J Mol Evol (1996) 43:304–311

JOURNAL OF MOLECULAR EVOLUTION

Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference

Bruce Rannala, Ziheng Yang

Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA

Received: 12 November 1995 / Accepted: 16 April 1996

Abstract. A new method is presented for inferring evolutionary trees using nucleotide sequence data. The birth–death process is used as a model of speciation and extinction to specify the prior distribution of phylogenies and branching times. Nucleotide substitution is modeled by a continuous-time Markov process. Parameters of the branching model and the substitution model are estimated by maximum likelihood. The posterior probabilities of different phylogenies are calculated and the phylogeny with the highest posterior probability is chosen as the best estimate of the evolutionary relationship among species. We refer to this as the maximum posterior probability (MAP) tree. The posterior probability provides a natural measure of the reliability of the estimated phylogeny. Two example data sets are analyzed to infer the phylogenetic relationship of human, chimpanzee, gorilla, and orangutan. The best trees estimated by the new method are the same as those from the maximum likelihood analysis of separate topologies, but the posterior probabilities are quite different from the bootstrap proportions. The results of the method are found to be insensitive to changes in the rate parameter of the branching process.

Key words: Maximum likelihood — Phylogeny — Nucleotide substitution — Posterior probability — Empirical Bayes estimation — MAP tree

Correspondence to: Z. Yang

Introduction

Felsenstein (1973, 1981) proposed a maximum likelihood (ML) method for inferring evolutionary trees using discrete characters (such as nucleotide sequences) of extant species. A Markov model is used to describe the evolutionary changes between character states and the tree topology and branch lengths are treated as parameters (see also Thompson 1975; Bishop and Friday 1985; Goldman 1990). Branch lengths and parameters in the substitution model are estimated by maximum likelihood for each tree topology, generating the (maximum) likelihood value for that topology. The tree with the highest (maximum) likelihood is chosen as the estimate of phylogeny. Computer simulations have demonstrated a general superiority of the ML method over other techniques of tree reconstruction under a variety of conditions (Hasegawa and Yano 1984; Fukami-Kobayashi and Tateno 1991; Hasegawa et al. 1991; Tateno et al. 1994; Kuhner and Felsenstein 1994; Yang 1994a, 1995; Gaut and Lewis 1995; Huelsenbeck 1995a,b). With improved computational power and increased acceptance of statistical methods by molecular systematists, the method is becoming more widely used.

Felsenstein's method differs from conventional maximum likelihood parameter estimation in that the functional form of the likelihood depends on the tree topology (Nei 1987:323–325), and the regularity conditions required for the asymptotic properties of maximum likelihood estimators are not satisfied (Yang 1994a, 1996). As a result, it is unclear whether this method of topology estimation shares all the asymptotic properties (especially efficiency) of maximum likelihood estimators of

parameters (Yang 1996). Another difficulty is the lack of a reliable method for evaluating the significance of the estimated tree. The method of nonparametric bootstrapping (Felsenstein 1985), intended to provide a measure of the sampling error of the estimated phylogeny, has been found to give somewhat unreliable results (Zharkikh and Li 1992; Hillis and Bull 1993), and its correct interpretation has been a subject of controversy (e.g., Felsenstein and Kishino 1993).

In this paper, we approach the problem of phylogenetic tree estimation from a somewhat different perspective. We use a birth–death process (Feller 1939) to specify the prior distribution of tree topologies and divergence times (or branch lengths) of the extant species and a Markov process to model nucleotide substitution. Parameters of the birth–death process and the substitution model are estimated by maximizing the likelihood (the probability of observing the data). The (posterior) probability of each tree topology, conditional on the nucleotide sequence data and the estimated parameters, is then calculated. The tree having the highest posterior probability is taken as the estimate of phylogeny. This is referred to as the maximum posterior probability (MAP) tree. The MAP method differs from the ML method of Felsenstein (1981) in that we treat topologies and branch lengths as random variables rather than parameters.

The structure of the model is very similar to a model studied by Cavalli-Sforza and Edwards (1967; see also Edwards 1970). These authors analyzed gene frequency data from human populations, using a Brownian motion model to approximate the process of genetic drift in different populations and a Yule pure-birth process (Yule 1925) to model population branching events. Their model was found to be mathematically intractable and remains unanalyzed. Felsenstein (1973, 1981), when adapting the method of Cavalli-Sforza and Edwards for use with discrete character data, used a Markov process to model character evolution and omitted the Yule process in order to simplify the mathematics. We note, however, that the nature of nucleotide sequence data and the independence of substitutions among sites assumed in the Markov model of character evolution makes the computation using either the Yule process or a birth–death process feasible, at least for small numbers of taxa; in this paper, we modify the model of Cavalli-Sforza and Edwards to derive a new method for phylogenetic analysis using nucleotide sequence data.

