Statistical Properties of Bootstrap Estimation of Phylogenetic Variability from Nucleotide Sequences. I. Four Taxa with a Molecular Clock[1]

*Andrey Zharkikh**,† *and Wen-Hsiung Li**

*Center for Demographic and Population Genetics, University of Texas, Houston; and †Institute of Cytology and Genetics, Novosibirsk

The statistical properties of sample estimation and bootstrap estimation of phylogenetic variability from a sample of nucleotide sequences are studied by using model trees of three taxa with an outgroup and by assuming a constant rate of nucleotide substitution. The maximum-parsimony method of tree reconstruction is used. An analytic formula is derived for estimating the sequence length that is required if P, the probability of obtaining the true tree from the sampled sequences, is to be equal to or higher than a given value. Bootstrap estimation is formulated as a two-step sampling procedure: (1) sampling of sequences from the evolutionary process and (2) resampling of the original sequence sample. The probability that a bootstrap resampling of an original sequence sample will support the true tree is found to depend on the model tree, the sequence length, and the probability that a randomly chosen nucleotide site is an informative site. When a trifurcating tree is used as the model tree, the probability that one of the three bifurcating trees will appear in ≥95% of the bootstrap replicates is <5%, even if the number of bootstrap replicates is only 50; therefore, the probability of accepting an erroneous tree as the true tree is <5% if that tree appears in ≥95% of the bootstrap replicates and if more than 50 bootstrap replications are conducted. However, if a particular bifurcating tree is observed in, say, <75% of the bootstrap replicates, then it cannot be claimed to be better than the trifurcating tree even if ≥1,000 bootstrap replications are conducted. When a bifurcating tree is used as the model tree, the bootstrap approach tends to overestimate P when the sequences are very short, but it tends to underestimate that probability when the sequences are long. Moreover, simulation results show that, if a tree is accepted as the true tree only if it has appeared in ≥95% of the bootstrap replicates, then the probability of failing to accept any bifurcating tree can be as large as 58% even when P = 95%, i.e., even when 95% of the samples from the evolutionary process will support the true tree. Thus, if the rate-constancy assumption holds, bootstrapping is a conservative approach for estimating the reliability of an inferred phylogeny for four taxa.

Introduction

The rapid accumulation of DNA sequence data has stimulated much activity in the reconstruction of phylogenetic relationships among organisms. It has also stimulated much interest in the development of methods for tree reconstruction and for evaluating the statistical confidence of an inferred phylogeny. Presently, among the statistical

1. Key words: phylogenetic reconstruction, sample estimation, bootstrap estimation, confidence level, sequence length, bootstrap replicates.

Address for correspondence and reprints: Wen-Hsiung Li, Center for Demographic and Population Genetics, University of Texas, P. O. Box 20334, Houston, Texas 77225.

Mol. Biol. Evol. 9(6):1119–1147. 1992.

0737-4038/92/0906-0009$02.00

methods for evaluating the reliability of inferred phylogenies [see the reviews by Felsenstein (1988) and Li and Gouy (1991)], the bootstrap method (Felsenstein 1985) is the simplest and the most frequently used method when the number of taxa under study is more than four. However, the statistical properties of this approach in the context of phylogenetic reconstruction have not been well studied, though its theoretical foundation in terms of general statistics has been examined thoroughly (Effron 1982). The present paper explores properties of bootstrap estimates based on the maximum-parsimony method of tree reconstruction. Recently, Hillis and Bull (accepted) have also studied this problem.

For simplicity we consider the case of three taxa with one outgroup and assume a constant rate for the evolution of nucleotide sequences. This simple case can be treated analytically, making it easier to clarify some of the conceptual aspects of bootstrap estimation. Moreover, it allows a close examination of the statistical properties of the distribution of informative sites in a sample of sequences, a study that was initiated by Saitou and Nei (1986), and our analytic results turn out to be very useful for investigating the statistical properties of bootstrap estimation. The simple case also makes it easier to study, theoretically, both bootstrap estimation of the confidence level of an inferred phylogeny and the dependence of the confidence level on both the amount of data under study and the number of bootstrap replications. Our ultimate aims are to know whether the bootstrap approach tends to overestimate or underestimate the confidence level of an inferred phylogeny and the probability of accepting an erroneous tree as the true tree.

Approaches and Results

To help readers understand the analysis to be given below, we explain here the approaches to be used. We also summarize the main results so that a reader can understand the essence of the present paper without going through the mathematical analysis.

We use a simple model tree in which there are three taxa with one outgroup. The three possible rooted bifurcating trees (I, II, and III) are shown in Figure 1a–c. We assume that the first tree (tree I) is the true tree and that the branching dates for the outgroup, species 3, and species 2 are, respectively, T_1, T_2, and T_3 before the present. The trifurcating tree (fig. 1d) is the best representation of the species phylogeny when we cannot make a decision about the branching order. We use either tree I or the trifurcating tree as the model tree in our analysis.

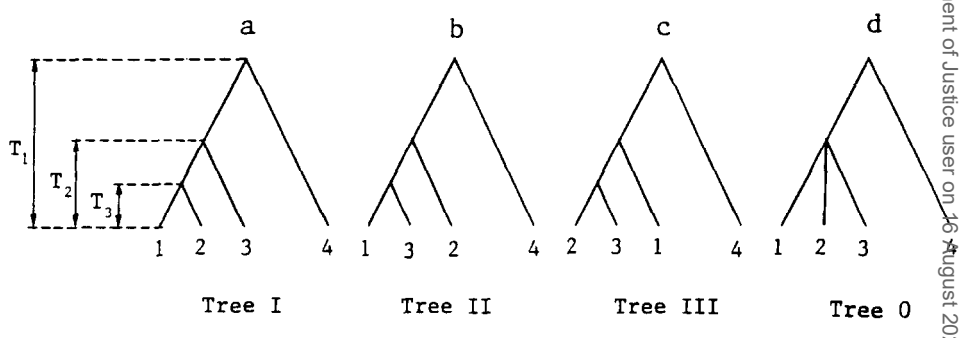

FIG. 1.—Three possible bifurcating trees (a–c) and the trifurcating tree (d), for three species with one outgroup. Tree I is assumed to be the true tree in the bifurcating models.

First, for a given model tree we study the evolution of a nucleotide sequence along each branch of the tree and the probability of having a particular configuration pattern of the nucleotides at the tips of the tree. Under the maximum-parsimony method, which is the tree-reconstruction method to be used in the present study, a configuration pattern is said to be informative if it is useful for distinguishing among the three possible bifurcating trees. An informative site is said to support tree i ($i = I, II,$ or III) if the number of nucleotide substitutions required to explain the observed configuration at that site is smaller under tree i than under either of the two other possible bifurcating trees. We derive a formula for the probability (p_i) that a randomly chosen site will support tree i. The probabilities p_I, p_{II}, and p_{III} are the basic quantities in subsequent analysis.

Next, we study the statistical properties of the distribution of the three types of informative sites in a sample of sequences of length N. We then derive an analytic formula for estimating the sequence length that is required if the probability of obtaining the true tree from the sampled sequences is to be equal to or higher than a given value, e.g., 95%. The analytic results obtained in this section are useful for studying the bootstrap technique.

Third, we use either tree I or the trifurcating tree in figure 1 as the model tree and study the bootstrap estimation of P_I, which is the probability of obtaining tree I from a random sample of sequences of length N. The bootstrap estimation is formulated as a two-step sampling procedure: (i) A random sample of sequences is taken from the evolutionary process. (ii) The sites of the sequences in the original sample are resampled with replacement (i.e., bootstrapped), and a tree is reconstructed from the resampled data. The second step is repeated N_b times, and the proportion of the bootstrap replicates that support tree I is taken as an estimate of P_I. Symbolically, the two-step procedure can be represented as

$$\begin{array}{ccccc} (p_I, p_{II}, p_{III}) & \rightarrow & (p_I^*, p_{II}^*, p_{III}^*) & \rightarrow & (p_I^{**}, p_{II}^{**}, p_{III}^{**}) \\ \downarrow & & \downarrow & & \downarrow \\ P_I & & P_I^* & & P_I^{**} \end{array}$$

where p_I, p_{II}, and p_{III} are the underlying probabilities of informative sites supporting tree I, tree II, and tree III, respectively; p_I^*, p_{II}^*, and p_{III}^* are the corresponding proportions of informative sites in a random sample of sequences from the evolutionary process and are considered as the underlying probabilities of informative sites for bootstrap resampling; and p_I^{**}, p_{II}^{**}, and p_{III}^{**} are the proportions of informative sites in a sample bootstrapped from the original sample. The probabilities p_i ($i = I, II, III$) determine the underlying probability P_I that a random sample of sequences from the evolutionary process will support tree I. In the same manner, the proportions p_i^* determine the probability P_I^* that a bootstrap resampling of an original sample will support tree I. The proportions p_i^{**} determine the most parsimonious tree in a bootstrap replicate, and P_I^{**} denotes the proportion of the bootstrap replicates in which tree I is chosen. Since P_I^* can be regarded as a random variable, P_I^{**} is actually a compound random variable (Johnson and Kotz 1969, p. 183). This formulation clearly shows that the variance of a bootstrap estimate consists of two components: the first one arises from sampling of sequence data from the evolutionary process, and the second arises from bootstrap resampling. The second component can be reduced to 0 by increasing N_b to infinity, but the first component is independent of bootstrap

resampling and can be reduced only by increasing the sequence length N. In order to understand the statistical properties of bootstrap estimation of P_I, we study the distribution of P_I^* by using, as the model tree, either tree I or the trifurcating tree in figure 1.

Fourth, since in practice we do not know a priori which tree is the true tree, we assume that the tree inferred from the sequence sample is the true tree. Denote this tree by X. In analogy with the preceding situation, let P_X^* be the probability that a bootstrap resampling of the original sample will support tree X and let P_X^{**} be the proportion of bootstrap replicates that support tree X. Note that, since the tree inferred can vary from sample to sample, X can be tree I, tree II, or tree III. For this reason, $P_X^* \geq P_I^*$ and $P_X^{**} \geq P_I^{**}$. As in the case of P_I^*, we study the distribution of P_X^* by using, as the model tree, tree I or the trifurcating tree.

Finally, and most important, we study whether P_X^{**} can be taken as the confidence level that tree X is the true tree. We show that, if $P_X^{**} \geq 95\%$, then the probability that tree X is an erroneous tree is <5%, even if N_b is as small as 50. In general, if $P_X^{**} \geq 80\%$ and $N_b \geq 100$, then considerable ($\geq$80%) confidence can be given to tree X as the true tree. However, if $P_X^{**} \leq 75\%$, then little confidence can be given to tree X, because it cannot be claimed to be better than the trifurcating tree. Further, we show that, if $P_I \sim \leq 78\%$, then P_X^{**} tends to overestimate P_I but that, if $P_I > 78\%$, then P_X^{**} actually tends to underestimate P_I. Indeed, when $P_I = 95.2\%$, the expected value of P_X^{**} is only 86.8% and the probability that $P_X^{**} \geq 95\%$ is only 42.0%. Even when $P_I = 99.6\%$, so that almost every sample from the evolutionary process will support tree I, the probability that $P_X^{**} \geq 95\%$ is still only 76.3%, though the expected value of P_X^{**} increases to 95.9%. Thus, the sequence length required for $P_X^{**} \geq 95\%$ is usually several times longer than that required for $P_I \geq 95\%$.

The above conclusions are obtained under the assumption of rate constancy. Under unequal rates of evolution among lineages, the maximum-parsimony method can be positively misleading (Felsenstein 1978), and so some of the above conclusions may not hold (Hillis and Bull, accepted; Zharkikh and Li, accepted).

Evolution of Nucleotides and Informative Sites

In this section we describe the model of nucleotide substitution and the method of phylogenetic reconstruction to be used in this study. We use Kimura's (1980) two-parameter model of nucleotide substitution, in which the rate of transition and the rate of each type of transversion are α and β substitutions per site per year, respectively; transitions are changes between either A and G or T and C, while all other types of changes are transversions. Under this model, the total rate of substitution per site is $\mu = \alpha + 2\beta$, because at each site there are one type of transition and two types of transversion.

Let us replace the parameters α and β in this model by their ratio $r = \alpha/\beta$ and by the total rate of substitution per site $\mu = \alpha + 2\beta$. Then, $\alpha = \mu r/(r+2)$ and $\beta = \mu/(r+2)$. For each time interval t, we can define the probabilities that the nucleotides at the two ends of this interval are X and Y, respectively (Li 1986):

$$\text{Prob}(X \rightarrow Y; t, \mu, r) = \tfrac{1}{4} + \tfrac{1}{4}e^{-4t\mu/(r+2)} - \tfrac{1}{2}e^{-2t\mu(r+1)/(r+2)}, \quad (1)$$

if $X \rightarrow Y$ is a transition;

$$\text{Prob}(X \rightarrow Y;\, t,\, \mu,\, r) = \tfrac{1}{4} - \tfrac{1}{4}e^{-4t\mu/(r+2)}, \tag{2}$$

if $X \rightarrow Y$ is a specific type of transversion; and

$$\text{Prob}(X = Y;\, t,\, \mu,\, r) = \tfrac{1}{4} + \tfrac{1}{4}e^{-4t\mu/(r+2)} + \tfrac{1}{2}e^{-2t\mu(r+1)/(r+2)}. \tag{3}$$

Note that we can replace both parameter t and parameter μ in these equations by the expected number of substitutions per site (M_i) for branch i in figure 2:

$$M_i = t_i\mu_i \qquad (i = 1, \ldots, 5). \tag{4}$$

So, the above probabilities can be redefined as functions of only two parameters, M and r: $\text{Prob}(X \rightarrow Y;\, M,\, r)$.

