

## Inferring the Root of a Phylogenetic Tree

JOHN P. HUELSENBECK, JONATHAN P. BOLLECK, AND AMY M. LEVINE

*Department of Biology, University of Rochester, Rochester, New York 14627, USA;  
E-mail: johnh@brahms.biology.rochester.edu*

**Abstract.**—Phylogenetic trees can be rooted by a number of criteria. Here, we introduce a Bayesian method for inferring the root of a phylogenetic tree by using one of several criteria: the outgroup, molecular clock, and nonreversible model of DNA substitution. We perform simulation analyses to examine the relative ability of these three criteria to correctly identify the root of the tree. The outgroup and molecular clock criteria were best able to identify the root of the tree, whereas the nonreversible model was able to identify the root only when the substitution process was highly nonreversible. We also examined the performance of the criteria for a tree of four species for which the topology and root position are well supported. Results of the analyses of these data are consistent with the simulation results. [Bayesian estimation; hierarchical Bayes; nonreversible models; outgroup; rooting.]

The root of a phylogenetic tree is usually determined by using the outgroup method; one or more of the species are assumed to fall outside of the species group of interest (denoted the ingroup) and the branch where the outgroup connects to the ingroup becomes the root of the ingroup tree. However, several other methods also can be used to root phylogenetic trees: nonreversible models of substitution (or asymmetric step matrices), the molecular clock, midpoint rooting, or paleontological (temporal) data.

The outgroup criterion has been studied extensively (Maddison et al., 1984; Wheeler, 1990a) and is expected to be a powerful method for rooting phylogenetic trees. The outgroup method makes minimal assumptions, the main one being that the outgroup species fall outside of the ingroup species. The quality of the rooting provided by the outgroup criterion should depend on the sampling of the outgroup species (Swofford et al., 1996) and on their phylogenetic proximity to the ingroup (Wheeler, 1990a). Yet, the outgroup criterion is of little help when there are no good outgroups, as is the case when the Tree of Life is rooted, or when the outgroup species are very distantly related to the ingroup species (Wheeler, 1990a).

Surprisingly little research has examined alternative methods for rooting phylogenetic trees. In a parsimony framework, asymmetric step matrices require consideration of rooted phylogenetic trees because different root positions imply different tree lengths (Sankoff, 1975; Sankoff and Rousseau, 1975; Swofford et al., 1996). One of the problems with using parsimony rooting with asym-

metric step matrices is the difficulty in determining what weights should be assigned to different character transformations. Several methods can potentially be used to determine step matrices, but these methods are ad hoc (e.g., Williams and Fitch, 1989; Wheeler, 1990b). Yang (1994a) compared the fit of the nonreversible model of DNA substitution to the most general time-reversible model by using likelihood ratio tests. He found that the nonreversible model did not provide a significant improvement over the general reversible model for two datasets of relatively closely related species (primates). Like asymmetric step matrices, the nonreversible model forces a root to the phylogenetic trees, although for the two datasets that Yang (1994a) examined, the likelihood did not change much when the tree was rooted at different branches. One of the advantages of using the nonreversible model in a likelihood framework is that the parameters of the model can be estimated by using maximum likelihood. To our knowledge, nothing has been published on nonreversible models of DNA substitution since Yang's (1994a) contribution.

Phylogenetic trees constructed under the molecular clock assumption are also rooted. The molecular clock is often used in phylogenetic analyses, although usually only in the course of testing the molecular clock assumption (using, for example, a likelihood ratio test). Typically, the molecular clock is rejected, suggesting that rates of DNA substitution change along branches. Although the clock is usually rejected, little is known about the sensitivity of phylogenetic methods to violation of the clock assumption or

about the quality of the rooting provided by the molecular clock.

In this paper, we compare three different methods for estimating the root of a phylogenetic tree: the outgroup criterion, the non-reversible model of DNA substitution, and the molecular clock. The Bayesian method we present for inferring the root of a phylogenetic tree provides the posterior probability that the root lies on any branch of the ingroup topology.

## METHODS

We have concentrated on the analysis of DNA sequences. However, the methods we present and our results can be extended to other types of data, such as amino acid sequences and morphological data, by using stochastic models of character change. The aligned DNA sequences are contained in a matrix  $X = \{x_{ij}\}$ , where  $i = 1, 2, \dots, s$  and  $j = 1, 2, \dots, c$ ;  $s$  is the number of species, and  $c$  is the length of the sequences. The  $j$ th site is contained in the vector  $x_j = \{x_{1j}, x_{2j}, \dots, x_{sj}\}$ . Each element of the matrix,  $x_{ij}$ , can take one of the four nucleotide states: A, C, G, or T.

We assume that the  $s$  species are related by a bifurcating tree,  $\tau$ . Figure 1a shows an example of an unrooted tree of  $s = 5$  species. An unrooted phylogenetic tree can be rooted along any of its  $b = 2s - 3$  branches. Figure 1b shows the seven possible rooted trees that can be produced from the tree of Figure 1a. The lengths of the branches of the tree are denoted  $\nu = \{v_1, v_2, \dots, v_b\}$  and are expressed in terms of the number of substitutions per site that are expected to occur along the branch. We label the tips of a rooted tree  $1, \dots, s$  and the internal nodes  $s + 1, \dots, 2s - 1$ ; the root of the tree is always labeled  $2s - 1$ . The ancestor of node  $k$  is denoted  $\sigma(k)$ .

A phylogenetic model includes not only a tree with branch lengths but also a mechanism of character change along the branches of the tree. We assume that DNA substitution follows a continuous-time Markov chain. The instantaneous rate of change,  $Q$ , for the chain is

$$Q = \{q_{ij}\} = \begin{pmatrix} - & a & b & c \\ g & - & d & e \\ h & i & - & f \\ j & k & l & - \end{pmatrix} \quad (1)$$

The diagonals of the rate matrix are specified by the requirement that the rows sum to 0. Because the instantaneous rate matrix does not change over the tree, the process is said to be time-homogeneous. This is the most general model of DNA substitution. The transition probabilities over a branch of length  $t$  are  $P(t) = \{p_{ij}(t)\} = e^{Qt}$ , where  $P(t)$  is a matrix containing the probability of a change from nucleotide  $i$  to nucleotide  $j$  over a branch of length  $t$ . If the matrix  $Q$  is irreducible (i.e., all states communicate), then the chain has a stationary distribution:

$$\lim_{t \rightarrow \infty} P(t) = \Pi_t = \begin{pmatrix} \pi_A \\ \pi_C \\ \pi_G \\ \pi_T \end{pmatrix} \quad (2)$$