Models and Estimation Theory

The Data

Let s be the number of sequences (species) and n the number of nucleotides in each sequence; insertions and deletions are ignored and it is assumed that the sequences are aligned with gaps removed. The data can be represented as an $s \times n$ matrix, $\mathbf{X} = \{x_{ij}\}$, where x_{ij} is the nucleotide at the jth site in the ith sequence. The jth column of the data matrix, $\mathbf{x}_j = \{x_{1j}, \ldots, x_{sj}\}'$, will be the nucleotides among species at the jth site. The sequences are descended through $s - 1$ speciation events, which occurred at times $t_1, t_2, \ldots, t_{s-1}$ in the past, with $t_1 > t_2 > \ldots > t_{s-1}$. The time of the first bifurcation is set to 1 (i.e., $t_1 = 1$) and parameters are then relative to this time scale; we let $\mathbf{t} = \{t_2, \ldots, t_{s-1}\}$. An example tree of four sequences ($s = 4$) is shown in Fig. 1A.

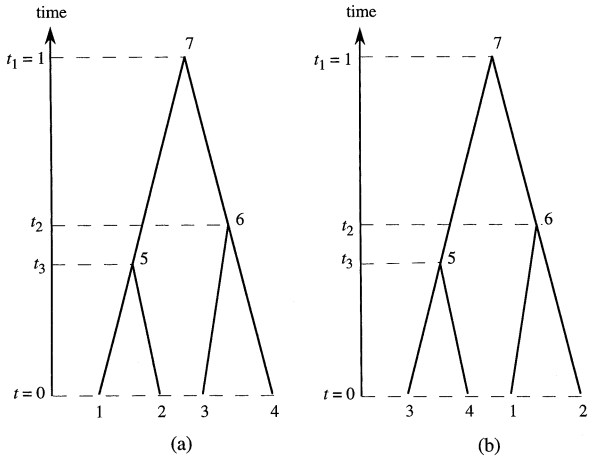

Fig. 1. Two ''labeled histories'' of four sequences that have the same tree topology. In **A,** the divergence event separating sequences 1 and 2 occurred after the divergence of sequences 3 from 4, while the opposite is true in **B.** In Table 1, these two labeled histories are represented as ((12)(34)) and ((34)(12)), respectively. The divergence times are t_1, t_2, and t_3, with $t_1 = 1 > t_2 > t_3$.

The Birth–Death Process

The birth–death process is a continuous-time process in which the probability that a speciation event occurs along any lineage during an infinitesimal time interval Δt is $\lambda\Delta t$, the probability that an extinction occurs is $\mu\Delta t$, and the probability that two or more events occur is of order $o(\Delta t)$. Parameters λ and μ are the branching and extinction rates per lineage, respectively. The number of species at present (s), the phylogenetic tree relating the species (τ), and the times of divergence **(t)** are all random variables under the model, with their distribution determined by parameters λ and μ. The phylogenetic tree (τ) is a ''labeled history,'' which is a topology with interior nodes ordered according to their times of divergence (Edwards 1970). For example, Figs. 1A and 1B have the same tree topology but are different labeled histories. Apart from the ordering of the internal nodes, there is no difference between a tree topology and a labeled history, and the two terms are used interchangeably in this paper when there is no possibility of confusion. For s sequences, there are $\xi = s!(s-1)!/2^{s-1}$ distinct labeled histories (Edwards 1970). According to the birth–death process, each of these labeled histories has equal probability of occurrence.

For the birth–death process, the probability that a lineage is extinct after time t is (Kendall 1949)

$$p_0(t) = \frac{\mu(1 - e^{-(\lambda-\mu)t})}{\lambda - \mu e^{-(\lambda-\mu)t}} \tag{1}$$

The probability that a lineage leaves one descendent after time t is

$$p_1(t) = \frac{(\lambda - \mu)^2 e^{-(\lambda-\mu)t}}{(\lambda - \mu e^{-(\lambda-\mu)t})^2} \tag{2}$$

and the probability that a lineage leaves i descendents after time t is

$$p_i(t) = (\lambda/\mu)^i p_1(t)[p_0(t)]^{i-1} \tag{3}$$

Thompson (1975) derived the joint density of τ, **t,** and s, conditional on t_1 (the time of the first speciation event among surviving lineages) as

$$f(\tau, \mathbf{t}, s; \lambda, \mu) = \frac{2^{s-1}\lambda^{s-2}[p_1(t_1)]^2 \prod_{i=2}^{s-1} p_1(t_i)}{s!} \tag{4}$$

Since the number of species s is fixed, the distribution of the other random variables should be conditioned on s. The probability of observing s lineages at present, descended from two ancestral lineages that arose at time t_1, with each ancestral lineage leaving at least one descendent, is

$$f(s; \lambda, \mu) = \sum_{i=1}^{s-1} p_i(t_1)p_{s-i}(t_1) \tag{5}$$

$$= (s-1)(\lambda/\mu)^{s-2}[p_0(t_1)]^{s-2}[p_1(t_1)]^2 \tag{6}$$

Setting $t_1 = 1$, we obtain the joint density of τ and **t,** conditional on observing s species at present, as

$$f(\tau, \mathbf{t}; \lambda, \mu) = f(\tau, \mathbf{t}, s; \lambda, \mu)/f(s; \lambda, \mu) \tag{7}$$

$$= \frac{2^{s-1}\mu^{s-2} \prod_{i=2}^{s-1} p_1(t_i)}{[p_0(1)]^{s-2}s!(s-1)} \tag{8}$$