Under the assumption of rate constancy, $\mu_i = \mu$ for all i, and all time spans in figure 2 and the corresponding expected numbers of substitutions will be defined as follows:

$$\begin{aligned} t_1 &= t_2 = T_3, & M_1 &= M_2 = \mu T_3; \\ t_3 &= T_2, & M_3 &= \mu T_2; \\ t_4 &= 2T_1 - T_2, & M_4 &= \mu(2T_1 - T_2); \\ t_5 &= T_2 - T_3, & M_5 &= \mu(T_2 - T_3). \end{aligned} \tag{5}$$

Let p_{X_i} be the probability of observing nucleotide X_i (A, T, G, or C) at a given site at node i (fig. 2). Then, the probability of observing nucleotides X_1, X_2, X_3, and X_4 at nodes 1, 2, 3, and 4, respectively, is (Saitou 1988)

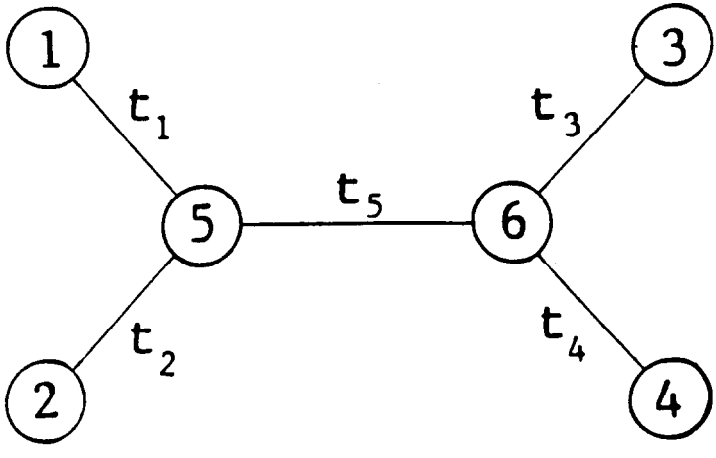

FIG. 2.—Unrooted model tree for four sequences. The branch lengths can be given either as the time spans (t_i, $i = 1, \ldots, 5$), if we assume a constant rate, or as the expected numbers of substitutions ($M_i = t_i\mu_i$) for the case of unequal evolutionary rates μ_i.

$$\text{Prob}(X_1, X_2, X_3, X_4) = \sum_{X_5} \sum_{X_6} p_{X_4} \text{Prob}(X_4 \rightarrow X_6; M_4, r)$$
$$\times \text{Prob}(X_6 \rightarrow X_3; M_3, r) \text{Prob}(X_6 \rightarrow X_5; M_5, r) \quad (6)$$
$$\times \text{Prob}(X_5 \rightarrow X_2; M_2, r) \text{Prob}(X_5 \rightarrow X_1; M_1, r) .$$

The pattern (X_1, X_2, X_3, X_4) is said to be informative if it helps to distinguish between different tree topologies. Different tree-making methods have different informative-site definitions. Some of them have been listed by Li et al. (1987). For example, in the case of the maximum-parsimony method, the pattern (X_1, X_2, X_3, X_4) is informative; that is, it supports one of the three bifurcating trees in figure 1:

$$\begin{array}{llllll} \text{tree I,} & \text{if} & X_1 = X_2, & X_2 \neq X_3, & \text{and} & X_3 = X_4 ; \\ \text{tree II,} & \text{if} & X_1 = X_3, & X_2 \neq X_3, & \text{and} & X_2 = X_4 ; \\ \text{tree III,} & \text{if} & X_1 = X_4, & X_2 \neq X_4, & \text{and} & X_2 = X_3 . \end{array} \quad (7)$$

For example, if $X_1 = X_2 =$ A and $X_3 = X_4 =$ G, then the site supports tree I. Using the above formulas, we can calculate the probability p_i that a randomly chosen site is an informative site supporting tree i, $i = I$, II, or III:

$$p_I = \sum_{X_4} \sum_{X_1 \neq X_4} \text{Prob}(X_1, X_1, X_4, X_4) ; \quad (8)$$

$$p_{II} = \sum_{X_4} \sum_{X_1 \neq X_4} \text{Prob}(X_1, X_4, X_1, X_4) ; \quad (9)$$

$$p_{III} = \sum_{X_2} \sum_{X_1 \neq X_2} \text{Prob}(X_1, X_2, X_2, X_1) ; \quad (10)$$

where, for example, the summation $\sum_{X_4} \sum_{X_1 \neq X_4}$ is over all possible nucleotide configurations in which $X_1 \neq X_4$, $X_1 = X_2$, and $X_3 = X_4$. Note that p_i is also the expected proportion of informative sites supporting tree i when a sample of sequences is taken from the four species.

The maximum-parsimony method is to choose the most parsimonious tree, i.e., the tree with the largest number of supporting sites. Other methods of tree reconstruction are based on more complicated scores (see Li et al. 1987; Nei 1987). Some of them (e.g., the evolutionary-parsimony method) are linear combinations of the numbers of informative sites. Presumably, such methods, in their statistical properties, share some similarities with the maximum-parsimony method. In this paper we shall consider only the statistical properties of the maximum-parsimony method. Other methods will be considered elsewhere.

Sample Estimation

In this section we consider the statistical properties of the distribution of the three types of informative sites in a sample of sequences of length N. The main purpose is to study the relationship between N and the probability of obtaining the true tree from the sequence sample.

For a given set of aligned sequences of length N, we can count the number of

informative sites, N_I, N_{II}, and N_{III}, supporting trees I, II, and III, respectively, and can calculate their sample proportions, $p_I^* = N_I/N$, $p_{II}^* = N_{II}/N$, and $p_{III}^* = N_{III}/N$. Under the assumption that each nucleotide site evolves independently and with the same rate of substitution, each of the numbers N_i, $i = I, II,$ or III, has a binomial distribution, and hence the mean and the variance of p_i^* are

$$\mathrm{E}(p_i^*) = p_i \qquad \text{and} \qquad \mathrm{Var}(p_i^*) = \frac{p_i(1-p_i)}{N} . \tag{11}$$

The observed proportions p_I^*, p_{II}^*, and p_{III}^* are said to support tree I, if $p_I^* > \max(p_{II}^*, p_{III}^*)$. For a sample of N sites, the probability of obtaining tree I, P_I, is

$$P_I = \mathrm{Prob}(p_I^* > \max(p_{II}^*, p_{III}^*)) . \tag{12}$$

When N is small, P_I can be obtained from the multinomial expansion of $(p_0+p_I+p_{II}+p_{III})^N$ (see Saitou and Nei 1986); $p_0 = 1-p_I-p_{II}-p_{III}$ is the proportion of noninformative sites. When N is large, the following approach is computationally much simpler. Define the difference function

$$\gamma_I^* = p_I^* - \max(p_{II}^*, p_{III}^*) . \tag{13}$$

If $\gamma_I^* > 0$, then the given set of sequences supports tree I. Therefore,

$$P_I = \mathrm{Prob}(\gamma_I^* > 0) = 1 - \mathrm{Prob}(\gamma_I^* \le 0) . \tag{14}$$

Expression (13) can be rewritten as follows:

$$\gamma_I^* = p_I^* - \left(\frac{p_{II}^* + p_{III}^*}{2} + \frac{|p_{II}^* - p_{III}^*|}{2} \right) . \tag{15}$$

Under the assumption of rate constancy and the assumption that tree I is the true tree, we have $p_{II} = p_{III}$ and $p_I > p_{II}$. So, the expectation of the first two terms in equation (15) is $\Delta p_I = p_I - p_{II}$. The last term of the equation involves the absolute difference $|x-y|$, the expected value of which is known as *Gini's mean difference* (Johnson and Kotz 1970, p. 67). If x and y are normally distributed with the same mean and with the variance σ^2, then $\mathrm{E}(|x-y|) \approx 2\sigma/\sqrt{\pi}$. When $N \gg 1/p_{II}$, we can use the normal approximation to the distribution of p_i^*. Note that the covariance between p_i^* and p_j^*, $i \neq j$, is $-p_i p_j/N$ (see Johnson and Kotz 1969, p. 284). Therefore, if $N \gg 1/p_{II}$, i.e., $1/N \ll p_{II}$, the covariances between p_I^*, p_{II}^*, and p_{III}^* are of the order of $p_I p_{II}/N$ and can be neglected; note that p_I, p_{II}, and p_{III} are usually much smaller than 1. We then obtain the following approximations for the mean and the variance of γ_I^*:

$$\mathrm{E}(\gamma_I^*) \approx \Delta p_I - \sqrt{\frac{\mathrm{Var}(p_{II}^*)}{\pi}} = \Delta p_I - \sqrt{\frac{p_{II}(1-p_{II})}{N\pi}} ; \tag{16}$$

and

$$\begin{aligned}\mathrm{Var}(\gamma_I^*) &\approx \mathrm{Var}(p_I^*)+\mathrm{Var}\left(\frac{p_{II}^*+p_{III}^*}{2}\right)+\mathrm{Var}\left(\frac{|p_{II}^*-p_{III}^*|}{2}\right)\\ &\approx \frac{p_I(1-p_I)}{N}+\frac{p_{II}(1-p_{II})}{N}\left(1-\frac{1}{\pi}\right).\end{aligned} \tag{17}$$

The case of $\Delta p_I = 0$ in equation (16) corresponds to the trifurcating model tree (fig. 1d). In figure 3, the plots for the probability density function of γ_I^* for different N are shown. The dashed line in the middle of each distribution indicates the mean value, $E(\gamma_I^*)$, which is always negative for this tree. As N increases, $E(\gamma_I^*)$ approaches 0 (fig. 3), and the width of the distribution of γ_I^* decreases in a manner such that the area for the right part of the distribution (i.e., $\gamma_I^*>0$) is approximately constant. The relative proportions of P_I, P_{II}, and P_{III} are equal to ⅓ and independent of N. Because of the nonzero probability of the equality $p_I^* = \max(p_{II}^*, p_{III}^*)$, the absolute value of P_I is actually <⅓. However, P_I approaches ⅓, as $N \to \infty$.

For $\Delta p_I > 0$ (tree I as the model tree), the picture is quite different (fig. 4). When $N < N_{0.5} = p_{II}(1-p_{II})/\pi(\Delta p_I)^2$, formula (16) implies that the expectation of γ_I^* is negative (fig. 4a). For $N = N_{0.5}$, $E(\gamma_I^*) = 0$ (fig. 4b). In this case, ~50% of the distribution of γ_I^* lies in the region of positive γ_I^*, i.e., $P_I \approx 0.5$. When $N > N_{0.5}$, $E(\gamma_I^*)$ is positive (fig. 4c), and P_I increases with N, approaching 1 as $N \to \infty$.