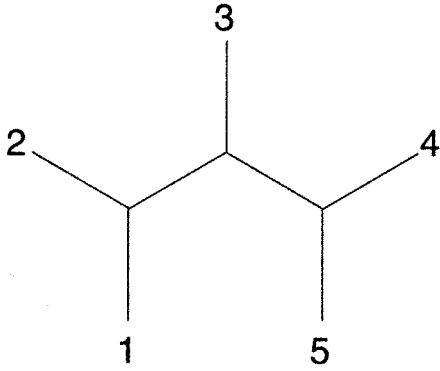
where  $\pi = \{\pi_A, \pi_C, \pi_G, \pi_T\}$  is the stationary frequency of the nucleotides. The stationary distribution will not change with time, so that  $\pi Q = 0$  (where 0 is a column vector of zeros). The substitution model is reversible if  $\pi_i q_{ij} = \pi_j q_{ji}$  for all  $i$  and  $j$ . A simple index of the degree to which a substitution matrix is nonreversible is  $I = \sum_{i,j} |\pi_i q_{ij} - \pi_j q_{ji}|$ . When the substitution process is reversible,  $I = 0$  and the likelihood is the same regardless of the placement of the root of the tree.

Note that the transition probabilities depend on the product of  $Q$  and  $t$ ; the rate matrix  $Q$  is known up to a multiplicative factor. When at least one speciation event on the tree is known and a molecular clock is assumed, then  $Q$  is the rate of substitution per unit time. Typically, however, time on the tree is unknown; branch lengths on the tree, then, are given in terms of expected number of substitutions per site ( $\nu$ ). The matrix  $Q$  can be rescaled such that branch lengths are expressed in terms of expected number of substitutions per site by adding the constraint that  $-\sum \pi_i q_{ii} = 1$ .

The Markov chain formed by the instantaneous rate matrix, discussed above, may not be time-reversible. Most models of DNA substitution used in phylogenetics, however, are time-reversible. The most general time-reversible model of DNA substitution is referred to as the GTR model, and has instantaneous rates

$$Q = \{q_{ij}\} = \begin{pmatrix} - & \pi_C a & \pi_G b & \pi_T c \\ \pi_A a & - & \pi_G d & \pi_T e \\ \pi_A b & \pi_C d & - & \pi_T f \\ \pi_A c & \pi_C e & \pi_G f & - \end{pmatrix} \quad (3)$$

(a)



(b)

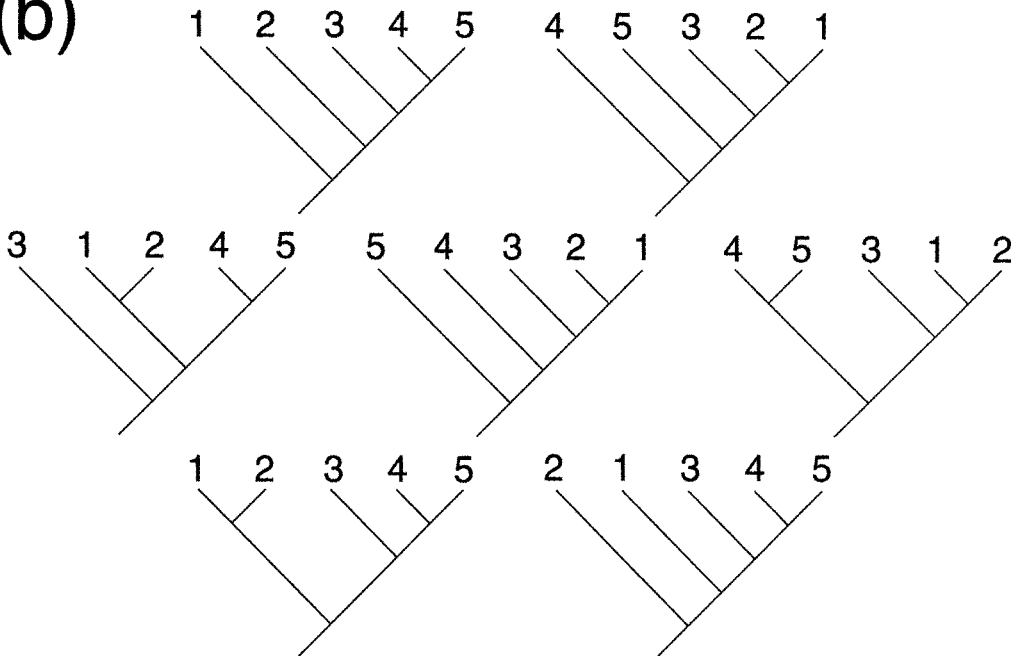


FIGURE 1. An unrooted tree (a) and the seven possible rooted trees (b) obtained by placing the root at the different branches of the unrooted tree.

(Tavaré, 1986). This model is a special case of the nonreversible model. Other models of DNA substitution are special cases of the GTR (and nonreversible) model. For example, the HKY85 model of DNA substitution (Hasegawa et al., 1984, 1985) sets  $b = e = \kappa$  and  $a = c = d = f = 1$ .

Rate variation across sites can be accommodated in several ways. Here, we assume that the rate of substitution at a site is a random variable drawn from a mean one-gamma distribution with shape parameter  $\alpha$  (Yang, 1993). We use the approximation suggested by Yang (1994b) in which the gamma distribution is broken into  $K$  discrete categories, each with equal weight. The mean rate from each category represents the rate for the entire category. Models that assume gamma-distributed rate variation among sites are denoted “+  $\Gamma$ ”. Parameters of the substitution model are contained in the vector  $\theta = \{a, b, c, d, e, f, g, h, i, j, k, l, \pi, \alpha\}$ .