If we set $\mu = 0$, Eq. 8 reduces to the density for a Yule pure-birth process, originally derived by Edwards (1970):

$$f(\tau, \mathbf{t}; \lambda) = \frac{2^{s-1}\lambda^{s-2}\exp\left\{-\lambda \sum_{i=2}^{s-1} t_i\right\}}{s!(s-1)(1-e^{-\lambda})^{s-2}} \tag{9}$$

Model of Nucleotide Substitution

A continuous-time Markov process is used to model nucleotide substitution. The model used in J. Felsenstein's DNAML program (since 1984, version 2.6) will be used in this paper, although other substitution models are applicable as well. The substitution rate matrix under the model is

$$Q = \begin{pmatrix} \cdot & (1+\kappa/\pi_Y)\pi_C & \pi_A & \pi_G \\ (1+\kappa/\pi_Y)\pi_T & \cdot & \pi_A & \pi_G \\ \pi_T & \pi_C & \cdot & (1+\kappa/\pi_R)\pi_G \\ \pi_T & \pi_C & (1+\kappa/\pi_R)\pi_A & \cdot \end{pmatrix} \varphi m \tag{10}$$

where Q_{ij} ($i \neq j$) is the instantaneous substitution rate from nucleotide i to j, with the nucleotides ordered T, C, A, and G. The substitution process is assumed to be at stationarity with nucleotide frequencies given by π_T, π_C, π_A, and π_G, with $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$. Parameters π_i ($i = T,C,A,G$) can be estimated using the average frequencies of nucleotides over all sequences. The parameter κ is the transition/transversion rate ratio; a κ greater than zero indicates that transitions occur with greater frequency than transversions. The diagonals of the matrix are determined by the mathematical requirement that sums of rows of Q are zero (Grimmett and Stirzaker 1992:239–246), and $-Q_{ii} = \Sigma_{j \neq i} Q_{ij}$ is the substitution rate of nucleotide i. The scale factor φ is determined as $\varphi = 1/[2\pi_T\pi_C(1 + \kappa/\pi_Y) + 2\pi_A\pi_G(1 + \kappa/\pi_R) + 2\pi_Y\pi_R]$, so that $m = -\Sigma_i \pi_i Q_{ii}$ is the average substitution rate. We assume the existence of a molecular clock (i.e., rate constancy across lineages).

The transition-probability matrix over time t is then $P(t) = \{p_{ij}(t)\} = e^{Qt}$, where $p_{ij}(t)$ is the probability that nucleotide i transforms to j in time t. The calculation can be performed by the diagonalization of the rate matrix Q (e.g., Hasegawa et al. 1985; Thorne et al. 1992).

We assume that substitutions occur independently at different nucleotide sites; the conditional probability of observing the sequence data, given the tree topology (τ) and the divergence times **(t),** is then a product over sites

$$f(\mathbf{X}|\tau, \mathbf{t}; m, \kappa) = \prod_{j=1}^{n} f(\mathbf{x}_j|\tau, \mathbf{t}; m, \kappa) \tag{11}$$

where $f(\mathbf{x}_j|\tau, \mathbf{t}; m, \kappa)$ is the probability of observing the nucleotides at the jth site, conditional on topology τ and divergence times **t.** The exact form of Eq. 11 depends on the tree topology (τ). Consider the topology of Fig. 1A, denoted as τ_1, and let x_{5j}, x_{6j}, x_{7j} be the unknown nucleotides at site j in the ancestral sequences at nodes 5, 6, and 7, respectively. Then

$$\begin{aligned} f(\mathbf{x}_j|\tau_1, t_2, t_3; m, \kappa) &= \sum_{x_{7j}} \sum_{x_{6j}} \sum_{x_{5j}} \pi_{x_{7j}} p_{x_{7j}x_{5j}}(1-t_3) p_{x_{7j}x_{6j}}(1-t_2) \\ &\quad \times p_{x_{5j}x_{1j}}(t_3) p_{x_{5j}x_{2j}}(t_3) p_{x_{6j}x_{3j}}(t_2) p_{x_{6j}x_{4j}}(t_2) \qquad (12) \\ &= \sum_{x_{5j}} \sum_{x_{6j}} \pi_{x_{5j}} p_{x_{5j}x_{6j}}(2-t_2-t_3) p_{x_{5j}x_{1j}}(t_3) \\ &\quad \times p_{x_{5j}x_{2j}}(t_3) p_{x_{6j}x_{3j}}(t_2) p_{x_{6j}x_{4j}}(t_2) \qquad (13) \end{aligned}$$

Because the substitution model is time-reversible (i.e., $\pi_i Q_{ij} = \pi_j Q_{ji}$ for any i,j), the ''pulley'' principle of Felsenstein (1981) is applied to ''move'' the root of the tree from node 7 (Eq. 12) to node 5, eliminating one summation (Eq. 13). The conditional probabilities of other possible labeled histories are calculated similarly.