Thus, if the bifurcating tree (tree I) represents the true phylogeny, then, by increasing the sequence length, we can reach any given proportion P_I. To estimate the sequence length required for obtaining tree I with a given probability $\hat{P}_I$, let us construct a new variable $\beta = \gamma_I^* - E(\gamma_I^*)/\sqrt{\mathrm{Var}(\gamma_I^*)}$, which for $N \gg 1/p_{II}$ has nearly the normal distribution with mean 0 and variance 1. In terms of this variable, we can rewrite definition (14) as follows:

$$P_I = \mathrm{Prob}\left(\beta > \frac{-E(\gamma_I^*)}{\sqrt{\mathrm{Var}(\gamma_I^*)}}\right) = \mathrm{Prob}(\beta > -\beta_{P_I}). \tag{18}$$

The correspondence between P_I and β_{P_I} can be obtained from the statistical table of the standard normal distribution. For example, for $P_I = 0.95$, $\beta_{0.95} \approx 1.65$. Defining $\beta_{\hat{P}_I}$ for a given $\hat{P}_I$ and using formulas (16) and (17) for $E(\gamma_I^*)$ and $\mathrm{Var}(\gamma_I^*)$, we can estimate the sequence length $N_{\hat{P}_I}$ that is required for the probability of obtaining tree I to be P_I:

$$N_{\hat{P}_I} \approx \left[\frac{\sqrt{p_{II}/\pi}+\beta_{\hat{P}_I}\sqrt{p_I+(1-(1/\pi))p_{II}}}{\Delta p_I}\right]^2. \tag{19}$$

Usually, this formula underestimates $N_{\hat{P}_I}$, because it does not take into account the discreteness of the model. Actually, there exists a nonzero probability of $\gamma_I^* = 0$ (the probability of having a trichotomy, P_0) that reduces P_I by approximately one half of P_0; that is, if we take $N_{\hat{P}_I}$ from formula (19), we actually obtain $P_I = \hat{P}_I - 0.5P_0$. As $N_{0.95}$ increases, P_0 decreases, and P_I approaches $\hat{P}_I$. A simple way to correct such an underestimation is to define $P_I = \mathrm{Prob}(\gamma_I^* > 1/N)$, rather than $P_I = \mathrm{Prob}(\gamma_I^* > 0)$;

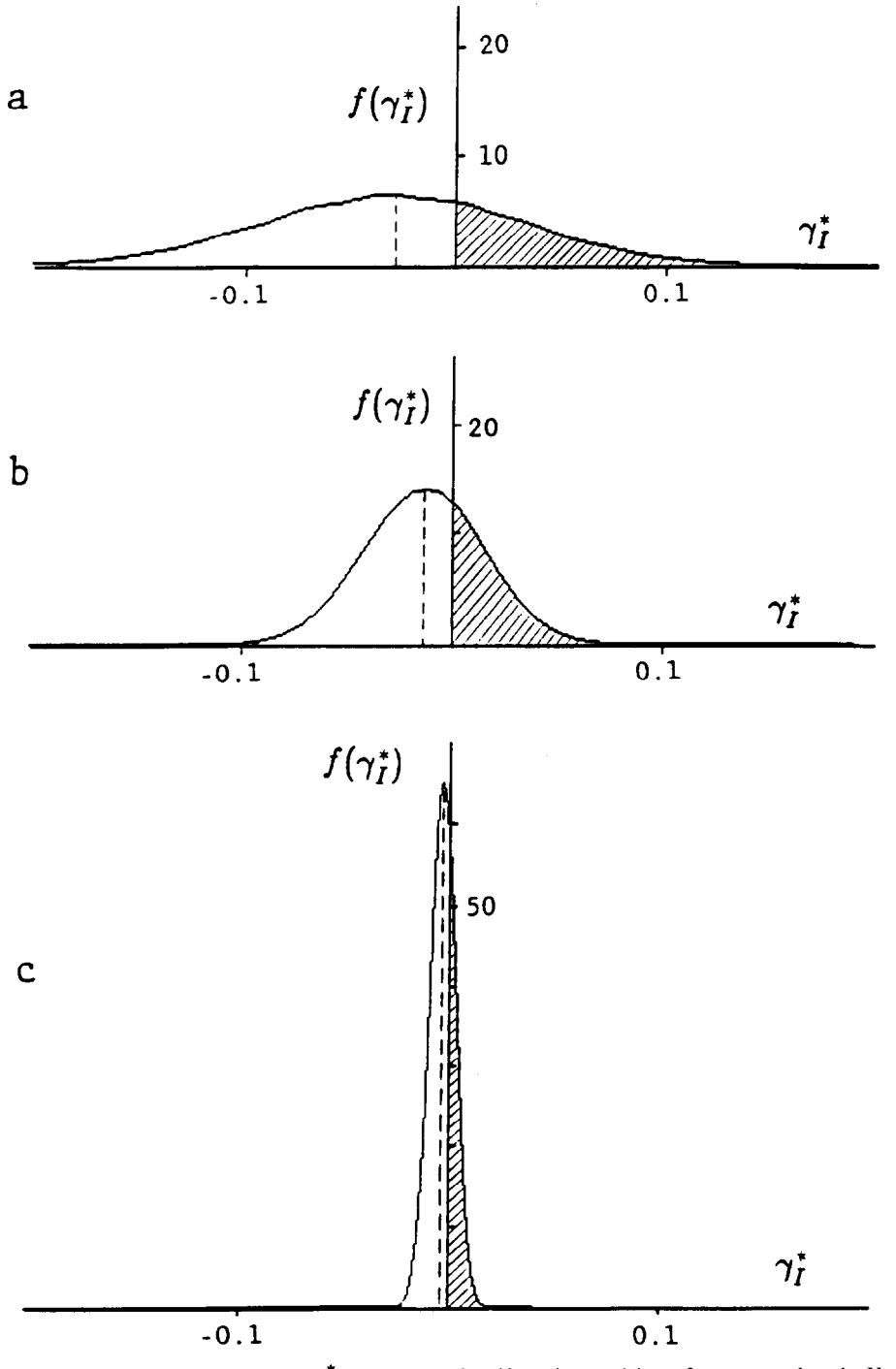

FIG. 3.—Probability density of γ_I^* for the case of a trifurcating model tree for sequence lengths $N = 20$ (a), $N = 82$ (b), and $N = 2{,}315$ (c). These values are chosen to provide a comparison with the cases of bifurcating trees shown in fig. 4. The probabilities are calculated using the multinomial distribution of the numbers of informative sites N_I, N_{II}, and N_{III} with expected proportions $p_I = p_{II} = p_{III} = 0.044$. This case corresponds approximately to the model in fig. 1a with time parameters $T_1 = 100$ Myr, $T_2 = T_3 = 50$ Myr, and the evolutionary rate of $\mu = 10^{-8}$ substitutions per site per year. For $N > 100$, the normal approximation of the binomial distribution was applied. The dashed line on each plot corresponds to the mean value of γ_I^*. The shaded part of each distribution represents the expected proportion of tree I—i.e., P_I—which is approximately the same for any length of sequences.

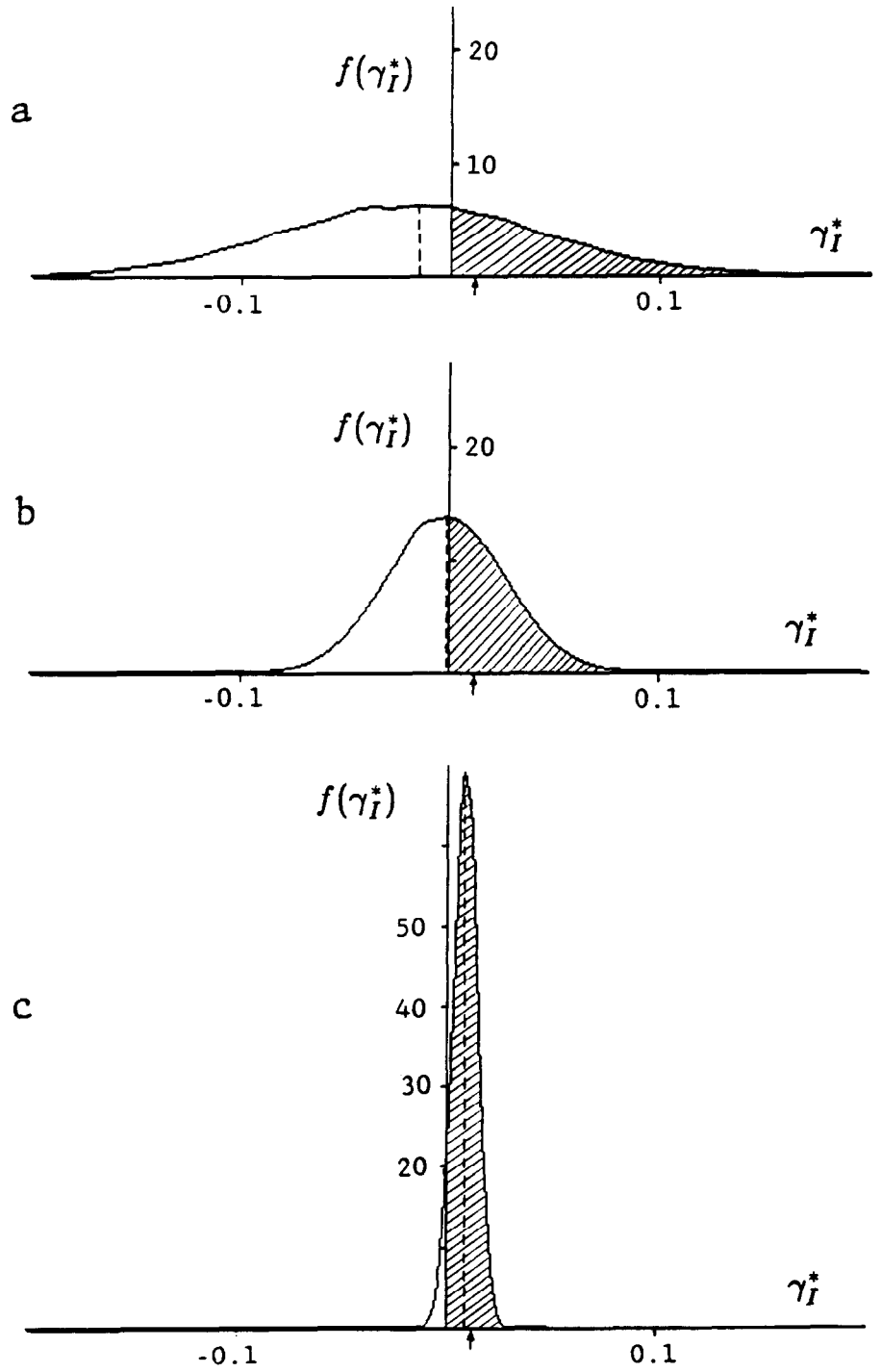

FIG. 4.—Probability density of γ_I^* for the case of a bifurcating tree (fig. 1a) with time parameters T_1 = 100 Myr, T_2 = 60 Myr, and T_3 = 50 Myr and with the evolutionary rate of $\mu = 10^{-8}$ substitutions per site per year ($\alpha=\beta$). These parameter values are chosen to give three qualitatively different types of the probability density plot: $E(\gamma_I^*) < 0$ (a), $E(\gamma_I^*) \approx 0$ (b), and $E(\gamma_I^*) > 0$ (c). The expected proportions of informative sites are $p_I = 0.0541$ and $p_{II} = p_{III} = 0.0417$. The difference $\Delta p_I = p_I - p_{II}$ is indicated by an arrow on the abscissa. From eq. (20), $N_{0.5} = 166$ and $N_{0.95} = 2{,}315$. The sequence lengths used are $N = 20$ (a), $N = 82$ (b), and $N = 2{,}315$ (c). The expected proportions of type I trees—i.e., P_I (shaded area)—increases with N: $P_I = 0.293$, 0.422, and 0.951 for plots a, b, and c, respectively.

for the probability of having tree I. This increases the estimate of $N_{\hat{P}_I}$ given by formula (19) approximately by $1/\Delta p_I$ (see Fleiss 1981, p. 42):

$$N_{\hat{P}_I} \approx \left[\frac{\sqrt{p_{II}/\pi}+\beta_{\hat{P}_I}\sqrt{p_I+(1-(1/\pi))p_{II}}}{\Delta p_I}\right]^2+\frac{1}{\Delta p_I}. \quad (20)$$

A detailed investigation of the relationship between sequence length and the probability of obtaining the correct tree was provided by Saitou and Nei (1986). Using various evolutionary models and applying various tree-making methods, they estimated the minimum sequence length that is required for having the probability $P_I = 0.95$ of obtaining the true phylogeny for three species with one or two outgroups. For short sequences ($N<100$), they applied the exact multinomial formula for the calculation of P_I. This approach becomes extremely tedious for long sequences. For this reason they used simulation when $N > 100$. In one of their model trees for three species with an outgroup, they selected the following parameters: $T_1\mu = 0.09$, $T_2\mu = 0.05$, and $T_3\mu = 0.045$ (fig. 1a). Two models of nucleotide substitution were used: the one-parameter model with $\alpha = \beta$ and Kimura's two-parameter model with $\alpha = 20\mu/22$ and $\beta = \mu/22$. For these two models, they obtained $N_{0.95} = 2{,}100$ and $N_{0.95} = 3{,}300$, respectively, for the maximum parsimony method. Our formula (20) gives similar estimates: $N_{0.95} = 2{,}153$ and $N_{0.95} = 3{,}312$ for the one- and two-parameter models, respectively. A good agreement between formula (20) and simulation results will be seen later (in table 4).

In tables 1 and 2 we present values of p_I and $p_{II} = p_{III}$ calculated from formulas (8) and (9) for tree I, with the time for the outgroup-branching-point $T_1 = 100$ Myr and T_2 and T_3 varying from 0 to 100 Myr, and with the corresponding values of $N_{0.95}$ and $N_{0.5}$ given by formula (20) for $\hat{P}_I = 0.95$ and $\hat{P}_I = 0.5$, respectively. For the evolutionary rate, we used two different values: $\mu = 10^{-9}$ and 10^{-8}; the former is similar to the average rate of nonsynonymous substitution, while the latter is approximately two times higher than the average rate of synonymous substitution for commonly studied mammalian genes (Li and Graur 1991).