The likelihood is the probability of observing the DNA sequences, given a phylogenetic model ( $\tau$ ,  $\nu$ , and  $\theta$ ). The probability of observing the data at the  $i$ th site is

$$f(x_i | \tau, \nu, \theta) = \sum_{j=1}^K \left\{ \sum_y \left[ \pi_{y_{\sigma(2s-1)i}} \left( \prod_{k=1}^s p_{y_{\sigma(k)i} x_{ki}}(v_k r_j) \right) \times \left( \prod_{k=s+1}^{2s-1} p_{y_{\sigma(k)i} y_{ki}}(v_k r_j) \right) \right] \right\} \frac{1}{K} \quad (4)$$

where  $v_k$  is the length of the  $k$ th branch,  $r_j$  is the rate for the  $j$ th gamma category, and  $y$  is a generic data vector for the (unobserved) states at the internal nodes of the tree. The likelihood of a site is a sum over all  $4^{s-1}$  possible assignments of nucleotides to the internal nodes of the tree. Felsenstein (1981) describes a pruning algorithm for efficiently performing this summation. Assuming independence of the substitutions across sites, the probability of observing the aligned sequences is

$$f(X | \tau, \nu, \theta) = \prod_{i=1}^c f(x_i | \tau, \nu, \theta) \quad (5)$$

Bayesian inference is based on the posterior probability of a parameter. The joint

posterior probability density of the tree topology, branch lengths, and substitution parameters is

$$f(\tau, \nu, \theta | X) = \frac{f(X | \tau, \nu, \theta) f(\tau, \nu, \theta)}{f(X)} \quad (6)$$

where

$$f(X) = \sum_{\tau} \int_{\nu, \theta} f(X | \tau, \nu, \theta) f(\tau, \nu, \theta) d\nu d\theta \quad (7)$$

and the summation is over all possible trees, and the integral is over the space of  $\nu$  and  $\theta$ . Typically, the phylogeny ( $\tau$ ) is of interest; inferences of phylogeny are then based on the marginal posterior probabilities of the trees (Li, 1996; Mau, 1996; Rannala and Yang, 1996; Mau and Newton, 1997; Yang and Rannala, 1997; Larget and Simon, 1999; Mau et al., 1999; Newton et al., 1999). For this study, however, we are interested in the root of the tree and base our inferences on the marginal posterior probability density of the root position. We assume that the topology is fixed, thereby eliminating the summation over all possible trees. However, the machinery we use to evaluate the joint posterior probability density works well if the topology is also treated as a random variable.

We evaluated the joint posterior probability density of parameters of the phylogenetic model by using Markov chain Monte Carlo (MCMC). Specifically, we used the Metropolis–Hastings–Green (MHG) algorithm (Metropolis et al., 1953; Hastings, 1970; Green, 1995). The MHG algorithm works by constructing a Markov chain as follows: (1) The current state of the chain is  $\psi = \{\tau, \nu, \theta\}$ . If this is the first step of the chain, then  $\psi$  is initialized by choosing arbitrary values for the parameters. (2) A new state for the chain is proposed,  $\psi'$ . The probability of proposing the new state, given the old state, is  $f(\psi' | \psi)$ . The probability of making the reverse move (which is not actually made) is  $f(\psi | \psi')$ . (3) The new state is accepted with the following probability:

$$R = \min \left( 1, \frac{f(X | \psi') f(\psi') / f(X)}{f(X | \psi) f(\psi) / f(X)} \times \frac{f(\psi | \psi')}{f(\psi' | \psi)} \right) \quad (8)$$

$$= \min \left( 1, \frac{f(\mathbf{X}|\psi')}{f(\mathbf{X}|\psi)} \times \frac{f(\psi')}{f(\psi)} \times \frac{f(\psi|\psi')}{f(\psi'|\psi)} \right) \quad (9)$$

In words, the acceptance probability is equal to the likelihood ratio times the prior ratio times the proposal ratio. Steps 1 to 3 are repeated a large number of times. The states of the chain are sampled and a histogram of any parameter of the sampled chain is an approximation to the posterior probability density of that parameter.

Our approach is to update parameters individually. We change branch lengths and the root position by using the LOCAL mechanism of Larget and Simon (1999). We assume a uniform (0.001, 100) prior for the elements of the  $Q$  matrix and propose a new state for the rate matrix by randomly picking one of the elements to change and then changing the parameter by using a sliding window. The sliding window mechanism was described by Larget and Simon (1999). A window of fixed width is centered around the current value of the parameter. A new state is proposed by uniformly and at random choosing a state from the window. If the proposed state is outside of a preset bound (e.g., the substitution rate is negative), the excess is reflected back into the valid parameter space. When the  $Q$  matrix is nonreversible, we determine the base frequencies ( $\pi$ ) by solving  $\pi Q = 0$ . However, when the model is reversible, we propose new base frequencies from a Dirichlet distribution (see Huelsenbeck et al., 2000).

We evaluated the ability of the outgroup criterion, the nonreversible model, and the molecular clock to correctly infer the root of the phylogenetic tree, using simulation. The eight-taxon tree shown in Figure 2 was the model tree for the simulations. The length of each branch was 0.1 expected substitutions per site. The root of the tree is indicated by the dot. Note that the lengths of the branches satisfy the clock assumption. The topology of the tree can be described by its taxon bipartitions. A taxon bipartition is formed by removing a branch, thereby separating the species into two groups—those to the left and those to the right of the removed branch. The taxon bipartitions for the simulated tree are {1},{2,3,4,5,6,7,8}; {2},{1,3,4,5,6,7,8}; {3},{1,2,4,5,6,7,8}; {4},{1,2,3,5,6,7,8}; {5},{1,2,3,4,6,7,8};

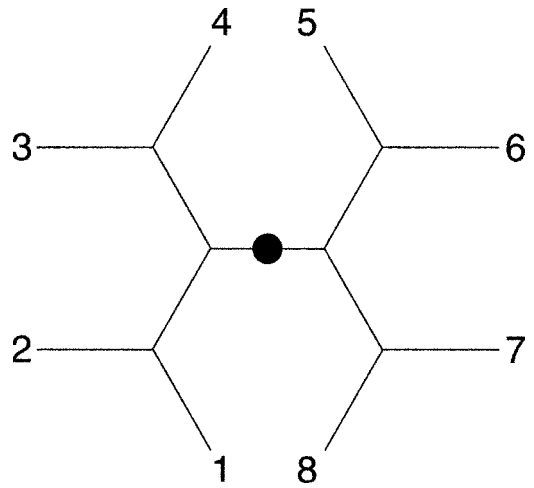


FIGURE 2. The model tree for the simulations. All of the branches were  $v = 0.1$  expected substitutions per site in length. The root of the tree is at the dot.