Maximum Likelihood Estimation of Parameters

The probability of observing the sequence data (i.e., the likelihood function) is

$$\begin{aligned} L(\lambda, \mu, m, \kappa|\mathbf{X}) &= f(\mathbf{X}; \lambda, \mu, m, \kappa) \\ &= \sum_{\tau} \int_{t_2=0}^{1} \cdots \int_{t_{s-1}=0}^{t_{s-2}} f(\mathbf{X}|\tau, \mathbf{t}; m, \kappa) \\ &\quad \times f(\tau, \mathbf{t}; \lambda, \mu)\, dt_{s-1} \cdots dt_2 \end{aligned} \tag{14}$$

where the summation is over the labeled histories and the integrations are over the divergence times. Analytical methods appear quite hopeless for evaluating the preceding equations and we instead used numerical integration to calculate approximate values of the likelihood function (see below). The dimension of the integration for a tree of s sequences is $s - 2$, and the computation is feasible only for a small number of species. Maximum likelihood estimates (MLEs) of parameters λ, μ, m, and κ are obtained by maximizing the log-likelihood function, $\ell = \log\{L\}$. A numerical optimization algorithm was used for this purpose.

Posterior Distribution of Phylogenetic Trees

After parameters λ, μ, *m*, and κ have been estimated, the (posterior) probability of the labeled history (τ), conditional on the observed sequence data, can be calculated as

$$f(\tau|\mathbf{X};\hat{\lambda},\hat{\mu},\hat{m},\hat{\kappa}) = \frac{f(\mathbf{X},\tau;\hat{\lambda},\hat{\mu},\hat{m},\hat{\kappa})}{f(\mathbf{X};\hat{\lambda},\hat{\mu},\hat{m},\hat{\kappa})} \quad (15)$$

where $f(\mathbf{X};\hat{\lambda},\hat{\mu},\hat{m},\hat{\kappa})$ in the denominator is given by Eq. 14, with $\hat{\lambda}$, $\hat{\mu}$, $\hat{m}$, and $\hat{\kappa}$ to be parameter estimates, and the joint probability in the numerator is

$$f(\mathbf{X},\tau;\hat{\lambda},\hat{\mu},\hat{m},\hat{\kappa}) = \int_{t_2=0}^{1}\cdots\int_{t_{s-1}=0}^{t_{s-2}}[f(\mathbf{X}|\tau,\mathbf{t};\hat{m},\hat{\kappa}) \times f(\tau,\mathbf{t};\hat{\lambda},\hat{\mu})]\,dt_{s-1}\cdots dt_2 \quad (16)$$

This is the ''contribution'' of tree topology τ to the likelihood function (see Eq. 14). The tree topology having the maximum posterior probability (or the greatest contribution to the likelihood function) is the MAP estimate of phylogeny. The posterior probability can be interpreted as the probability that the estimated tree is the true tree (under the models), providing a measure of the reliability of the estimated phylogeny. One can also construct a (minimum) set of most probable trees with the sum of their posterior probabilities constrained to be no less than a specified value, say 0.99. This is known as the highest posterior density ''credible set'' (Berger 1985:140–145) of phylogenies and is similar to the usual confidence interval of a parameter estimate.

Computational Methods

Calculation of the likelihood function (Eq. 14) involves evaluation of an $(s-2)$-dimensional integral for each labeled history. We have used the numerical approach of repeated one-dimensional integration, evaluating several algorithms for this procedure (see, e.g., Press et al. 1992: 129–164). The calculation does not seem practical with data for more than five species (i.e., a three-dimensional integral). One difficulty is that the functional form of the integrand in Eq. 14 changes with the tree topology (labeled history), although a computer algorithm was devised for performing the integration for any number of species and any tree topology. A further complication is that the probabilities in Eqs. 11 and 13 are very small and cause underflows in the computer. The probabilities also vary greatly for different values of **t,** which makes the choice of a scaling factor (to prevent underflows) difficult. To overcome this problem we calculated the log likelihood (ℓ) using the formula

$$\ell = C + \log\left\{\sum_{\tau}\int_{t_2=0}^{1}\cdots\int_{t_{s-1}=0}^{t_{s-2}}\exp\left(\sum_{j=1}^{n}\log\{f(\mathbf{x}_j|\tau,\mathbf{t};m,\kappa)\} - C\right) \times f(\tau,\mathbf{t};\lambda,\mu)\,dt_{s-1}\cdots dt_2\right\}, \quad (17)$$

where *C* is a scaling factor which may differ according to the values of λ, μ, *m*, and κ but is the same for different topologies (τ) and divergence times **(t).**

Analysis of Example Data Sets

We analyzed two data sets, one of nuclear DNAs and another of mitochondrial DNAs from several primate species. Both the Yule pure-birth process and the birth–death process were used as branching models. The birth–death process did not provide a better fit to the data (in terms of the likelihood score) than the Yule process for either data set, and the estimated death rate $\hat{\mu}$ was zero in both cases. The results presented below are therefore obtained using the Yule process model.