Bootstrap Estimation

Equation (20) can be used also for estimating β, from which one can infer the expected proportion of type I trees, P_I, if the sequence length, N, and the proportions of informative sites, p_I, p_{II}, and p_{III} are given. However, such a direct estimation of P_I for a tree with more than four species is a difficult task. For this purpose, one can use the bootstrap technique, which was introduced into phylogenetic studies by Felsenstein (1985). The characters under study are assumed to evolve independently. In the bootstrap estimation procedure, the sites of the sequences under study are resampled randomly with replacement, and a tree is reconstructed for each resampled data set. It is supposed that the resampled data have the same distribution of informative sites as do repeated samples from the original process. For example, in the case of four species, the proportions P_I^{**}, P_{II}^{**}, and P_{III}^{**} of trees I, II, and III among the bootstrap replicates are the estimates of proportions P_I, P_{II}, and P_{III}, respectively.

As mentioned above, the bootstrap estimate of P_I is a result of two steps of sampling:

$$P_I \rightarrow P_I^* \rightarrow P_I^{**}, \quad (21)$$

where the first step is the sampling of sequences from the evolutionary process and

Table 1
Proportions (%) of Informative Sites p_I and $p_{II} = p_{III}$ (on and above the Diagonal)—and Sequence Lengths $N_{0.95}$ and $N_{0.5}$ Required for Having Probability, $P_I = 0.95$ and 0.5, Respectively (below the Diagonal), of Obtaining Tree I

	T_3										
T_2	0	10	20	30	40	50	60	70	80	90	100
0	0.00	4.71	8.36	11.18	13.37	15.08	16.40	17.43	18.24	18.87	19.37
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	78	1.86	5.27	7.92	9.98	11.57	12.81	13.78	14.54	15.13	15.59
	21	1.86	1.66	1.52	1.41	1.33	1.27	1.22	1.18	1.16	1.14
20	44	211	3.01	5.50	7.44	8.94	10.11	11.02	11.73	12.29	12.73
	11	31	3.01	2.74	2.53	2.38	2.26	2.17	2.11	2.06	2.03
30	33	92	417	3.72	5.55	6.97	8.07	8.94	9.61	10.14	10.55
	8	16	47	3.72	3.43	3.21	3.05	2.93	2.83	2.77	2.72
40	27	62	163	761	4.15	5.50	6.55	7.37	8.01	8.51	8.90
	7	12	23	71	4.15	3.87	3.67	3.52	3.40	3.32	3.26
50	24	49	104	280	1,341	4.41	5.41	6.19	6.80	7.28	7.65
	6	10	16	33	108	4.41	4.17	3.99	3.85	3.75	3.68
60	22	42	80	173	473	2,315	4.56	5.31	5.89	6.35	6.71
	6	8	13	23	48	166	4.56	4.36	4.20	4.09	4.01
70	21	38	67	130	285	795	3,960	4.65	5.21	5.65	6.00
	5	8	12	19	33	71	258	4.65	4.48	4.36	4.27
80	20	35	60	107	210	470	1,333	6,742	4.71	5.13	5.46
	5	7	11	16	26	48	106	405	4.71	4.57	4.47
90	19	33	55	94	172	342	777	2,237	11,460	4.74	5.06
	5	7	10	15	23	37	69	160	646	4.74	4.64
100	19	32	51	86	150	278	561	1,291	3,763	19,474	4.77
	5	7	9	14	20	32	54	103	246	1,040	4.77

NOTE.—In each cell on and above the diagonal, the top number is the proportion (%) of informative sites p_I, and the bottom number is the proportion (%) of informative sites $p_{II} = p_{III}$. In each cell below the diagonal, the top number is $N_{0.95}$, and the bottom number is $N_{0.5}$. The diagonal elements correspond to the cases of trifurcating trees. All these values are calculated using expressions (8), (9), and (20), for $\mu = 10^{-8}$, $T_1 = 100$ Myr, and various combinations of the divergence times T_2 and T_3.

where the second step is the bootstrap resampling. P_I, as defined in the previous section, is the probability that a random sample of sequences from the evolutionary process will support tree I. Now suppose that a sample of sequences is taken. Bootstrapping of this original sample of sequences produces new samples (bootstrap replicates) each of which supports tree I with probability P_I^*. From the resampled data sets (i.e., the bootstrap replicates), one calculates the proportion P_I^{**} of the bootstrap replicates that support tree I. This proportion is actually an estimate of P_I^* rather than of P_I.

For a given set of sequences, the proportion P_I^{**} has the binomial distribution with the mean and the variance

$$\mathrm{E}(P_I^{**} \mid P_I^*) = P_I^* \qquad \text{and} \qquad \mathrm{Var}(P_I^{**} \mid P_I^*) = \frac{P_I^*(1 - P_I^*)}{N_b}, \tag{22}$$

where N_b is the number of bootstrap replications. Because P_I^* is, in turn, a random variable, the distribution of P_I^{**} is actually a compound distribution (Johnson and Kotz 1969, p. 183), the mean and the variance of which are

Table 2
Proportions (%) of Informative Sites p_I and $p_{II} = p_{III}$ (on and above the Diagonal)—and Sequence Lengths $N_{0.95}$ and $N_{0.5}$ (below the Diagonal)—Calculated for $\mu = 10^{-9}$, $T_1 = 100$ Myr, and Various Combinations of Divergence Times T_2 and T_3

	T_3										
T_2	0	10	20	30	40	50	60	70	80	90	100
0	0.00	0.87	1.73	2.57	3.39	4.19	4.98	5.75	6.51	7.25	7.97
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	424	0.06	0.91	1.74	2.56	3.35	4.14	4.90	5.65	6.39	7.11
	114	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05
20	214	532	0.11	0.94	1.75	2.54	3.32	4.08	4.83	5.56	6.27
	57	120	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.10	0.10
30	144	250	616	0.16	0.97	1.76	2.53	3.29	4.03	4.75	5.46
	38	60	125	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.15
40	109	163	276	700	0.21	1.0	1.76	2.52	3.25	3.97	4.68
	29	40	62	131	0.21	0.21	0.21	0.21	0.21	0.21	0.20
50	88	122	177	302	787	0.26	1.03	1.77	2.50	3.22	3.92
	23	30	41	64	138	0.26	0.26	0.26	0.26	0.25	0.25
60	74	97	131	191	329	878	0.31	1.05	1.78	2.49	3.19
	20	24	31	43	67	144	0.31	0.31	0.30	0.30	0.30
70	64	81	104	140	205	357	974	0.35	1.08	1.78	2.48
	17	20	25	32	44	69	151	0.35	0.35	0.35	0.35
80	56	70	86	110	149	219	385	1,076	0.40	1.10	1.79
	15	17	21	26	33	46	72	159	0.40	0.39	0.39
90	51	61	74	91	117	158	235	416	1,184	0.44	1.12
	13	15	18	21	26	34	47	75	166	0.44	0.44
100	46	55	65	78	96	124	168	250	448	1,298	0.48
	12	14	16	19	22	27	35	49	77	174	0.48

NOTE.—In each cell on and above the diagonal, the top number is the proportion (%) of informative sites p_I, and the bottom number is the proportion (%) of informative sites $p_{II} = p_{III}$. In each cell below the diagonal, the top number is $N_{0.95}$, and the bottom number is $N_{0.5}$.

$$\mathrm{E}(P_I^{**}) = \mathrm{E}(P_I^{*}) \tag{23}$$

and

$$\begin{aligned}\mathrm{Var}(P_I^{**}) &= \mathrm{Var}[\mathrm{E}(P_I^{**}|P_I^{*})]+\mathrm{E}[\mathrm{Var}(P_I^{**}|P_I^{*})] \\ &= \mathrm{Var}(P_I^{*})+\frac{1}{N_b}\mathrm{E}[P_I^{*}(1-P_I^{*})]\,.\end{aligned} \tag{24}$$

We can see that the variance consists of two components: the first one, $\mathrm{Var}(P_I^{*})$, represents the variance of sampling of sequence data from the evolutionary process, and the second represents the variance arising from bootstrap resampling. Note that the second component decreases to 0 as $N_b \to \infty$ but that the first component is independent of N_b and remains constant even as $N_b \to \infty$. However, the distribution of P_I^{**} approaches the distribution of P_I^{*} as $N_b \to \infty$. The variance $\mathrm{Var}(P_I^{*})$ refers to the effects of sampling of sequences (with finite length N) from the evolutionary process and can be reduced to 0 only by increasing N to infinity. For finite N, P_I^{*} will vary among samples, and so will P_I^{**}, regardless of the number of bootstrap replications conducted. Therefore, to understand the full variation of P_I^{**}, one needs to consider

not only the variation over bootstrap replicates but also the variation over samples taken from the evolutionary process.

Obviously, to understand the distribution of P_I^{**}, we need to study the distribution of P_I^*. We now characterize the distribution of P_I^*. We begin by recalling the distribution of γ_I^* that was described in the previous section. In figure 5a, an example of the distribution of γ_I^* among the original data sets is shown. For this distribution, the probability of obtaining tree I is defined by equation (14). For our purpose, it is more convenient to write it in the continuous mode:

$$P_I = 1 - \int_{-1}^{0} f(\gamma_I^*) d\gamma_I^*, \tag{25}$$

where $\int_{-1}^{x} f(\gamma_I^*) d\gamma_I^* = \text{Prob}(\gamma_I^* \leq x)$, i.e., $f(\gamma_I^*)$ represents the probability density (frequency) function of γ_I^*, with the mean $\bar{\gamma}_I^* = \text{E}(\gamma_I^*)$.

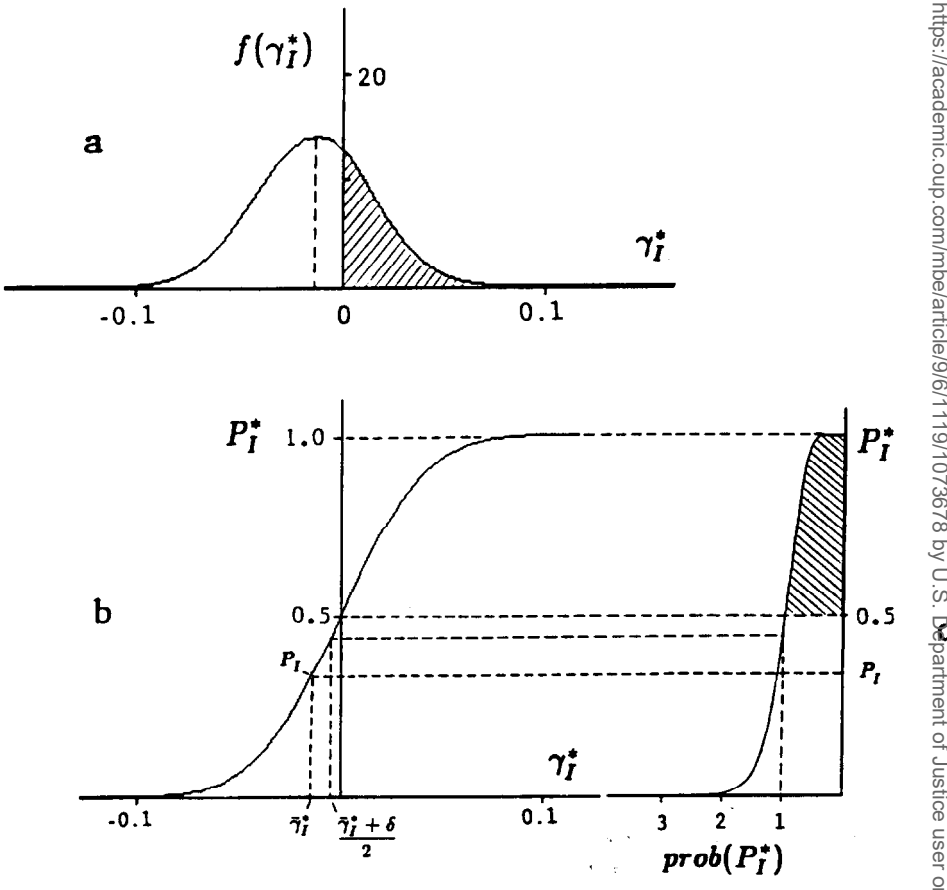

FIG. 5.—Graphic representation of the frequency-function inference for the expected proportion of type I trees, P_I^*. a, Probability density function of γ_I^*. The parameter values used are the same as in fig. 3. The sequence length used is $N = 82$. The shaded area is equal to P_I. b, Expected proportion of type I trees, P_I^*, for the sampled data that are characterized by the given value γ_I^* [eq. (28)]. For $\gamma_I^* = 0$, this proportion is $P_I^*(0) \approx 0.5$; for $\gamma_I^* = \bar{\gamma}_I^*$, $P_I^*(\bar{\gamma}_I^*) \approx P_I$. c, Probability density function of P_I^* [eq. (31)]. To correspond with plot b, the axes have been rotated by 90° counterclockwise. The dashed line corresponding to $\gamma_I^* = (\bar{\gamma}_I^* + \delta)/2$ gives the probability density $\text{prob}(P_I^*) \approx 1$.