{6},{1,2,3,4,5,7,8}; {7},{1,2,3,4,5,6,8}; {8},{1,2,3,4,5,6,7}; {1,2},{3,4,5,6,7,8}; {3,4},{1,2,5,6,7,8}; {5,6},{1,2,3,4,7,8}; {7,8},{1,2,3,4,5,6}; and {1,2,3,4},{5,6,7,8}. We will calculate the posterior probability that the root falls on any one of the taxon bipartitions. The true root position is on taxon bipartition {1,2,3,4},{5,6,7,8}.

What is the probability of a random rooting of this tree? If the root is placed at random and uniformly along the branches according to their length, then the probability of any single branch being the root is simply  $1/13 = 0.077$  for the simulated data. If some of the branches are much longer than others, then the probability of the root being placed along that branch is a priori higher. For the simulated data, we simply examine the posterior probability of the root being placed on any particular branch of the tree. This makes sense because the prior probability of all root positions is equal, given that the branches are of equal length. For the empirical data, on the other hand, some branches are a priori more probable to be the root of the tree because they are longer. We calculate this prior probability and report the ratio of the posterior to prior probabilities of the root being placed on any single branch. Ratios  $> 1.0$  mean that a rooting on that branch is more probable after observing the data.

The molecular clock analysis simulated sequences along the tree of Figure 2. The simulations for the outgroup analysis included one additional taxon that was generated from the ancestor (the dot of Fig. 2).

The length of the branch was 0.0, 0.25, 1.0, and  $\infty$  in length. For both the molecular clock and outgroup simulations, the Jukes–Cantor (Jukes and Cantor, 1969) model of DNA substitution was assumed. The instantaneous rate matrix ( $Q$ ) for the Jukes–Cantor model is (before rescaling):

$$Q = \{q_{ij}\} = \begin{pmatrix} - & 1 & 1 & 1 \\ 1 & - & 1 & 1 \\ 1 & 1 & - & 1 \\ 1 & 1 & 1 & - \end{pmatrix} \quad (10)$$

and the stationary distribution is  $\pi = \{1/4, 1/4, 1/4, 1/4\}$ . The nonreversible model analysis also simulated sequences along the tree of Figure 2. However, the model of DNA substitution was generated by taking the instantaneous rate matrix for the Jukes–Cantor model and multiplying each entry by a mean one-gamma-distributed random variable with shape parameter  $\gamma$ :

$$Q = \{q_{ij}\} = \begin{pmatrix} - & r_1 & r_2 & r_3 \\ r_7 & - & r_4 & r_5 \\ r_8 & r_9 & - & r_6 \\ r_{10} & r_{11} & r_{12} & - \end{pmatrix} \quad (11)$$

The expected value of the random variable,  $E(r_i)$ , is 1 and the variance,  $\text{Var}(r_i)$ , is  $\frac{1}{\gamma}$ . As  $\gamma \rightarrow \infty$ , the instantaneous rate matrix converges on the Jukes–Cantor model, and the substitution process becomes more reversible. The substitution process becomes less reversible when  $\gamma$  is small. We simulated data with  $\gamma = 1.0, 10.0, 100.0$ , and  $\infty$ . The resulting index of nonreversibility was  $E(I) = 0.33, 0.12, 0.03$ , and  $0.00$ , respectively.

We also violated the assumptions of the molecular clock analysis when simulating data with a compound Poisson process (Huelsenbeck et al., 2000). A Poisson-distributed number of events of substitution rate change were placed onto the tree. At each event, the substitution rate changed multiplicatively; the old substitution rate was multiplied by a gamma-distributed random variable with parameter  $\alpha_p$  (using the parameterization of the gamma distribution discussed in Huelsenbeck et al., 2000). The parameter of the Poisson distribution was  $\lambda = 10$  and the gamma parameter was varied from 1.0 to  $\infty$ .

For each combination of parameters a total of at least 500 datasets of  $c = 500$  sites were

simulated. Each simulated dataset was analyzed with programs written by J.P.H. and J.P.B. For each simulated dataset, the MCMC analysis was run for 100,000 generations, the first 50,000 being discarded as the burn-in for the chain.

We examined the posterior probability density of the root placement under the three criteria for several DNA sequence datasets. For five of the datasets, the topology and root position can be considered known; the data included representatives from Actinopterygii, Amphibia, Aves, Rodentia, and Primates. The tree is assumed to be (Actinopterygii, (Amphibia, (Aves, (Rodentia, Primates))). Because the Actinopterygii were used as an outgroup, the correct root position for the ingroup tree was along the branch leading to Amphibia. Table 1 summarizes the DNA sequences used in the analysis and their sources.

Posterior probabilities of root positions were calculated with programs that implemented the MCMC procedure. One potential concern of MCMC is that the chains can fail to converge. In this study, we were very cautious in our use of MCMC. First, our problem (finding the posterior probabilities of different root positions) is much simpler than the usual problem of finding the posterior probabilities of trees. In our case, the unrooted topology of the trees is fixed, and the problem is reduced to integrating over root positions, branch lengths, and substitution model parameters. Second, our problems were small compared to those typically analyzed for the phylogeny problem. The largest dataset examined comprised nine taxa (the simulations with the outgroup). Third, we ran the Markov chains much longer than was probably necessary. In all cases that we examined, apparent stationarity of the chain was reached after a few thousand cycles of the MCMC algorithm. However, we ran the chains for a total of 100,000 cycles and discarded all samples made before 50,000 cycles were completed. Finally, there were a few natural checks on our results. We know what the results should look like for an outgroup connected by an infinitely long branch (outgroup criterion) and for a gamma shape parameter of infinity (for the nonreversible simulations). In both cases, the rooting chosen by the MCMC algorithm should be equivalent to randomly choosing a branch. In these

TABLE 1. Five coding DNA sequences for five organisms were sampled in this study.