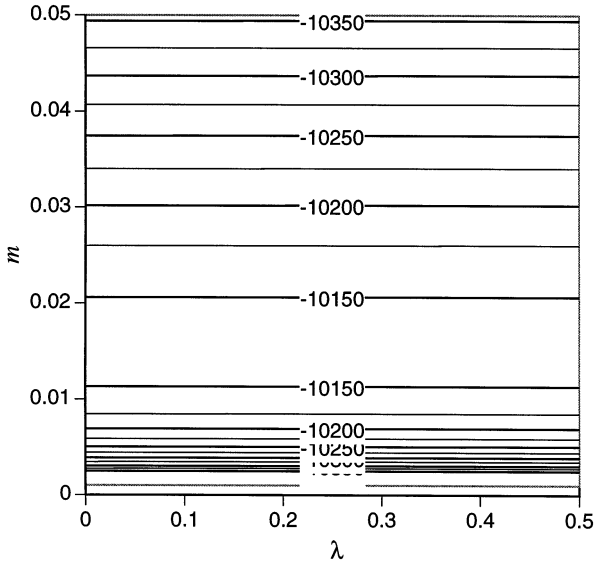

Fig. 2. The log-likelihood (ℓ) contour as a function of the branching rate λ of the Yule process and the substitution rate *m.* The transition/transversion rate ratio κ is fixed at its maximum likelihood estimate ($\hat{\kappa} = 2.127$). The results are for the ψη-globin pseudogenes of human, chimpanzee, gorilla, and orangutan.

ψη-globin Pseudogenes

The ψη-globin pseudogenes of human, chimpanzee, gorilla, and orangutan (Miyamoto et al. 1987) are analyzed. There are 6,166 nucleotide sites in each sequence, 5,891 (95.5%) of which are identical across species. The average nucleotide frequencies estimated from these data are $\hat{\pi}_T = 0.3074$, $\hat{\pi}_C = 0.1853$, $\hat{\pi}_A = 0.3072$, and $\hat{\pi}_G = 0.2000$.

Maximum likelihood estimates of the parameters are $\hat{\lambda} = 0.000$, $\hat{m} = 0.015 \pm 0.001$, and $\hat{\kappa} = 2.127 \pm 0.345$ (standard errors were calculated numerically by inverting the second-order derivatives of the log likelihood with respect to the parameters). The log-likelihood surface is shown in Fig. 2 as a function of parameters λ and *m*, with the transition/transversion rate ratio κ fixed at the MLE ($\hat{\kappa} = 2.12679$). The estimate of λ is subject to a large sampling error, as is clear from the log-likelihood surface which is almost flat with respect to λ. In contrast, estimates of *m* and κ are much more reliable.

Table 1 lists the conditional log-likelihood value for each topology, calculated as $\log\{f(\mathbf{X}|\tau;\hat{\lambda},\hat{m},\hat{\kappa})\} = \log\{\xi \times f(\mathbf{X},\tau;\hat{\lambda},\hat{m},\hat{\kappa})\}$ at the maximum-likelihood estimates of parameters under the MAP method. This is comparable with the log-likelihood value calculated using the ML method. The former is an average (integration) over the distribution of divergence times t_2 and t_3 and is always smaller than the latter, which is calculated using the MLEs of t_2 and t_3 for the particular topology considered. The order of the three best trees is identical by the two methods (Table 1). These trees all have the orangutan as the first species to diverge. The remaining 15 labeled histories have very small posterior probabilities by the MAP method, while in the ML analysis, each of the 12 remaining topologies (corresponding to the 15 labeled histories) has at least one zero interior branch length.

We also calculated the bootstrap proportions of different phylogenies (Felsenstein 1985) using the approximate method of resampling estimated log likelihoods (RELL, Kishino and Hasegawa 1989). The

Table 1. Comparison of different labeled histories (tree topologies) using the MAP method of this paper and the ML method[a]

	MAP method		ML method					
Topology	$\log\{f(\mathbf{X}\|\tau)\}$	$f(\tau\|\mathbf{X})$	ℓ	P(RELL)	$\hat{m}$	$\hat{t}_2$	$\hat{t}_3$	$\hat{\kappa}$
τ_1 = (((12)3)4)	-10,137.01	0.842	-10,132.33	0.587	0.015	0.524	0.495	2.133
τ_2 = (((23)1)4)	-10,138.86	0.133	-10,133.90	0.334	0.016	0.521	0.500	2.107
τ_3 = (((13)2)4)	-10,140.51	0.025	-10,135.56	0.079	0.016	0.517	0.503	2.070
τ_4 = (((34)1)2)	-10,185.92	0.000	-10,171.34	0.000	0.012	1.000	0.981	2.122
τ_5 = (((24)1)3)	-10,188.58	0.000	-10,173.54	0.000	0.012	1.000	1.000	2.076
τ_6 = (((14)3)2)	-10,186.20	0.000	-10,171.93	0.000	0.012	1.000	0.980	2.096
τ_7 = (((14)2)3)	-10,186.35	0.000	-10,171.93	0.000	0.012	1.000	0.980	2.096
τ_8 = (((24)3)1)	-10,189.86	0.000	-10,173.54	0.000	0.012	1.000	1.000	2.076
τ_9 = (((34)2)1)	-10,187.34	0.000	-10,171.34	0.000	0.012	1.000	0.981	2.122
τ_{10} = (((12)4)3)	-10,178.83	0.000	-10,167.08	0.000	0.012	1.000	0.897	2.127
τ_{11} = (((13)4)2)	-10,183.21	0.000	-10,171.57	0.000	0.012	1.000	0.922	2.055
τ_{12} = (((23)4)1)	-10,183.03	0.000	-10,169.43	0.000	0.012	1.000	0.928	2.104
τ_{13} = ((12)(34))	-10,180.25	0.000	-10,167.08	0.000	0.012	1.000	0.897	2.127
τ_{14} = ((34)(12))	-10,186.02	0.000						
τ_{15} = ((13)(24))	-10,184.89	0.000	-10,171.57	0.000	0.012	1.000	0.922	2.055
τ_{16} = ((24)(13))	-10,189.74	0.000						
τ_{17} = ((23)(14))	-10,183.17	0.000	-10,169.43	0.000	0.012	1.000	0.928	2.104
τ_{18} = ((14)(23))	-10,186.48	0.000						