Once a sample is taken from the evolutionary process, it is characterized by a particular value of $\gamma_I^* = p_I^* - \max(p_{II}^*, p_{III}^*) = \Delta p_I^*$. Suppose that $p_{II}^* \geq p_{III}^*$. Then $\Delta p_I^* = p_I^* - p_{II}^*$. Now consider bootstrap resampling of the original sample. Denote the difference function for a resampled data set by γ_I^{**}. The distribution of γ_I^{**} is characterized by the frequency function $g(\gamma_I^{**} | \gamma_I^*)$ with expectation $\bar{\gamma}_I^{**}$. By analogy with equation (16), the value of $\bar{\gamma}_I^{**}$ can be defined as

$$\bar{\gamma}_I^{**} = \mathrm{E}(\gamma_I^{**} | \gamma_I^*) \approx \Delta p_I^* - \alpha \sqrt{\frac{p_{II}^*(1-p_{II}^*)}{N}} = \gamma_I^* - \delta , \tag{26}$$

where $\delta = \alpha\sqrt{p_{II}^*(1-p_{II}^*)/N}$. Because the proportions p_{II}^* and p_{III}^* in a sample are often unequal, the value of α in this case is likely to differ from $1/\sqrt{\pi}$, unlike the case of equation (16). From equation (15), if $p_{III}^* \to 0$, then $\mathrm{E}(\gamma_I^*) \to \mathrm{E}(p_I^* - p_{II}^*) = \Delta p_I$ and $\alpha \to 0$. In general, $0 \leq \alpha \leq \pi^{-1/2}$.

For long sequences, the distribution $g(\gamma_I^{**} | \gamma_I^*)$ of γ_I^{**} among the resampled data sets will have approximately the same shape as does the original distribution $f(\gamma_I^*)$. The two distributions differ from each other only by the shift $\bar{\gamma}_I^* - \bar{\gamma}_I^{**}$ in the abscissa, which is the difference between the mean of $f(\gamma_I^*)$ and the mean of $g(\gamma_I^{**} | \gamma_I^*)$. That is,

$$g(\gamma_I^{**} | \gamma_I^*) \approx f(\gamma_I^{**} + \bar{\gamma}_I^* - \bar{\gamma}_I^{**}) . \tag{27}$$

Thus, by analogy with equation (25) we can write the particular distribution of γ_I^{**} given γ_I^* and define the expected proportion P_I^* of type I trees among the resampled data sets as a function of γ_I^* (fig. 5b):

$$\begin{aligned} P_I^*(\gamma_I^*) &= 1 - \int_{-1}^{0} g(\gamma_I^{**} | \gamma_I^*) d\gamma_I^{**} \approx 1 - \int_{-1}^{0} f(\gamma_I^{**} + \bar{\gamma}_I^* - \bar{\gamma}_I^{**}) d\gamma_I^{**} \\ &= 1 - \int_{-1}^{\bar{\gamma}_I^* - \bar{\gamma}_I^{**}} f(\gamma_I^{**}) d\gamma_I^{**} \approx 1 - \int_{-1}^{\bar{\gamma}_I^* - \bar{\gamma}_I^* + \delta} f(\gamma_I^{**}) d\gamma_I^{**} . \end{aligned} \tag{28}$$

From probability theory, it is known that, if x is a random variable with the frequency function $f(x)$ and if $y = u(x)$ is a monotonic function, then the frequency function of y can be expressed as follows:

$$w(y) = \frac{f(x)}{u'(x)} . \tag{29}$$

Taking $u(x)$ as $P_I^*(\gamma_I^*)$, we can derive the frequency function of P_I^*. From equation (28),

$$\frac{dP_I^*}{d\gamma_I^*} \approx f(\bar{\gamma}_I^* - \gamma_I^* + \delta) . \tag{30}$$

So, the frequency function of P_I^* is (fig. 5c)

$$\text{prob}(P_I^*) \approx \frac{f(\gamma_I^*)}{f(\bar{\gamma}_I^* - \gamma_I^* + \delta)} . \tag{31}$$

Note that for $\gamma_I^* = (\bar{\gamma}_I^* + \delta)/2$, $\bar{\gamma}_I^* - \gamma_I^* + \delta = (\bar{\gamma}_I^* + \delta)/2 = \gamma_I^*$, and so the value of the above function is equal to 1. This point is indicated on the abscissa of the plot in figure 5c.

Combining equations (30) and (31), we obtain

$$\text{Prob}(P_I^* < y) = \int_0^y \text{prob}(P_I^*) dP_I^* = \int_{-1}^x f(\gamma_I^*) d\gamma_I^* = \text{Prob}(\gamma_I^* < x) , \tag{32}$$

where $y = P_I^*(x)$, in correspondence with definition (28). For large N, equation (32) gives the following two characteristic points of the distribution of P_I^*:

$$\text{Prob}(P_I^* > ½) \approx \text{Prob}(\gamma_I^* > 0) \approx P_I , \tag{33}$$

and

$$\text{Prob}(P_I^* > P_I) \approx \text{Prob}(\gamma_I^* > \bar{\gamma}_I^*) \approx ½ ; \tag{34}$$

that is, for large N, the probability for P_I^* (the expected proportion of obtaining tree I among bootstrap replicates) to be >½ is approximately P_I, and the probability for $P_I^* > P_I$ is ~½. The prob(P_I^*) values corresponding to these two points are shown in figure 5.

In figure 6, the frequency functions of P_I^* for a trifurcating and a bifurcating model tree are presented. Figure 6a is calculated for the case of the trifurcating tree (fig. 1d). The proportions of all types of informative sites in this case are equal: $p_I = p_{II} = p_{III}$. The probabilities of different types of sampled trees will also be equal, $P_I = P_{II} = P_{III}$, and its value approaches ⅓ when $N \rightarrow \infty$. Although equation (31) is inferred for long sequences, the main properties of the frequency function prob(P_I^*) hold also for short sequences. For the parameter values used, if $N > 20$, then the frequency function prob(P_I^*) is practically independent of the sequence length. Note that the frequency function of P_I^* has a negative slope; that is, it decreases with increasing P_I^*. Therefore, the probability for P_I^* to be ≥95% is small, and so is the probability for tree I to appear in ≥95% of the bootstrap replicates.

Shown in figure 6b–d are graphs corresponding to the bifurcating tree (fig. 1a). Figure 6c represents the case where the distribution $f(\gamma_I^*)$ is symmetrical, i.e., $f(\bar{\gamma}_I^* - \gamma_I^* + \delta) = f(\bar{\gamma}_I^* + \gamma_I^* - \delta)$. In this case, if $\bar{\gamma}_I^* = \delta$, then, from equation (31), $\text{prob}(P_I^*) = f(\gamma_I^*)/f(\gamma_I^*) \equiv 1$. According to equations (16) and (26), this condition occurs when $E(\gamma_I^{**}) = 0$ and $E(\gamma_I^*) = \Delta p_I - \sqrt{p_{II}(1-p_{II})/N\pi} = \delta$; therefore, $\Delta p_I = \sqrt{p_{II}(1-p_{II})/N\pi} + \delta$. So, for a sequence length close to $N' = (\alpha + \pi^{-0.5})^2 p_{II}(1-p_{II})/(\Delta p)^2$, we will have a nearly constant frequency function of P_I^*, i.e., a nearly uniform distribution. For the proportions of informative sites, $p_I = 0.0541$ and $p_{II} = p_{III} = 0.0417$, used in the model tree, $N' = 200$ (fig. 6c). For short sequences ($N < N'$), the frequency function, prob(P_I^*), has a negative slope (fig. 6b). When $N > N'$, this

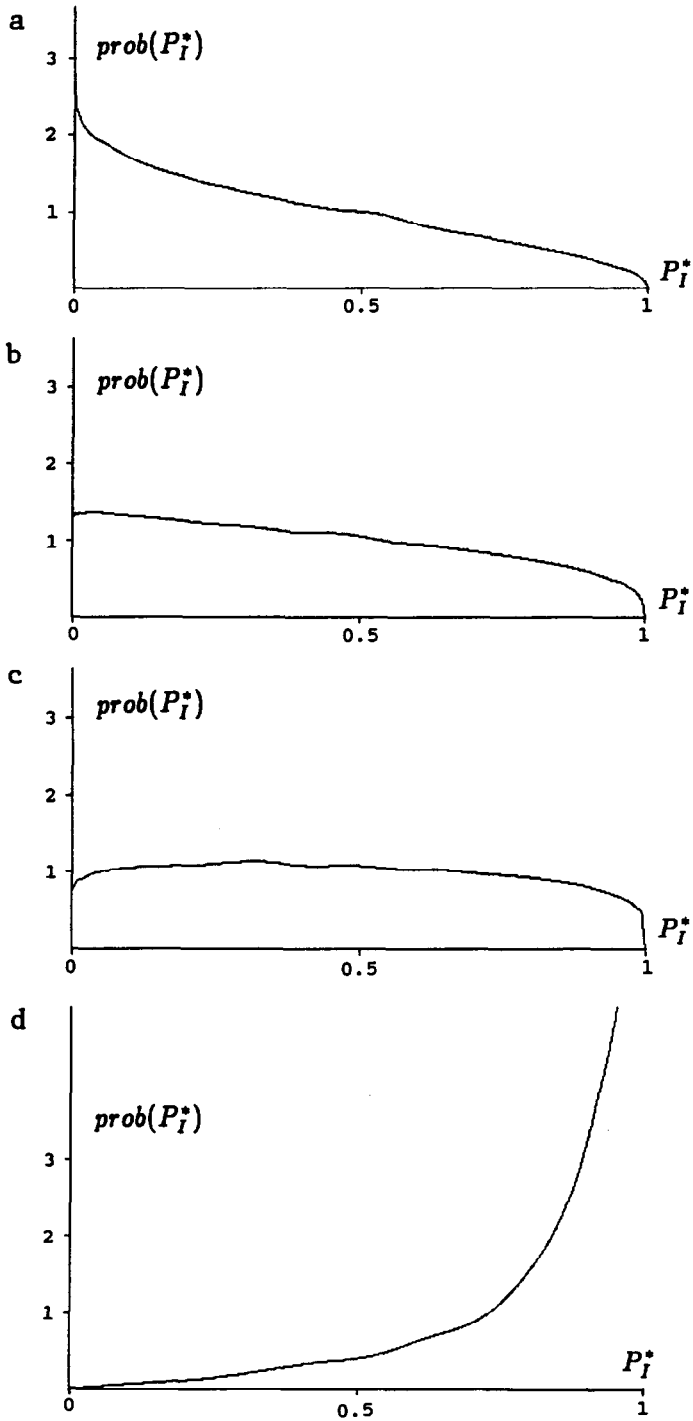

FIG. 6.—Distribution of the expected proportion of type I trees, P_I^*, among bootstrap replicates for (a) a trifurcating model tree and (b–d) a bifurcating model tree, with the same parameters as in figs. 3 and 4, respectively. In the case of trifurcation, the plot of the distribution is nearly the same for all sequence lengths $N \geq 20$. For the bifurcating tree the distribution depends on N: when $N < 200$ (N=82), the plot has a negative slope (b); when $N = 200$, the frequency function is approximately constant, prob(P_I^*) $\cong 1$ (c); and when $N > 200$ (N=2,315), the plot has a positive slope (d).

function has a positive slope. In figure 6d, $N = 2{,}315$, which is much larger than $N' = 200$, and prob(P_I^*) increases with P_I^*, particularly after P_I^* becomes >75%. Note, however, that even in this case, where $P_I = 95\%$, the probability for P_I^* to be ≥95% is still not large. To see the difference between P_I and P_I^*, let us consider a hypothetical example. Suppose that a sample of sequences is taken and that there are 10, 8, and 8 informative sites supporting trees I, II, and III, respectively. In this sample, tree I will be chosen as the true tree, but P_I^* is certainly <95%, because the number of informative sites supporting tree I is only slightly higher than the number of those supporting trees II and III, so that the probability that a resampling of the original sample will fail to support tree I is >5%.

Phylogenetic Inference

All the above analyses assume that we know a priori the true phylogeny (tree I). It means that, for any kind of sample, we always estimate the probability of having tree I. Let the probabilities of obtaining a sample supporting tree Y, $Y = I, II, III$ (the bifurcating trees) or 0 (the trifurcating tree), be P_I, P_{II}, P_{III}, and P_0, respectively. Then the probability of obtaining tree I from a bootstrap resampling of a random sample of sequence is

$$P_I^* = \text{Prob}(R_I|S_I)P_I + \text{Prob}(R_I|S_{II})P_{II} + \text{Prob}(R_I|S_{III})P_{III} + \text{Prob}(R_I|S_0)P_0\,, \quad (35)$$

where $\text{Prob}(R_X|S_Y)$ is the conditional probability that a resampling of sample Y will support tree X.