| Gene                       | <i>c</i> | Taxonomic group | Representative species   | Accession no. |
|----------------------------|----------|-----------------|--------------------------|---------------|
| Albumin<br>(nuclear)       | 1,383    | Actinopterygii  | <i>Salmo salar</i>       | X52397        |
|                            |          | Amphibia        | <i>Xenopus laevis</i>    | M18350        |
|                            |          | Aves            | <i>Gallus gallus</i>     | X60688        |
|                            |          | Rodentia        | <i>Rattus norvegicus</i> | J00698        |
|                            |          | Primates        | <i>Homo sapiens</i>      | L00132        |
| ATPase6<br>(mitochondrial) | 603      | Actinopterygii  | <i>Cyprinus carpio</i>   | X61010        |
|                            |          | Amphibia        | <i>Xenopus laevis</i>    | M10217        |
|                            |          | Aves            | <i>Gallus gallus</i>     | X52392        |
|                            |          | Rodentia        | <i>Mus musculus</i>      | J01420        |
|                            |          | Primates        | <i>Homo sapiens</i>      | J01415        |
| c-myc<br>(nuclear)         | 891      | Actinopterygii  | <i>Salmo gairdneri</i>   | M13048        |
|                            |          | Amphibia        | <i>Xenopus laevis</i>    | M14455        |
|                            |          | Aves            | <i>Gallus gallus</i>     | M20006        |
|                            |          | Rodentia        | <i>Rattus norvegicus</i> | Y00396        |
|                            |          | Primates        | <i>Homo sapiens</i>      | V00568        |
| COI<br>(mitochondrial)     | 1,506    | Actinopterygii  | <i>Cyprinus carpio</i>   | X61010        |
|                            |          | Amphibia        | <i>Xenopus laevis</i>    | M10217        |
|                            |          | Aves            | <i>Gallus gallus</i>     | X52392        |
|                            |          | Rodentia        | <i>Mus musculus</i>      | J01420        |
|                            |          | Primates        | <i>Homo sapiens</i>      | J01415        |
| COII<br>(mitochondrial)    | 648      | Actinopterygii  | <i>Cyprinus carpio</i>   | X61010        |
|                            |          | Amphibia        | <i>Xenopus laevis</i>    | M10217        |
|                            |          | Aves            | <i>Gallus gallus</i>     | X52392        |
|                            |          | Rodentia        | <i>Mus musculus</i>      | J01420        |
|                            |          | Primates        | <i>Homo sapiens</i>      | J01415        |

cases, the MCMC algorithm was equivalent to a random choice of a root.

## RESULTS

Tables 2, 3, and 4 summarize the results of the simulations. The tables show the average posterior probabilities that the root will lie on any of the 13 branches of the eight-taxon tree. The outgroup and molecular clock criteria were able to correctly identify the root of the phylogenetic tree a large proportion of the time. As expected, the outgroup criterion performed best when the length of the branch leading to the outgroup was small (Wheeler, 1990a). When the outgroup was the direct ancestor ( $v = 0$ ) or when the outgroup

branch was short ( $v = 0.25$ ), the criterion was able to correctly identify the root branch; the posterior probability of a root on the correct branch averaged  $\sim 0.98$ . However, as the branch was increased in length, the performance of the outgroup criterion decreased; when the outgroup was a random sequence ( $v = \infty$ ), the root was essentially random on the ingroup topology.

The molecular clock was also able to correctly identify the root of the tree in the simulations. The molecular clock criterion performed particularly well when the rates of substitution were constant across lineages. The ability of the clock criterion to root the tree decreased as the substitution process deviated from the clock assumption. We

TABLE 2. Results for the nonreversible simulations showing the average posterior probabilities of rooting the unrooted tree of  $s = 8$  species on each of the 13 branches. The branches are indicated by the taxon bipartitions. The bipartition {1, 2, 3, 4} is the true (simulated) root of the tree. The branches {1, 2}, {3, 4}, {5, 6}, and {7, 8} are all one removed from the true root. The remaining eight branches are two removed from the correct root. The degree of nonreversibility in the model is related to the parameter  $\gamma$ .

| $\gamma$ | Taxon bipartitions |       |       |       |       |       |           |       |       |       |       |       |       |
|----------|--------------------|-------|-------|-------|-------|-------|-----------|-------|-------|-------|-------|-------|-------|
|          | {1}                | {2}   | {3}   | {4}   | {1,2} | {3,4} | {1,2,3,4} | {5,6} | {7,8} | {5}   | {6}   | {7}   | {8}   |
| 1        | 0.019              | 0.020 | 0.022 | 0.020 | 0.089 | 0.089 | 0.438     | 0.096 | 0.110 | 0.025 | 0.022 | 0.026 | 0.025 |
| 10       | 0.064              | 0.064 | 0.061 | 0.058 | 0.095 | 0.089 | 0.137     | 0.107 | 0.087 | 0.062 | 0.066 | 0.053 | 0.059 |
| 100      | 0.087              | 0.086 | 0.086 | 0.088 | 0.071 | 0.073 | 0.070     | 0.068 | 0.068 | 0.080 | 0.070 | 0.074 | 0.080 |
| $\infty$ | 0.081              | 0.085 | 0.089 | 0.093 | 0.066 | 0.064 | 0.063     | 0.070 | 0.067 | 0.087 | 0.083 | 0.073 | 0.080 |

TABLE 3. Results for the Outgroup simulations showing the average posterior probabilities of rooting the unrooted tree of  $s = 8$  species on each of the 13 branches. The branches are indicated by the taxon bipartitions. The bipartition  $\{1, 2, 3, 4\}$  is the true (simulated) root of the tree. The branches  $\{1, 2\}$ ,  $\{3, 4\}$ ,  $\{5, 6\}$ , and  $\{7, 8\}$  are all one removed from the true root. The remaining eight branches are two removed from the correct root. The length of the branch leading to the outgroup is  $v$ .

| $v$      | Taxon bipartitions |       |       |       |       |       |           |       |       |       |       |       |       |
|----------|--------------------|-------|-------|-------|-------|-------|-----------|-------|-------|-------|-------|-------|-------|
|          | {1}                | {2}   | {3}   | {4}   | {1,2} | {3,4} | {1,2,3,4} | {5,6} | {7,8} | {5}   | {6}   | {7}   | {8}   |
| 0.00     | 0.000              | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.999     | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.25     | 0.000              | 0.000 | 0.000 | 0.000 | 0.004 | 0.004 | 0.984     | 0.002 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1.00     | 0.003              | 0.004 | 0.003 | 0.005 | 0.104 | 0.109 | 0.513     | 0.130 | 0.112 | 0.004 | 0.005 | 0.005 | 0.003 |
| $\infty$ | 0.089              | 0.097 | 0.078 | 0.087 | 0.062 | 0.060 | 0.054     | 0.063 | 0.059 | 0.097 | 0.093 | 0.087 | 0.076 |

simulated sequences in such a way that the rates of substitution changed according to a step function (Huelsenbeck et al., 2000). When the rates of substitution changed very little over the tree ( $\alpha_p = 100$ ), the clock criterion was able to correctly identify the root of the tree, with an average posterior probability of  $\sim 0.75$  across simulations. However, even when the clock criterion was severely violated ( $\alpha_p < 10$ ), the molecular clock criterion did better than a random rooting of the tree; under the most extreme conditions of rate heterogeneity simulated ( $\alpha_p = 1$ ), the correct root was roughly four times as likely to be identified under the clock criterion than under a random rooting criterion.