[a] The ψη-globin pseudogenes (6166 bp) of human (1), chimpanzee (2), gorilla (3) and orangutan (4) are analyzed. MLEs of parameters under the MAP approach (assuming the Yule process prior) are $\hat{\lambda} = 0.000$, $\hat{m} = 0.015$ and $\hat{\kappa} = 2.127$, with $\ell = -10{,}139.73$. In the ML method, parameters are estimated independently for each topology, and ℓ is the log likelihood. In the labeled history ((12)(34)), the separation of species 1 and 2 occurred after the separation of species 3 and 4, while in the labeled history ((34)(12)) the opposite is true (see Fig. 1). Labeled histories τ_{13} and τ_{14} have the same topology, as do τ_{15} and τ_{16}, and τ_{17} and τ_{18}

bootstrap proportions for the three best trees, i.e., ((human, chimpanzee), gorilla), ((chimpanzee, gorilla), human), and ((human, gorilla), chimpanzee)—with the orangutan diverging first in all cases—are 0.59, 0.33, and 0.08, while the posterior probabilities are 0.84, 0.13, and 0.03.

Estimates of κ from the likelihood analysis of different topologies are very similar and are close to the estimate obtained from the present method. The divergence times in the tree can be estimated in the present model by using the conditional mean $E(t_2,t_3|\mathbf{X}; \hat{\lambda}, \hat{m}, \hat{\kappa})$, but this is not pursued in this study as our primary interest is the tree topology. The total tree length (i.e., the sum of branch lengths along the tree), calculated as $2\hat{m}(1 + \hat{t}_2 + \hat{t}_3)$ is much more similar across tree topologies than separate estimates of t_2 and t_3. The estimates of m from the three best trees obtained using the ML method are similar to one another and to the estimate obtained using the MAP method.

Mitochondrial tRNA Genes

The second data set to be analyzed comprises 11 mitochondrial tRNA genes (739 nucleotides in each sequence) of the human, common chimpanzee, pygmy chimpanzee, gorilla, and orangutan. The data, together with some protein-coding genes, were used by Horai et al. (1992) and Takezaki et al. (1995) to calculate the divergence times of the species. The average nucleotide frequencies estimated from these data are $\hat{\pi}_T = 0.2625$, $\hat{\pi}_C = 0.2472$, $\hat{\pi}_A = 0.3378$, and $\hat{\pi}_G = 0.1526$.

Maximum-likelihood estimates of parameters are $\hat{\lambda} = 1.430 \pm 2.227$, $\hat{m} = 0.051 \pm 0.006$, and $\hat{\kappa} = 20.045 \pm 8.642$, with $\ell = -1{,}563.78$. Compared with the estimates obtained for the ψη-globin genes, the mitochondrial genes have a high substitution rate (m) and a very biased transition/transversion rate ratio (κ). Once again, the estimate of λ involves a large sampling error while the estimate of m is much more reliable.

The MAP tree is shown in Fig. 4, which is probably also the correct phylogeny of these species. The posterior probability for this tree, calculated using the MLEs for the parameters, is close to one (0.9999). In an ML analysis of separate topologies, the same substitution model and the molecular clock are assumed. The ML tree is also the one shown in Fig. 4, with log likelihood $\ell = -1{,}554.75$ and parameter estimates $\hat{m} = 0.051$, $\hat{\kappa} = 20.033$, $\hat{t}_2 = 0.591$ for the divergence of the gorilla, $\hat{t}_3 = 0.370$ for the divergence of the human, and $\hat{t}_4 = 0.168$ for the separation of the two chimpanzee species. We have also used the RELL approximate method of Kishino and Hasegawa (1989) to calculate the bootstrap proportion for the ML tree. This is 0.89 and is much lower than the posterior probability calculated from the MAP method. The second best tree in the ML analysis has topology ((human, (common chimpanzee, pigmy chimpanzee)), (gorilla, orangutan)), with a bootstrap proportion of 0.09.