In usual practice, we infer from a given set of sequences a phylogeny that can be classified as any one of the three possible bifurcating trees in figure 1 (for long sequences, samples supporting the trifurcating tree are usually rare and are neglected in this analysis). We then conduct bootstrapping and compute the proportion of bootstrap replicates that support the inferred tree. The probability that the tree obtained in a single bootstrap replicate is the same as the inferred tree is given by

$$P_X^* \approx \text{Prob}(R_I|S_I)P_I + \text{Prob}(R_{II}|S_{II})P_{II} + \text{Prob}(R_{III}|S_{III})P_{III}\,. \quad (36)$$

Obviously, P_X^* tends to be $> P_I^*$.

In terms of the difference function, we consider tree I as the true tree only when the number of type I informative sites, N_I, is the largest, i.e., $\gamma_I^* > 0$. Otherwise, we assume the true tree to be tree II, if $\gamma_{II}^* > 0$, or tree III, if $\gamma_{III}^* > 0$. Since in sample estimation all three types of decisions may be made, we call such decisions "mixed decisions." To study the statistical properties of P_X^*, let us construct a new difference function

$$\gamma_X^* = \frac{N_{\max}}{N} - \frac{N_{\text{med}}}{N}\,, \quad (37)$$

where $N_{\max} = \max(N_I, N_{II}, N_{III})$, and where N_{med} is the second largest of the three numbers. The function γ_X^* is defined in the region $0 \le \gamma_X^* \le 1$ and is characterized by the frequency function $f_X(\gamma_X^*)$. We will use the function f with the subscripts X, I, II, and III to distinguish among the frequency functions of γ_X^*, γ_I^*, γ_{II}^*, and γ_{III}^*, respectively. Because $\gamma_I^* > 0$, $\gamma_{II}^* > 0$, and $\gamma_{III}^* > 0$ are mutually exclusive events, the function $f_X(\gamma_X^*)$ is simply the sum of all the functions $f_I(\gamma_X^*)$, $f_{II}(\gamma_X^*)$, and $f_{III}(\gamma_X^*)$ taken in the positive region of their arguments:

$$f_X(\gamma_X^*) = f_I(\gamma_X^*)+f_{II}(\gamma_X^*)+f_{III}(\gamma_X^*)\,, \tag{38}$$

where $\gamma_X^* > 0$.

In figure 7, the frequency functions of γ_X^* for various sequence lengths are drawn. The upper set of the plots (fig. 7a–c) corresponds to the trifurcating model tree. As in the case of the frequency function of γ_I^* (fig. 3), the plots for different sequence lengths can be transformed to each other by rescaling the axes x and y. The lower set of the plots (fig. 7d–f) corresponds to the case of bifurcation. As the sequence length increases, the distribution becomes narrower. For $N > N_{0.95}$, the function $f_X(\gamma_X^*)$ (fig. 7f) becomes similar to the function $f_I(\gamma_I^*)$ (fig. 4c).

According to equation (31), each of the terms in equation (38) gives the corresponding component of the probability density function of P_X^*:

$$\text{prob}_X(P_X^*) \approx \text{prob}_I(P_X^*)+\text{prob}_{II}(P_X^*)+\text{prob}_{III}(P_X^*)\,. \tag{39}$$

As in the case of $f_X(\gamma_X^*)$, we use the notation $\text{prob}_X(P_X^*)$ to distinguish it from the previously defined functions $\text{prob}_Y(P_Y^*)$, $Y = I$, II, and III. Because γ_X^* is always >0, equation (33) implies that $P_X^* > 1/2$ (fig. 8c). Graphically, the function $P_X^*(\gamma_X^*)$ shown in figure 8b defines the correspondence between the probability density functions $f_X(\gamma_X^*)$ and $\text{prob}_X(P_X^*)$ in the following manner: if $y = P_X^*(x)$, then $\text{Prob}(\gamma_X^* \geq x) = \text{Prob}(P_X^* \geq y)$.

For the trifurcating model tree (fig. 1d) all the three components in equation (39) are identical, and we have $\text{prob}_X(P_X^*) = 3\text{prob}_I(P_X^*)$ (fig. 9a). In this case, the properties of the distribution of P_X^* are very similar to those of the distribution of P_I^*. The plots for both distributions fit each other well if the latter is scaled by the multiplier 0.5 in the abscissa and by 2 in the ordinate and is shifted to the region [0.5, 1.0]. In particular, the mean value $E(P_I^*) \approx 1/3$ corresponds in this way to $E(P_X^*) \approx 0.5+(0.5\times 1/3) \approx 0.66$. This is very close to the value obtained by simulations (results not shown). This means that, under the trifurcating model tree, the expected proportion of bootstrap replicates supporting an observed bifurcating tree is close to 66%. As in the case of the frequency functions of P_I^* (fig. 6a), the plots for $\text{prob}_X(P_X^*)$ are approximately the same for different sequence lengths (fig. 9a).

The frequency functions of P_X^* in figure 9b–d correspond to the case of a bifurcating model tree. In figure 9b, $N = 200$, and $\text{prob}_X(P_X^*)$ is considerably higher than $\text{prob}_I(P_I^*)$, though the difference decreases as P_X^* increases from 0.5 to 1. Thus, in a sample of short sequences, P_X^* can be considerably larger than P_I^*. As the sequence length increases, the proportion of correct decisions P_I grows and the terms $f_I(\gamma_X^*)$ and $\text{prob}_I(P_X^*)$ in equations (38) and (39), respectively, become dominant. For $N > N_{0.95}$ (fig. 9d), the function $\text{prob}_X(P_X^*)$ is very similar to the function $\text{prob}_I(P_I^*)$.

Note that the condition for $\text{prob}_X(P_X^*)$ to be nearly constant requires a longer sequence length than does the corresponding condition for $\text{prob}_I(P_I^*)$ to be nearly constant. For example, when $N = 200$, the plot for $\text{prob}_I(P_I^*)$ is approximately constant (fig. 6c), but the corresponding plot for $\text{prob}_X(P_X^*)$ still has a negative slope (fig. 9b). Indeed, although the first term in equation (39)—$\text{prob}_I(P_X^*)$—is nearly constant for $N = 200$, the last two terms—$\text{prob}_{II}(P_X^*)$ and $\text{prob}_{III}(P_X^*)$— always have a negative slope, if tree I is the true tree. Thus, the sum of these functions will also have a negative slope. The slope disappears only when $N \approx 550$ (fig. 9c). For larger values of N, the plot for $\text{prob}_X(P_X^*)$ will have a positive slope (fig. 9d).

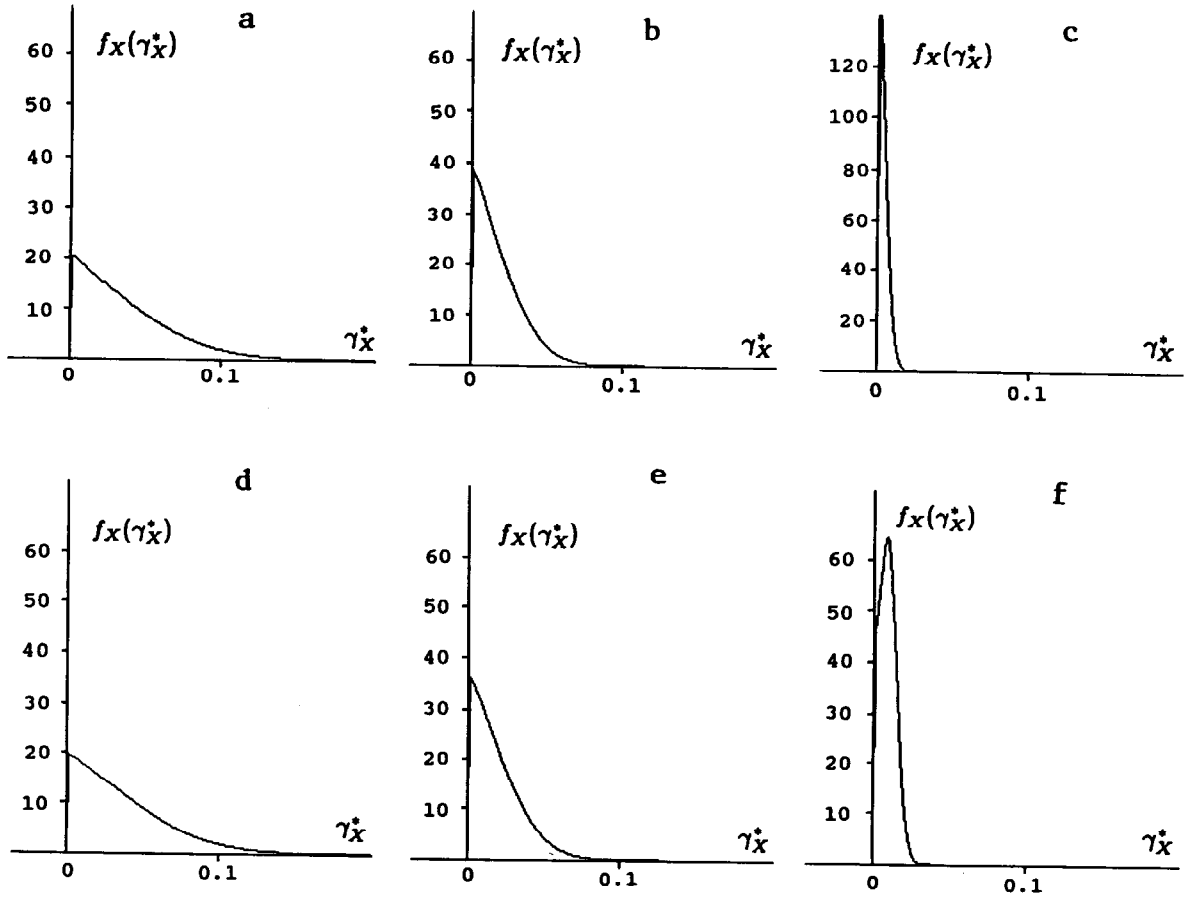

FIG. 7.—Probability density functions of γ_X^* for the case of a trifurcating tree (a–c) and a bifurcating tree (d–f), calculated in the same way and with the same parameters as for the density function of γ_I^* in figs. 3 and 4, respectively. The sequence lengths used are $N = 20$ (a and d); $N = 82$ (b and e); and $N = 2{,}315$ (c and f).

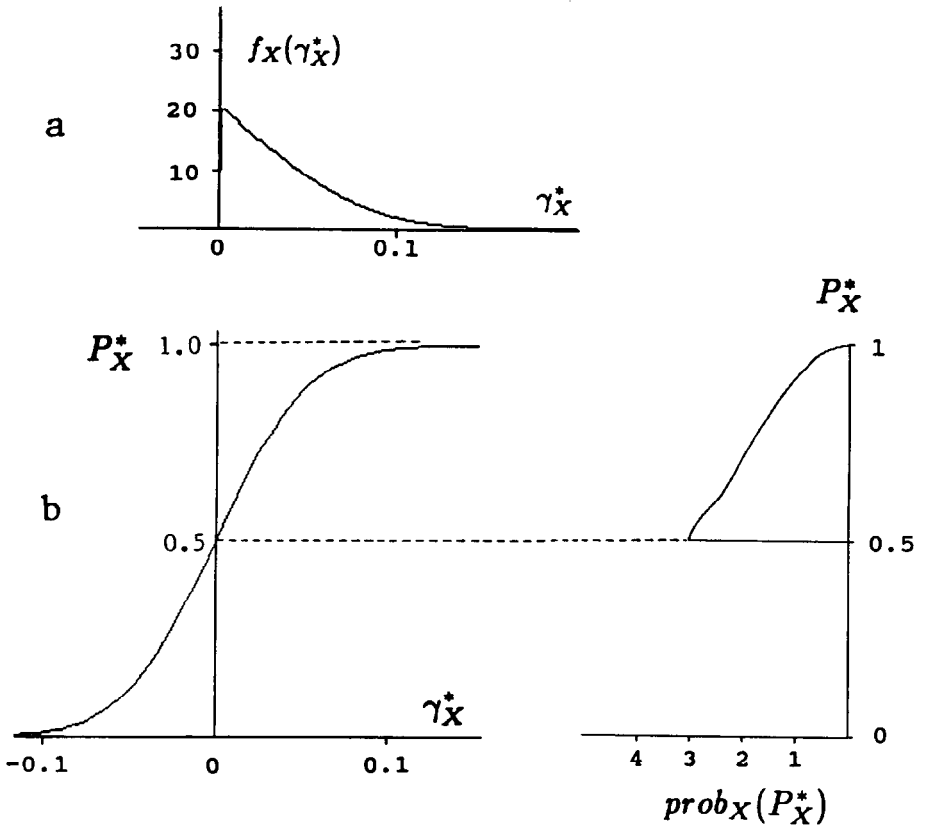

FIG. 8.—Inference of the frequency function of P_X^*, the expected proportion of bootstrap replicates supporting the inferred tree X. a, Frequency function of γ_X^* for the case of a trifurcating tree (N=20). The parameters used are the same as in fig. 3. b, Expected proportion of tree X (P_X^*) for the sampled value γ_X^*. This plot is the same as in fig. 5b. c, Probability density function of P_X^* [eq. (39)].