The nonreversible rooting criterion performed poorly under most circumstances. When the substitution process was highly nonreversible ( $\gamma = 1, I = 0.33$ ), the nonreversible model was able to correctly identify the root of the tree; the posterior probability on the correct root branch was, on average,  $\sim 0.45$ . However, when the model of DNA substitution was nearly reversible ( $\gamma \geq 10, E[I] < 0.3$ ), the method did about as well as a random rooting criterion (the correct branch should have a posterior probability of  $\sim 0.075$  under a random rooting criterion).

Table 5 summarizes the results of the analyses of the five-taxon data sets. For

these data, the root of the unrooted phylogenetic tree is assumed to fall along the branch leading to amphibians: (Amphibia, Aves, (Rodentia, Primates)). The table shows the ratio of the probabilities of rooting the tree along a specific branch when using a rooting criterion to the probability of rooting the tree along that branch if the root were determined randomly:

$$\Lambda_i = \frac{\text{Pr}[\text{Root at branch } i \text{ under criterion } |X]}{\text{Pr}[\text{Root at branch } i \text{ at random } |X]}$$

For  $\Lambda_i = 1.0$ , the rooting criterion is performing no better than rooting the tree at random. For  $\Lambda_i > 1.0$ , the criterion is identifying the  $i$ th branch as the root more often than expected under a random rooting criterion. In all analyses, the partition denoted  $\{0111\}$  is the correct root branch. Figure 3 shows the posterior probability density of rooting the tree of the albumin sequences. The width of the branches is proportional to the posterior probability of a root at that position on the tree. Note that for the albumin sequences, the molecular clock and outgroup criteria correctly identify the root of the tree a high proportion of the time. The nonreversible model, on the other hand, places the root at a large number of positions. Even so, using the nonreversible model, some positions on the

TABLE 4. Results for the molecular clock simulations showing the average posterior probabilities of rooting the unrooted tree of  $s = 8$  species on each of the 13 branches. The branches are indicated by the taxon bipartitions. The bipartition  $\{1, 2, 3, 4\}$  is the true (simulated) root of the tree. The branches  $\{1, 2\}$ ,  $\{3, 4\}$ ,  $\{5, 6\}$ , and  $\{7, 8\}$  are all one removed from the true root. The remaining eight branches are two removed from the correct root. The clock was violated according to the compound Poisson process with parameters  $\lambda = 10$  and  $\alpha_p$ .

| $\alpha_p$ | Taxon bipartitions |       |       |       |       |       |           |       |       |       |       |       |       |
|------------|--------------------|-------|-------|-------|-------|-------|-----------|-------|-------|-------|-------|-------|-------|
|            | {1}                | {2}   | {3}   | {4}   | {1,2} | {3,4} | {1,2,3,4} | {5,6} | {7,8} | {5}   | {6}   | {7}   | {8}   |
| 1          | 0.033              | 0.020 | 0.035 | 0.028 | 0.105 | 0.095 | 0.347     | 0.117 | 0.116 | 0.027 | 0.022 | 0.030 | 0.026 |
| 10         | 0.012              | 0.012 | 0.011 | 0.004 | 0.159 | 0.148 | 0.340     | 0.148 | 0.136 | 0.003 | 0.012 | 0.006 | 0.011 |
| 100        | 0.000              | 0.000 | 0.000 | 0.000 | 0.062 | 0.058 | 0.765     | 0.061 | 0.055 | 0.000 | 0.000 | 0.000 | 0.000 |
| $\infty$   | 0.000              | 0.000 | 0.000 | 0.000 | 0.006 | 0.009 | 0.972     | 0.006 | 0.007 | 0.000 | 0.000 | 0.000 | 0.000 |

TABLE 5. The ratio of the probabilities of rooting the tree on a branch (indicated by the partition) when using one of the three criteria to the probability of rooting the tree at that branch at random.

| Method          | Partition   | Gene    |         |              |      |      |
|-----------------|-------------|---------|---------|--------------|------|------|
|                 |             | Albumin | ATPase6 | <i>c-myc</i> | COI  | COII |
| Molecular clock | <b>0111</b> | 2.44    | 0.00    | 2.48         | 1.72 | 0.07 |
|                 | 0100        | 0.00    | 0.00    | 0.01         | 0.35 | 0.29 |
|                 | 0010        | 0.00    | 0.00    | 0.00         | 0.01 | 0.05 |
|                 | 0001        | 0.00    | 0.00    | 0.00         | 0.01 | 0.33 |
|                 | 0011        | 0.00    | 2.34    | 0.03         | 7.69 | 8.08 |
| Outgroup        | <b>0111</b> | 2.16    | 2.05    | 2.20         | 3.31 | 4.62 |
|                 | 0100        | 0.40    | 2.27    | 0.11         | 0.01 | 0.12 |
|                 | 0010        | 0.00    | 0.00    | 0.00         | 0.00 | 0.00 |
|                 | 0001        | 0.00    | 0.00    | 0.00         | 0.00 | 0.00 |
|                 | 0011        | 0.06    | 0.46    | 0.27         | 0.00 | 1.73 |
| Nonreversible   | <b>0111</b> | 0.83    | 0.10    | 0.15         | 0.46 | 0.41 |
|                 | 0100        | 1.03    | 0.16    | 0.10         | 2.02 | 2.45 |
|                 | 0010        | 0.31    | 1.35    | 7.29         | 0.57 | 0.51 |
|                 | 0001        | 1.34    | 1.29    | 0.25         | 0.88 | 0.41 |
|                 | 0011        | 2.08    | 1.48    | 0.21         | 0.78 | 0.66 |

tree have a low posterior probability of being the root. For example, the tips of the branches leading to the frog and to the mouse have a low probability of being the root under the nonreversible model.

