta_1}\right)^{\alpha-1} e^{-\beta(\theta_0^* - \theta_0 + \theta_1^* - \theta_1)} \times \frac{\theta_0^*\theta_1^*}{\theta_0\theta_1}\right\}. \quad \text{(A9)}$$

Next, speciation date parameters γ_0 and γ_1 are updated while θ_0 and θ_1 as well as branch lengths b_{i0} and b_{i1} for all loci are fixed. At each locus the following constraints have to be satisfied: $0 < \gamma_1 < b_1 < \gamma_0 + \gamma_1 < b_0 + b_1 < \infty$ in gene tree T_0 and $0 < b_0 < \infty$ and $\gamma_0 + \gamma_1 < b_1 < \infty$ in gene tree T_1, T_2, or T_3 (Figure 1). To allow the chain to move more freely, the gene tree is allowed to change, if necessary, from T_0 to T_1, T_2, or T_3 or vice versa. Thus only the following constraints have to be satisfied when new values are proposed for γ_0 and γ_1: $0 < \gamma_1 < b_1$ and $\gamma_0 + \gamma_1 < b_0 + b_1$ at every locus; that is, $\gamma_1 < \min\{b_{i1}\} = c$ and $\gamma_1 < \gamma_0 + \gamma_1 < \min\{b_{i0} + b_{i1}\} = d$. The following transformation is used to propose new states,

$$y_0 = \gamma_0/(d - \gamma_1),$$
$$y_1 = \gamma_1, \quad \text{(A10)}$$

with constraints $0 < y_0 < 1$ and $0 < y_1 < c$. Note that y_0 is the ratio of the distance CD to AD in Figure 1, and changing y_0 will slide the speciation date C (Figure 1) between A and D. Note that $\gamma_0 = y_0(d - \gamma_1)$, $\gamma_1 = y_1$, and

$$\left|J\right| = \begin{vmatrix} \frac{\partial \gamma_0}{\partial y_0} & \frac{\partial \gamma_0}{\partial y_1} \\ \frac{\partial \gamma_1}{\partial y_0} & \frac{\partial \gamma_1}{\partial y_1} \end{vmatrix} = d - \gamma_1.$$

Sliding windows are used to propose new values,

$$y_0^* \sim U(y_0 - H_3/2, y_0 + H_3/2),$$
$$y_1^* \sim U(y_1 - H_3/2, y_1 + H_3/2), \quad \text{(A11)}$$

where H_3 is a small fine-tuning parameter. If the new values y_0^* and y_1^* are out of the range, they are reflected back into the range. The proposal ratio for the transformed variables y_0 and y_1 is 1. The proposal ratio for variables γ_0 and γ_1 is $(d - \gamma_1^*)/(d - \gamma_1)$. Next, all loci are scanned to see whether the gene tree needs updating. If the current gene tree is T_0 but $\gamma_0^* + \gamma_1^* < b_{i1}$, the gene tree T_i^* is set to be one of T_1, T_2, or T_3, chosen at random. The proposal ratio will be multiplied by 3 since there are three trees to move to and only one tree to move back. If the current tree is T_1, T_2, or T_3 but $\gamma_0^* + \gamma_1^* > b_{i1}$, the gene tree is set to $T_i^* = T_0$, and the proposal ratio is divided by 3. Otherwise the gene tree for the locus remains unchanged: $T_i^* = T_i$. In sum the acceptance ratio is

$$R = \min\left\{1, \prod_{i=1}^{L} \frac{P(D_i|T_i^*, b_{i0}, b_{i1})f(T_i^*, b_{i0}, b_{i1}|\theta_0, \theta_1, \gamma_0^*, \gamma_1^*)}{P(D_i|T_i, b_{i0}, b_{i1})f(T_i, b_{i0}, b_{i1}|\theta_0, \theta_1, \gamma_0, \gamma_1)} \times \frac{f(\gamma_0^*, \gamma_1^*)}{f(\gamma_0, \gamma_1)} \times \frac{d - \gamma_1^*}{d - \gamma_1} c_T\right\}, \quad \text{(A12)}$$

where $f(\gamma_0^*, \gamma_1^*)/f(\gamma_0, \gamma_1) = (\gamma_0^*\gamma_1^*/\gamma_0\gamma_1)^{\alpha-1} e^{-\beta(\gamma_0^* - \gamma_0 + \gamma_1^* - \gamma_1)}$ from the gamma priors and c_T is the proposal ratio due to changes to gene trees at some loci (a product of threes and one-thirds).

Mixing step: A mixing step is found to be effective

in speeding up convergence when a poor starting point is chosen for the chain. In this step, the gene trees remain unchanged, but parameters θ_0, θ_1, γ_0, and γ_1 and branch lengths b_{i0} and b_{i1} for all loci are multiplied by a constant

$$c = e^{H_4(r-0.5)}, \tag{A13}$$

where r is a random number from $U(0, 1)$ and H_4 is a small fine-tuning parameter. The proposal ratio is c^{4+2L}. The acceptance ratio is

$$R = \min\left\{1, \prod_{i=1}^{L}\left[\frac{P(D_i|T_i, b_{i0}^*, b_{i1}^*) \times f(T_i, b_{i0}^*, b_{i1}^*|\theta_0^*, \theta_1^*, \gamma_0^*, \gamma_1^*)}{P(D_i|T_i, b_{i0}, b_{i1}) \times f(T_i, b_{i0}, b_{i1}|\theta_0, \theta_1, \gamma_0, \gamma_1)}\right] \times \frac{f(\Theta^*)}{f(\Theta)} \times c^{4+2L}\right\},$$

where $f(\Theta^*)/f(\Theta) = c^{4(\alpha-1)}e^{-\beta(\theta_0^*-\theta_0+\theta_1^*-\theta_1+\gamma_0^*-\gamma_0+\gamma_1^*-\gamma_1)}$, given by the gamma prior distributions.

Performance of the algorithm: The performance of the MCMC algorithm is noted to depend on the choice of priors (values of α and β in the gamma distributions). For some priors, the algorithm is noted not to mix well, and in particular, parameters θ_0 and γ_0 appear to change slowly. The high correlation among parameters and the constraints seem to cause difficulties for the algorithm. In such cases, the chain has to be run much longer than usual to achieve stable estimates. In proposing new values for γ_0 and γ_1, only small steps are taken (with H_3 in the range 0.01–0.05) to achieve an acceptance rate of ~0.1–0.3. For other variables, even large steps (with H_1, H_2, H_4 in the range 0.1–0.5) are accepted at high frequencies (>50%). The mixing step seems rather effective so that ~1000 generations seem enough for the burn-in. For some priors, <500,000 generations appear sufficient, which takes a few minutes on a Pentium III PC at 1.2 GHz for the data of Chen and Li (2001). For other priors, the algorithm has to be run much longer.