Bootstrap Estimation of Confidence Level

In the case of selecting one of the three alternative bifurcating trees for three taxa with one outgroup, the common practice of estimating the confidence level for a selected tree by bootstrapping is as follows: Let P_X^{**} be the proportion of bootstrap replicates in which tree X is chosen. Then, P_X^{**} is taken as the confidence level for three X. A common confidence level for accepting a tree is 95%. We investigate below the probability of accepting a tree at a given confidence level $\hat{P}_X$. This probability obviously depends on the number of bootstrap replications and on the sequence length. We shall also study the distribution of P_X^{**}.

An important question is, What is the probability of accepting an erroneous tree as the true tree? If the trifurcating tree is used as the model tree, then trees I–III are all considered as erroneous trees. Therefore, this model tree gives the largest probability of accepting an erroneous tree as the true tree. In this case, the probability is given by

$$\text{Prob}(P_X^{**} \geq \hat{P}_X) = \int_{\hat{P}_X}^{1} \text{prob}_X(P_X^{**}) dP_X^{**} . \quad (40)$$

In figure 10 the plots for $\text{Prob}(P_X^{**} \geq \hat{P}_X)$ are shown for various numbers of bootstrap

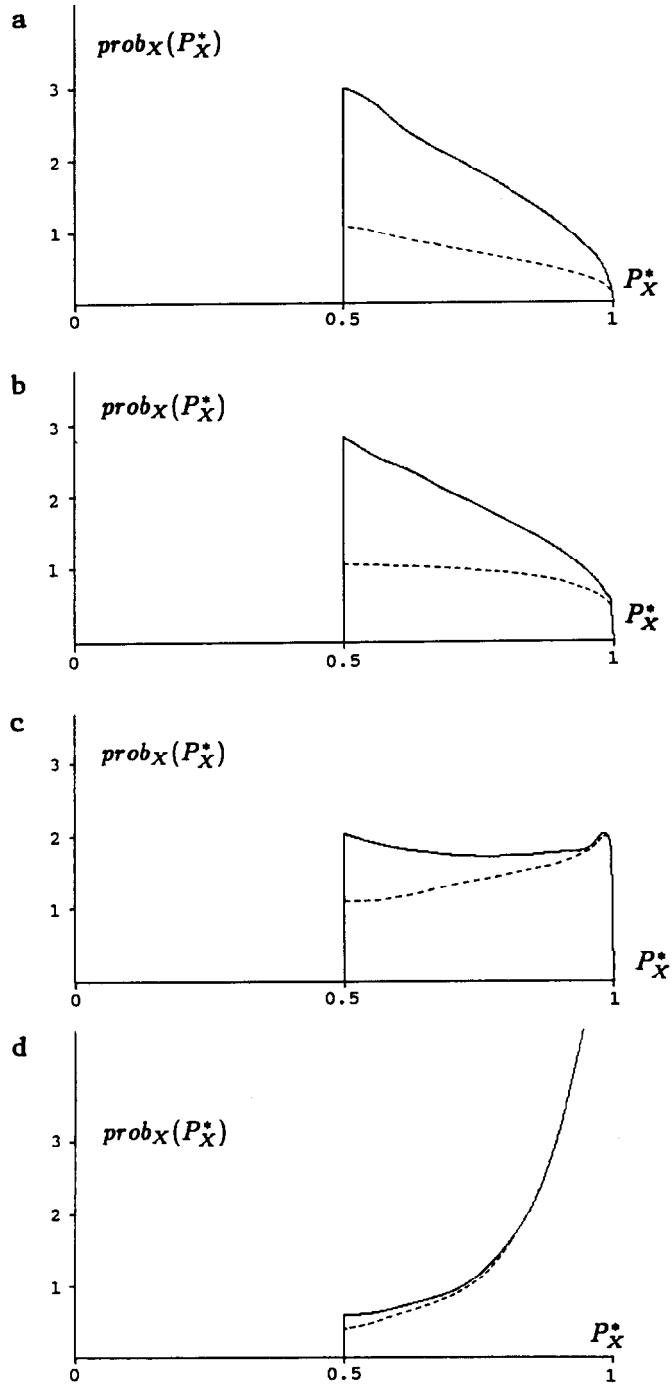

FIG. 9.—Distribution of P_X^*, the expected proportion of bootstrap replicates supporting the inferred tree X for (a) a trifurcating tree and (b–d) a bifurcating tree. The parameter values used are the same as in fig. 6. For the case of trifurcation, the plot of the distribution is the same for any sequence length. For the bifurcating tree, the sequence lengths used are $N = 200$ (b), $N = 550$ (c), and $N = 2{,}315$ (d). In each plot, the dashed line represents $\text{prob}_I(P_I^*)$, where P_I^* is the expected proportion of type I trees among bootstrap replicates.

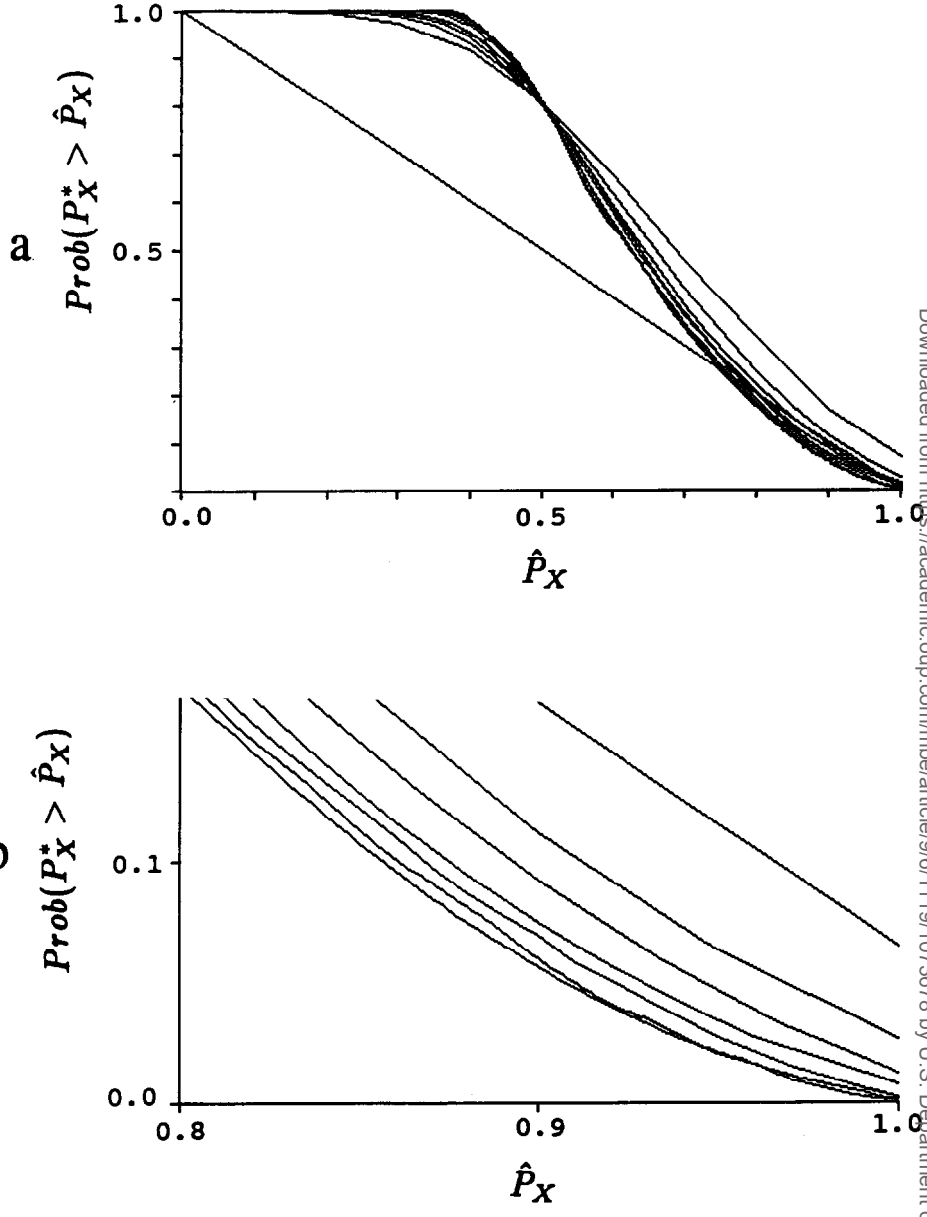

FIG. 10.—Probability of $P_X^{**} \geq \hat{P}_X$ inferred from simulating the case of a trifurcating tree with the same parameters as in fig. 3. The sequence length is $N = 100$. a, Entire region of 0.0–1.0 for $\hat{P}_X$. The straight line represents $1 - \hat{P}_X$. b, More detailed plot for $0.8 \leq \hat{P}_X \leq 1.0$. The numbers of bootstrap replications are (from the top curve to the bottom) N_b = 10, 20, 30, 50, 100, 300, and 1,000. Ten thousand simulation replicates were conducted for each curve.

replications, N_b. These plots are practically the same for different sequence lengths, so we show them only for $N = 100$. In table 3 the numerical estimates of this probability are also presented for several values of $\hat{P}_X$. As N_b increases (fig. 10), the plot for

Table 3
Prob($P_X^{**} \geq \hat{P}_X$), Inferred from Simulation with Various Numbers of Bootstrap Replications N_b

	Prob($P_X^{**} \geq \hat{P}_X$) FOR N_b OF				
$\hat{P}_X$	10	20	50	100	1,000
0.80	0.316	0.246	0.198	0.187	0.178
0.85	0.166	0.173	0.116	0.122	0.113
0.90	0.166	0.112	0.045	0.069	0.059
0.95	0.065	0.063	0.027	0.027	0.020
0.96	0.065	0.026	0.027	0.021	0.017
0.97	0.065	0.026	0.017	0.015	0.009
0.98	0.065	0.026	0.017	0.010	0.005
0.99	0.065	0.026	0.007	0.006	0.002
1.00	0.065	0.026	0.007	0.002	0.000

NOTE.—The parameters of the simulation are given in the legend to fig. 10.

Prob($P_X^{**} \geq \hat{P}_X$) gradually approaches the asymptotic plot for $N_b = \infty$, i.e., Prob($P_X^{*} \geq \hat{P}_X$). The straight line in figure 10a represents $1-\hat{P}_X$. Note that if $\hat{P}_X$ is high, say, ≥90%, then Prob($P_X^{**} \geq \hat{P}_X$) quickly becomes smaller than $1-\hat{P}_X$ as N_b increases. For example, if $N_b \geq 50$, then the probability for $P_X^{**} \geq 95\%$ is <5% (fig. 10b and table 3). Therefore, in the case of three taxa with one outgroup, if $\hat{P}_X = 95\%$, then the probability of accepting an erroneous tree as the true tree is <5% as long as $N_b \geq 50$. Note that in the region of $\hat{P}_X > 0.9$, the plot for $N_b = 300$ is practically the same as that for $N_b = 1{,}000$ (fig. 10a). This means that, in the case of three taxa with one outgroup, for estimating the confidence level of $\hat{P}_X \geq 90\%$, it is sufficient to use 300 bootstrap replications.

Figure 10a reveals also that, if $\hat{P}_X$<75%, then, even if $N_b = 1{,}000$, Prob($P_X^{**} \geq \hat{P}_X$) is larger than $1-\hat{P}_X$ or, in other words, the probability for P_X^{**} to be ≥75% is ≥25%. This is not surprising, because, as shown in the last section, the expected proportion of bootstrap replicates supporting an observed bifurcating tree is close to 66%, when the trifurcating tree is used as the model tree. An implication of these results is that, if a bifurcating tree is observed in less than, say, 75% of the bootstrap replicates, then one cannot claim that it is better than the trifurcating tree.

We now consider the distribution of P_X^{**} when tree I in figure 1 is used as the model tree. According to equation (20), for a bifurcating model tree, any given value of P_I can be reached by increasing the sequence length N. To study the sequence length required for P_X^{**} to be equal to or higher than a given value, we have conducted simulations with the parameters given in table 4. The number of bootstrap replications in each of these simulations is $N_b = 300$. In table 4 the $\bar{P}_I$ value was obtained from the average over 10,000 simulation replicates and therefore should be an accurate estimate of P_I, which is the probability that a random sample of sequences will support tree I. The first set of simulations (table 4a) demonstrates that, for very short sequences, e.g., $N \leq 40$ for the parameter values used in table 4a, the mean value of P_X^{**} ($\bar{P}_X^{**}$) is larger than P_I. The two values become equal when $N \approx 40$, i.e., $\bar{P}_X^{**} \approx P_I \approx 0.777$. For $N > 40$, $\bar{P}_X^{**}$ underestimates P_I.