Discussion

Prior Distribution of Phylogenies

The birth–death process has been widely used as a model of the speciation and extinction process, both by evolutionary biologists (Yule 1925; Thompson 1975; Nee et al. 1994) and by paleontologists (Raup 1985). However, neither the simple birth–death process, nor a submodel of it, the Yule process, is likely to accurately describe the actual process of speciation and extinction, especially when we consider the additional effect of species sampling by biologists. A more realistic prior branching process may be used in the present method as long as the joint distribution of phylogeny and divergence times can be derived. The birth–death process assigns equal probability to each labeled history, which seems reasonable

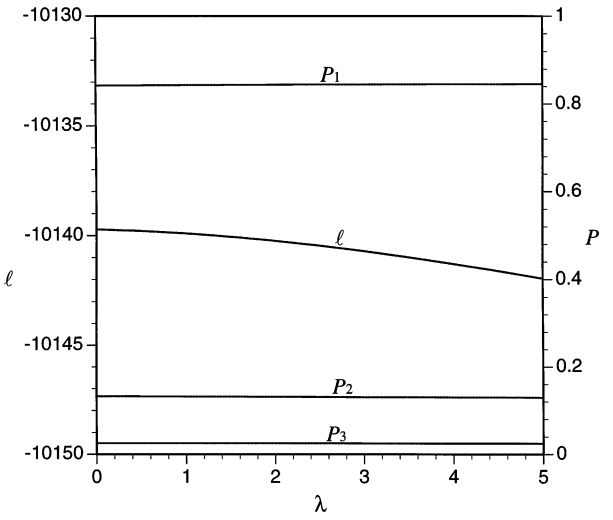

Fig. 3. The log-likelihood (ℓ) value and the posterior probabilities of the three best trees, τ_1 = ((human, chimpanzee), gorilla), τ_2 = ((chimpanzee, gorilla), human), and τ_3 = ((human, gorilla), chimpanzee)—with the orangutan diverging first in all cases (see table 1)—plotted as a function of the branching rate λ of the Yule process. Parameters *m* and κ are fixed at their maximum-likelihood estimates (0.015 and 2.127, respectively). The results are for the ψη-globin pseudogenes.

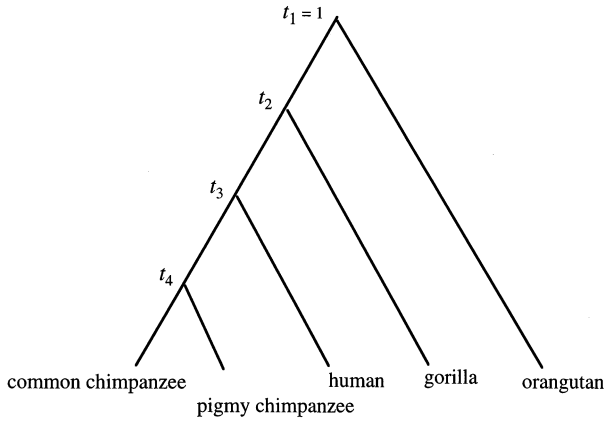

Fig. 4. The MAP (and also ML) tree for the mitochondrial tRNA genes of human, common chimpanzee, pygmy chimpanzee, gorilla, and orangutan.

given that we usually have no prior knowledge about the phylogeny.

An important question is how sensitive the calculated posterior probabilities are to the prior distribution. We examined this problem by studying the change in the posterior probabilities of the three best trees from the analysis of the ψη-globin pseudogenes (see Table 1) when the branching rate λ of the Yule process is varied, parameters *m* and κ being fixed at their MLEs. The results are shown in Fig. 3. The posterior probabilities appear insensitive to the value of λ, suggesting that most of the information concerning the tree topology comes from the sequence data rather than the prior distribution specified by the Yule process. It is not clear whether this insensitivity is due to the small number of species in the study or is a more general property of the method.

Extensions of the Method

The substitution model used in this paper accounts for different nucleotide frequencies and transition/transversion rate bias. It is easy to extend the model to account for among-site rate variation (Yang 1993, 1994b); for example, the discrete-gamma model of Yang (1994b) can be incorporated into the present method in a straightforward manner. Indeed, the conditional probability of Eq. 11 is the likelihood function for topology τ in the ML analysis of topology, so any substitution model developed for maximum-likelihood phylogenetic analysis can be used in the MAP method.

For the ψη-globin genes and mitochondrial tRNA genes analyzed in this paper, the assumption of a molecular clock is statistically acceptable (Yang et al. 1995, Z. Yang unpublished results). For data from more distantly related species, the molecular clock assumption is often violated, and it is well known that ignoring rate variation among lineages can lead to incorrect phylogenetic estimates. To relax the molecular clock assumption, one might consider using independent rate parameters for different branches in the tree, as is done in the likelihood analysis of separate topologies. Another possibility might be to construct a stochastic process for the change of substitution rate over lineages, but this appears difficult.

Since parameter estimates obtained using other methods can also be used to evaluate the posterior probabilities of trees, it may be worthwhile to use approximate methods to estimate parameters. In fact, apart from λ and μ, other parameters such as *m* and κ can be reliably estimated using traditional ML methods, and ad hoc methods for estimating these parameters are not difficult to devise. Another possible method for approximating the likelihood is Monte Carlo integration combined with importance sampling (Robert 1994). As most of the many possible phylogenies (labeled histories) contribute little to the likelihood (see Table 1), sampling phylogenies might also be useful for estimation purposes (Kuhner et al. 1995).