The results of the analyses of the five-species sequence data are consistent with the simulation results (Table 5). The molecular clock and outgroup criteria are better able to correctly identify the root of the tree than the nonreversible model. The outgroup criterion consistently outperformed the other rooting criteria. The ratio  $\Lambda_{\{0111\}}$  is  $>2.0$  for all five genes, indicating that the outgroup criterion favors the correct root with twice the probability it would a random root. The molecular clock criterion worked well for three of the genes (albumin, *c-myc*, and cytochrome oxidase I [COI]); for these genes, the correct root was roughly twice as likely to be identified under the molecular clock criterion than under the random rooting criterion. For two of the genes (ATPase6 and COII), the molecular clock criterion failed to correctly identify the root. A likelihood-ratio test of the molecular clock hypothesis indicates that the clock would be rejected for all of the genes except *c-myc* (see Table 6; all likelihoods were calculated under the GTR +  $\Gamma$  model of DNA substitution by using PAUP\*; Swofford, 1999). For both ATPase6 and COII, the molecular clock was rejected, suggesting that failure of the data to obey the molecular clock assumption might be partially responsible for the poor performance of the molecular clock rooting criterion. However, in two of the other genes, the molecular clock was also

rejected and yet the criterion worked well for finding the correct root position of the tree.

The nonreversible substitution model performed poorly as a rooting criterion for the five-species datasets. The method performed worse than a random rooting of the tree for all five genes. Interestingly, the index of nonreversibility that was estimated for these genes was in the range of the simulated values (above). The index of nonreversibility for the five genes is summarized in Table 6.

## DISCUSSION

Of the three criteria examined for rooting a phylogenetic tree—the outgroup, molecular clock, and nonreversible model criteria—the outgroup criterion was consistently able to identify the root. As expected, only when the outgroup was very distantly related to the ingroup taxa did the criterion fail (Wheeler, 1990a). The molecular clock criterion might also be usefully applied to identify the root of a phylogenetic tree; the method seems somewhat robust to violation of the clock assumption, suggesting that the root of phylogenetic trees can be correctly identified even if the substitution process is not strictly clocklike.

On the basis of an analysis of two genes, Yang (1994a) suggested that the nonreversible model would have little utility for rooting phylogenetic trees. The simulations and analyses of five genes in this paper confirm that the nonreversible rooting criterion does a poor job of distinguishing among possible rooting positions. However, perhaps this criterion will be useful for more

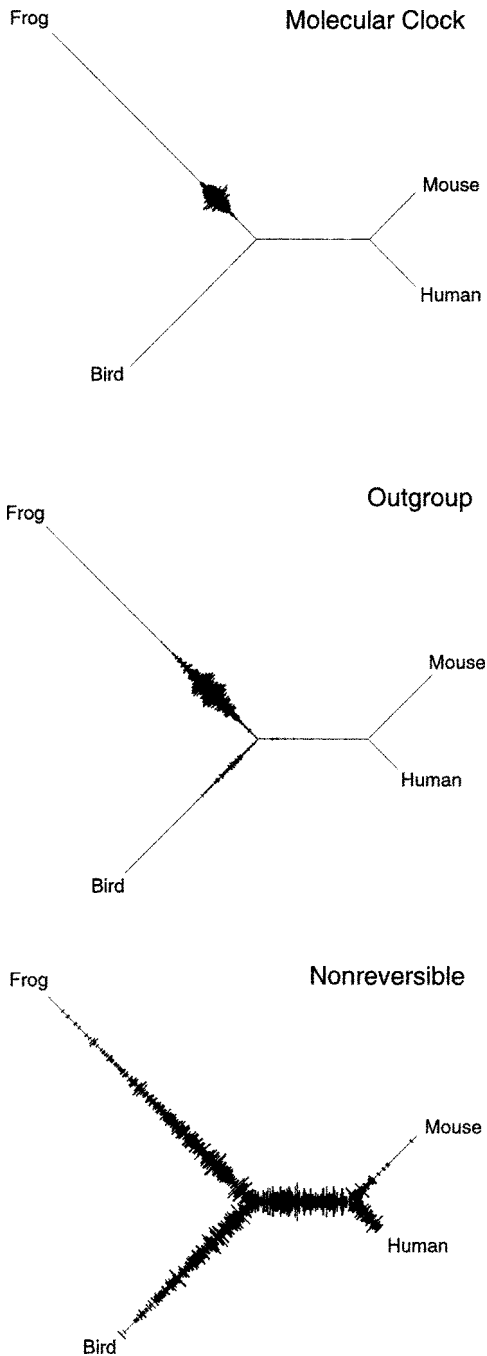


FIGURE 3. The posterior probabilities of rooting the albumin tree if using the three criteria discussed in this paper. The width of the branch is proportional to the posterior probability that the root is at that point. The lengths of the branches are the mean of the posterior density.

distantly related species (such as the relationships among the main lineages of life). In fact, Yang and Roberts (1995) found that for very divergent sequences with different base frequencies, a nonreversible model could be useful for identifying the root. This suggests that a well-supported phylogeny with more divergent sequences might be useful for examining the practical utility of the nonreversible criterion for rooting trees.

We described a Bayesian method for inferring the root of a phylogenetic tree. This method has two advantages: (1) It can determine the probabilities of a root falling at different positions on a tree and (2) it is relatively fast compared with maximum likelihood. The analysis under the nonreversible model of DNA substitution, for example, was much faster than the equivalent analysis using maximum likelihood. Maximum likelihood and Bayesian inference both use the likelihood function, but in different ways. Maximum likelihood estimates parameters by maximizing the likelihood function. Bayesian inference, on the other hand, uses the likelihood function and the scientist's prior beliefs about a parameter to calculate the posterior probability of the parameter. Inferences are then based on the posterior probability distribution of the parameter. Algorithmically, it seems easier to calculate posterior probabilities by using MCMC than to maximize likelihoods by using standard hill-climbing algorithms such as those used in PAUP\* (Swofford, 1999). The difference in speed is especially apparent for complicated models.