Posterior Probabilities and Bootstrap Proportions

The results of MAP analyses of the ψη-globin pseudogenes and the mitochondrial tRNA genes appear reasonable. The ordering of the posterior probabilities of trees generated by the method appear to correspond with accepted theories concerning the pattern of hominoid evolution. The posterior probabilities obtained from the

MAP method are generally more extreme than the bootstrap proportions. Previous studies suggest that the bootstrap method provides a conservative test of the significance of the estimated phylogeny; it underestimates the probability when the probability is high and overestimates the probability when the probability is low (Zharkikh and Li 1992; Hillis and Bull 1993). The patterns of posterior probabilities and bootstrap proportions found in this study (e.g., Table 1) suggest that the posterior probabilities calculated in this paper may be more reliable in measuring the accuracy of the estimated phylogeny. Nevertheless, it is probably worthwhile to perform simulations to evaluate the accuracy of the posterior probabilities calculated in the present method when the birth–death process model is not assumed.

Program availability: The method described in this paper has been implemented in the PAML (Phylogenetic Analysis by Maximum Likelihood) program package, which is available by anonymous ftp at ftp.bio.indiana.edu in the directory molbio/evolve.

Acknowledgments. We thank Drs. J. Felsenstein, M. Hasegawa, J.A. Hartigan, and J.P. Huelsenbeck for their comments on an earlier version of this paper. Support for B.R. was provided by a Natural Sciences and Engineering Research Council (NSERC) of Canada postdoctoral fellowship. Support for Z.Y. was provided by a grant from the National Institute of Health (GM40282) to M. Slatkin.

References

Berger JO (1985) Statistical decision theory and Bayesian analysis. Springer-Verlag, New York

Bishop MJ, Friday AE (1985) Evolutionary trees from nucleic acid and protein sequences. Proc R Soc Lond [Biol] 226:271–302

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Evolution 21:550–570

Edwards AWF (1970) Estimation of the branch points of a branching diffusion process (with discussion). J R Stat Soc B 32:155–174

Feller W (1939) Die grundlagen der volterraschen theorie des kampfer ums dasein in wahrsheinlichkeits theoretischen behandlung. Acta Biotheor 5:1–40

Felsenstein J (1973) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst Zool 22:240–249

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

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Felsenstein J, Kishino H (1993) Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Syst Biol 42:193–200

Fukami-Kobayashi K, Tateno Y (1991) Robustness of maximum likelihood tree estimation against different patterns of base substitution. J Mol Evol 32:79–91

Gaut BS, Lewis PO (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. Mol Biol Evol 12:152–162

Goldman N (1990) Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analysis. Syst Zool 39:345–361

Grimmett GR, Stirzaker DR (1992) Probability and Random Processes. 2nd ed. Clarendon Press, Oxford

Hasegawa M, Yano T (1984) Maximum likelihood method of phylogenetic inference from DNA sequence data. Bull Biomet Soc Jpn 5:1–7

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

Hasegawa M, Kishino H, Saitou N (1991) On the maximum likelihood method in molecular phylogenetics. J Mol Evol 32:443–445

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing the confidence in phylogenetic analysis. Syst Biol 42:182–192

Horai S, Satta Y, Hayasaka K, Kondo R, Inoue T, Ishida T, Hayashi S, Takahata N (1992) Man's place in Hominoidea revealed by mitochondrial DNA genealogy. J Mol Evol 35:32–43

Huelsenbeck JP (1995a) The performance of phylogenetic methods in simulation. Syst Biol 44:17–48

Huelsenbeck JP (1995b) The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol Biol Evol 12:843–849

Kendall DG (1949) Stochastic processes and population growth. J R Stat Soc B 11:230–264

Kishino H, Hasegawa M (1989) Evaluation of maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol 29:170–179

Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459–468

Kuhner MK, Yamato J, Felsenstein J (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 140:1421–1430

Miyamoto MM, Slighton JL, Goodman M (1987) Phylogenetic relations of humans and African apes from DNA sequences in the ψη-globin region. Science 238:369–373

Nee S, May RM, Harvey PH (1994) The reconstructed evolutionary process. Philos Trans R Soc Lond Biol 344:305–311

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C: the art of scientific computing. 2nd ed. Cambridge University Press, Cambridge

Raup DM (1985) Mathematical models of cladogenesis. Paleobiology 11:42–52

Robert CP (1994) The Bayesian choice: a decision-theoretic motivation. Springer-Verlag, New York

Takezaki N, Rzhetsky A, Nei M (1995) Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol 12:823–833

Tateno Y, Takezaki N, Nei M (1994) Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol Biol Evol 11:261–277

Thompson EA (1975) Human evolutionary trees. Cambridge University Press, Cambridge, England

Thorne JL, Kishino H, Felsenstein J (1992) Inching toward reliability: an improved likelihood model of sequence evolution. J Mol Evol 34:3–16

Yang Z (1993) Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396–1401

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

Yang Z (1994b) Maximum likelihood phylogenetic estimation from

DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314

Yang Z (1995) Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. J Mol Evol 40:689–697

Yang Z (1996) Phylogenetic analysis by parsimony and likelihood methods. J Mol Evol 42:294–307

Yang Z, Goldman N, Friday AE (1995) Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst Biol 44:384–399

Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. Philos Trans R Soc Lond Biol 213:21–87

Zharkikh A, Li W-H (1992) Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: I. four taxa with a molecular clock. Mol Biol Evol 9:1119–1147