One potential concern of a Bayesian analysis is the use of prior probability distributions on parameters. In some cases, such as reconstructing ancestral states for a single character (Schultz and Churchill, 1999), the choice of prior can make a difference in the analysis. However, for most phylogenetic problems, the prior appears to be overwhelmed by the character data. In this analysis, we assumed a uniform prior on the root position; this means, first of all, that the posterior probability distribution will look much like the likelihood surface and, moreover, that we are not biasing the potential root position of the tree. In some cases, however, the investigator may have strong opinions about the root placement of a tree and want to bias the root position. In such cases, the Bayesian analysis can be modified to take into account these

TABLE 6. Statistics for the five data sets examined in this study.  $I$ , index of reversibility;  $L$ , tree length (in terms of expected number of substitutions per site), calculated under the GTR +  $\Gamma$  model of DNA substitution;  $-\log_e L_C$  and  $-\log_e L_{NC}$ , log likelihoods calculated under the clock constraint and without the clock constraint, respectively (again, calculated under the GTR +  $\Gamma$  model);  $-2 \log_e \Lambda$ , likelihood ratio test statistic for a test of the molecular clock hypothesis; and  $P$  is the  $P$ -value for the test of the molecular clock hypothesis. For all but *c-myc*, the molecular clock hypothesis can be rejected.

| Data         | $I$   | $L$  | $-\log_e L_C$ | $-\log_e L_{NC}$ | $-2 \log_e \Lambda$ | $P$     |
|--------------|-------|------|---------------|------------------|---------------------|---------|
| Albumin      | 0.182 | 1.73 | -5894.61      | -5890.46         | 8.30                | 0.016   |
| ATPase6      | 0.372 | 2.40 | -2332.40      | -2320.30         | 24.20               | <0.0001 |
| <i>c-myc</i> | 0.190 | 0.62 | -2794.87      | -2792.72         | 4.30                | 0.117   |
| COI          | 0.220 | 1.12 | -4991.20      | -4986.11         | 10.18               | 0.006   |
| COII         | 0.253 | 1.33 | -2378.35      | -2373.02         | 10.66               | <0.005  |

prior beliefs about the root. Information that might cause a biologist to modify his or her beliefs about the root position of a tree include the results from other genes or stratigraphic information.

The Bayesian method for inferring the root of a phylogenetic tree, introduced here, is compatible with other methods for rooting trees that were not examined in this paper. For example, the root of a phylogenetic tree can also be determined by using temporal data (Huelsenbeck, 1994), the midpoint criterion, and gene duplications (Iwabe et al., 1989; Mathews and Donoghue, 1999). These methods, in fact, may prove more powerful than the methods examined in this paper for rooting a tree of distantly related species.

#### ACKNOWLEDGMENTS

This paper was improved through the comments of Z. Yang and an anonymous reviewer. This work was supported by National Science Foundation grants DEB-0075406 and MCB-0075404 awarded to J.P.H.

#### REFERENCES

- FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17:368-376.
- GREEN, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82:711-732.
- HASEGAWA, M., H. KISHINO, AND T. YANO. 1985. Dating the human-ape split by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22:160-174.
- HASEGAWA, M., T. YANO, AND H. KISHINO. 1984. A new molecular clock of mitochondrial DNA and the evolution of Hominoids. *Proc. Jpn. Acad. Ser. B* 60:95-98.
- HASTINGS, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97-109.
- HUELSENBECK, J. P. 1994. Measuring and testing the fit of the stratigraphic record to phylogenetic trees. *Paleobiology* 20:470-483.
- HUELSENBECK, J. P., B. LARGET, AND D. L. SWOFFORD. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* 154:1879-1892.
- IWABE, N., K.-I. KUMA, M. HASEGAWA, S. OSAWA, AND T. MIYATA. 1989. Evolutionary relationship of archaeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. USA* 86:9355-9359.
- JUKES, T., AND C. CANTOR. 1969. Evolution of protein molecules. Pages 21-132 in *Mammalian protein metabolism* (H. Munro, ed.). Academic Press, New York.
- LARGET, B., AND D. SIMON. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16:750-759.
- LI, S. 1996. Phylogenetic tree construction using Markov chain Monte Carlo. Ph.D. dissertation, Ohio State Univ., Columbus.
- MADDISON, W. P., M. J. DONOGHUE, AND D. R. MADDISON. 1984. Outgroup analysis and parsimony. *Syst. Zool.* 33:83-103.
- MATHEWS, S., AND M. J. DONOGHUE. 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* 286:947-950.
- MAU, B. 1996. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Ph.D. Dissertation, Univ. of Wisconsin, Madison.
- MAU, B., AND M. NEWTON. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J. Comput. Graph. Stat.* 6:122-131.
- MAU, B., M. NEWTON, AND B. LARGET. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* 55:1-12.
- METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, AND E. TELLER. 1953. Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21:1087-1091.
- NEWTON, M., B. MAU, AND B. LARGET. 1999. Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. In *Statistics in molecular biology* (F. Seillier-Moseiwitch, T. P. Speed, and M. Waterman, eds.). Monograph Series of the Institute of Mathematical Statistics.
- RANNALA, B., AND Z. YANG. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J. Mol. Evol.* 43:304-311.
- SANKOFF, D. 1975. Minimal mutation trees of sequences. *SIAM J. Appl. Math.* 28:35-42.
- SANKOFF, D., AND P. ROUSSEAU. 1975. Locating the vertices of a Steiner tree in arbitrary space. *Math. Prog.* 9:240-246.

- SCHULTZ, T. R., AND G. A. CHURCHILL. 1999. The role of subjectivity in reconstructing ancestral character states: A Bayesian approach to unknown rates, states, and transformation asymmetries. *Syst. Biol.* 48: 651–664.
- SWOFFORD, D. L. 1999. PAUP\*. Phylogenetic analysis using parsimony (\*and other methods), version 4. Sinauer Associates, Sunderland, Massachusetts.
- SWOFFORD, D., G. OLSEN, P. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–511 in *Molecular systematics*, 2nd edition (D. Hillis, C. Moritz, and B. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.
- TAVARÉ, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect. Math. Life Sci.* 178:57–86.
- WHEELER, W. C. 1990a. Nucleic acid sequence phylogeny and random outgroups. *Cladistics* 6:363–368.
- WHEELER, W. C. 1990b. Combinatorial weights in phylogenetic analysis: A statistical parsimony procedure. *Cladistics* 6:269–275.
- WILLIAMS, P. L., AND W. M. FITCH. 1989. Finding the minimal change in a given tree. Pages 453–470 in *The hierarchy of life* (B. Fernholm, K. Bremer, and H. Jörnvall, eds.). Elsevier, Amsterdam.
- 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. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39:105–111.
- 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., AND B. RANNALA. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte carlo method. *Mol. Biol. Evol.* 14:717–724.
- YANG, Z., AND D. ROBERTS. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.* 12:451–458.

*Received 6 June 2000; accepted 16 August 2000*

*Associate Editor: R. Olmstead*