The above phenomenon can be explained by considering P_I^* as a function of γ_I^* [eq. (28)]. For the expectation of a function of a random variable, $u(x)$, one can

Table 4

Average Values of P_I, P_{II}, and P_X^{**} ($\bar{P}_I$, $\bar{P}_{II}$, and $\bar{P}_X^{**}$), Inferred from Simulation by Using a Bifurcating Model Tree with Various Divergence Times (fig. 1a)

T_2, T_3	$N_{0.95}$	N	$\bar{P}_I$	$\bar{P}_{II}$	$\bar{P}_X^{**}$
$N_{0.5} < N < N_{0.95}$:					
50, 20	104	30	0.7040	0.0591	0.7588
50, 20	104	40	0.7766	0.0457	0.7769
50, 20	104	50	0.8295	0.0365	0.7962
50, 20	104	60	0.8595	0.0301	0.8146
50, 20	104	70	0.8965	0.0219	0.8327
50, 20	104	80	0.9198	0.0187	0.8505
50, 20	104	90	0.9393	0.0149	0.8646
50, 20	104	100	0.9508	0.0115	0.8772
50, 20	104	110	0.9611	0.0082	0.8907
50, 20	104	120	0.9691	0.0067	0.9028
$N = N_{0.95}$:					
50, 10	49	49	0.9637	0.0053	0.9071
50, 20	104	104	0.9528	0.0093	0.8847
50, 30	280	280	0.9495	0.0148	0.8691
50, 40	1,341	1,341	0.9485	0.0209	0.8742
50, 45	5,847	5,847	0.9467	0.0241	0.8673
30, 10	92	92	0.9614	0.0076	0.8922
30, 20	417	417	0.9521	0.0158	0.8706
30, 25	1,762	1,762	0.9527	0.0195	0.8672
80, 50	470	470	0.9495	0.0185	0.8731
80, 60	1,333	1,333	0.9481	0.0204	0.8694
80, 70	6,741	6,741	0.9533	0.0212	0.8654
$N = 1.7N_{0.95} \approx N_{0.99}$:					
50, 10	49	70	0.9902	0.0013	0.9438
50, 20	104	165	0.9893	0.0025	0.9396
50, 30	280	475	0.9895	0.0038	0.9417
30, 20	417	730	0.9907	0.0029	0.9407

NOTE.—In all cases T_1 = 100 Myr, whereas T_2 and T_3 are given in the table. The values of $N_{0.95}$ were estimated using eq. (8), (9), and (20) for $\mu = 10^{-8}$ and $\alpha = \beta$. N is the actual sequence length used in the simulation. The number of bootstrap replications N_b = 300. Ten thousand simulation replicates were conducted for each set of parameter values.

use the approximation $E(u(x)) = u(E(x)) + \frac{1}{2}u''(E(x))\mathrm{Var}(x)$. In this equation the first term on the right-hand side is $u(E(x)) = P_I^*(\bar{\gamma}_I^*) \approx P_I$ [fig. 5b and eq. (34)], and the factor $\mathrm{Var}(x) = \mathrm{Var}(\gamma_I^*)$ in the second term is positive. When $\bar{\gamma}_I^* < 0$, the second derivative of the function $P_I^*(\gamma_I^*)$ at $\gamma_I^* = \bar{\gamma}_I^*$ is positive, and so the expectation of P_I^*, i.e., $E(u(x))$, overestimates P_I, whereas, when $\bar{\gamma}_I^* > 0$, the second derivative at $\gamma_I^* = \bar{\gamma}_I^*$ is negative, and so the expectation of P_I^* underestimates P_I. The two values become equal when the sequence length N is approximately $N_{0.5}$, i.e., $\bar{\gamma}_I^* = 0$. As stated above [see eq. (35) and (36)], P_X^* is always greater than P_I^* and approaches P_I^* for large N. Since $\bar{P}_X^{**}$ is not far from $\bar{P}_X^*$ for $N_b \geq 300$, the condition $\bar{P}_X^{**} \geq P_I$ requires $N > N_{0.5}$. For several models, it was found that the condition holds when $N \approx N_{0.78}$. Therefore, if $N > N_{0.78}$, then P_X^{**} is expected to underestimate P_I.

When we take $N = N_{0.95}$, i.e., P_I = 95%, the corresponding values of $\bar{P}_X^{**}$ are 87%–89% (table 4b). In order to reach P_I = 99% and, correspondingly, $\bar{P}_X^{**}$ = 94%, N should be ~1.7 times larger (table 4c), i.e., $N = 1.7N_{0.95}$. This relation can be approximated by a simple equation:

$$E(P) \approx 1 - Ce^{-aN/N_{0.95}}, \tag{41}$$

where C and a can be estimated from the simulation data. For $P = P_I$, $C \approx 0.6$ and $a \approx 2.48$; and, for $P = P_X^{**}$, $C \approx 0.3$ and $a \approx 1.01$.

Note that this estimation of P_I and P_X^{**} depends only on the ratio $N/N_{0.95}$. This can be explained by considering eq. (19). Expressing β_P from this equation, we can write

$$\frac{\beta_P}{\beta_{0.95}} = \frac{\sqrt{N_P}\Delta p - \sqrt{p_{II}/\pi}}{\sqrt{N_{0.95}}\Delta p - \sqrt{p_{II}/\pi}}. \tag{42}$$

Since $\sqrt{N_{0.5}}\Delta p \approx \sqrt{p_{II}/\pi}$, we obtain

$$\beta_P = \beta_{0.95} \frac{\sqrt{N_P/N_{0.95}} - \sqrt{N_{0.5}/N_{0.95}}}{1 - \sqrt{N_{0.5}/N_{0.95}}}. \tag{43}$$

For most cases, $N_{0.95} \gg N_{0.5}$. Neglecting terms containing the ratio $N_{0.5}/N_{0.95}$, we get a simple formula,

$$\beta_P \approx \beta_{0.95} \sqrt{\frac{N_P}{N_{0.95}}}, \tag{44}$$

in which β_P depends only on the ratio $N_P/N_{0.95}$.

One of the important statistical properties of bootstrap estimation is the probability of failing to accept the true tree, Prob($P_I^{**} < \hat{P}_X$) (when tree I is used as the model tree), which is evidently greater than the probability of failing to accept any of the three alternative trees, Prob($P_X^{**} < \hat{P}_X$). The two probabilities become equal as N becomes large. On the basis of the results of simulation (table 5), we estimate these probabilities for the confidence level $\hat{P}_X = 0.95$. To characterize further the distribution of P_X^{**} we also consider a left cut-off point P_L that gives Prob($P_X^{**} < P_L$) < 0.05 (table 5). All these characteristics demonstrate that, unless N is very large, the distribution of P_X^{**} is wide and, hence, using P_X^{**} as a criterion for accepting a tree leads to a very high probability of failing to accept any bifurcating tree. For example, even if the expected value of P_I is as high as 99.6%, so that the expected value of P_X^{**} is 95.9% (N=2,800 in table 5), there is a 5% probability that P_X^{**} is <80%, and the probability of failing to accept any bifurcating tree, i.e., Prob($P_X^{**} < 0.95$), is >22%. Note that when the expected value of P_I is 99.6%, almost all samples of sequences from the evolutionary process will support tree I but that, nevertheless, in a substantial proportion, i.e., 24%, of the samples, the support for tree I is not strong enough for P_X^{**} to reach 95%. It is clear from table 5 that, for Prob($P_X^{**} < 95\%$) to be <5%, the sequence length required is at least three times (almost four times) longer than that required for $P_I = 95\%$. To understand the preceding conclusion, it is useful to consider the distribution of P_X^*, which is the distribution of P_X^{**} when $N_b = \infty$. For example, in figure 9d, $P_I = 95\%$, but the probability for $P_X^* \geq 95\%$ is less than the probability for $P_X^* < 95\%$.

Table 5
Characteristics of the Distribution of P_I^{**} and P_X^{**}, Estimated by Simulating the Bifurcating Model Tree (T_1=100, T_2=50, T_3=40, $\mu=10^{-8}$, and $\alpha=\beta$), with Various Sequence Lengths N

N	$\bar{P}_I$	P_I^{**}	$\bar{P}_X^{**}$	S_I	S_X	P_L
200	0.620	0.509	0.686	0.956	0.938	0.433
400	0.738	0.695	0.727	0.905	0.892	0.443
600	0.819	0.695	0.764	0.828	0.832	0.467
800	0.877	0.756	0.798	0.779	0.773	0.487
1,000	0.916	0.792	0.834	0.719	0.679	0.510
1,200	0.928	0.837	0.856	0.628	0.623	0.530
1,400	0.952	0.859	0.868	0.586	0.580	0.544
1,600	0.971	0.873	0.886	0.526	0.526	0.565
1,800	0.981	0.900	0.911	0.460	0.459	0.618
2,000	0.984	0.913	0.923	0.401	0.400	0.650
2,200	0.989	0.933	0.932	0.340	0.339	0.700
2,400	0.991	0.940	0.943	0.312	0.311	0.732
2,600	0.994	0.947	0.948	0.285	0.285	0.755
2,800	0.996	0.958	0.959	0.239	0.237	0.802
3,000	0.999	0.967	0.965	0.222	0.220	0.820
3,500	0.999	0.974	0.974	0.141	0.141	0.863
4,000	1.000	0.983	0.983	0.088	0.088	0.920
4,500	1.000	0.989	0.989	0.061	0.061	0.937
5,000	1.000	0.993	0.993	0.029	0.029	0.963

NOTE.—One thousand simulation replicates with N_b = 300 bootstrap replications were conducted for each sequence length. $\bar{P}_I$, $\bar{P}_I^{**}$, and $\bar{P}_X^{**}$ denote the mean values of P_I, P_I^{**}, and P_X^{**} over simulation replicates. S_I is Prob(P_I^{**}<0.95), S_X is Prob(P_X^{**}<0.95), and P_L is defined by Prob(P_X^{**}<P_L) ≈ 5%.

Discussion

In this study we have considered four taxa and have assumed a constant rate of nucleotide substitution. Under this simple situation, one can draw the following conclusion: As long as a reasonable number of bootstrap replicates (say, ≥100) have been conducted, considerable (≥80%) confidence can be given to a tree that is supported by >80% of the replicates. In particular, the probability that a tree is an erroneous one is <5%, if it is supported by ≥95% of the replicates. Thus, one is on the safe side if he or she sets 95% as the level for accepting a tree. On the other hand, little confidence can be given to a tree that is supported by ≤75% of the replicates, for in this case the tree cannot be claimed to be better than the trifurcating tree.

It should be emphasized that, under the ideal conditions assumed in this study, it is rather simple to identify the true tree. In practice, deviations from ideal conditions are likely to occur, and identifying the true tree can be very difficult. We discuss below the conditions assumed in this study.

First, let us consider the assumption of a constant rate of nucleotide substitution. There is now strong evidence that this assumption is violated in many evolutionary lineages (e.g., see Wu and Li 1985; Britten 1986; Seino et al. 1992). As pointed out by Felsenstein (1978, 1985), unequal rates of evolution can mislead parsimony inferences, and bootstrapping does not correct this problem. Therefore, under unequal rates of evolution, the probability of accepting an erroneous tree is likely to be higher than that given in the present study. For the effects of unequal rates on bootstrap estimation, readers may refer to Hillis and Bull (accepted) and Zharkikh and Li (accepted).

Second, we consider the assumption of homogeneous sequences in which all sites are variable and evolve at the same rate. This assumption may hold approximately for nonfunctional sequences or for sequences with very weak selective constrains, e.g., intergenic regions and pseudogenes. In functional sequences there may be sites that do not change with time. Since such invariable sites cannot become informative, they do not contribute to the sequence length N used in the above analysis; in our formulation, we assumed that all sites are variable. Therefore, in practice, the effective sequence length can be considerably shorter than the actual length. In theory, invariable sites should be excluded from analysis, though such sites are usually difficult to identify in practice. Another problem is that in functional sequences not all sites evolve at the same rate. Obviously, how rate heterogeneity may affect bootstrap estimation is worth studying.

Third, in many cases we have used a fairly high rate of nucleotide substitution, i.e., $u = 10^{-8}$ substitutions per site per year. This high rate was used to facilitate computations and simulations because it leads to many informative sites in a relatively short time of divergence. If the rate is lower, then the sequence length required for P_X^{**} to reach a given confidence level will be different.

As an example of application of the present results, let us consider the sequence data used by Li et al. (1992) for determining the phylogenetic position of the guinea pig. The four taxa they used are (1) guinea pig, (2) myomorphs (mice and rats), (3) primates, and (4) marsupials or aves as an outgroup. Among the 2,413 amino acid sites under study, there are 109 informative sites, of which 50, 29, and 30 support tree III, tree I, and tree II, respectively, where tree I represents the traditional view that the guinea pig and the myomorphs are sister groups, tree II puts the guinea pig and the primates in one clade, and tree III assumes that the guinea pig is an outgroup to the myomorphs and the primates and that it therefore does not belong to the order Rodentia. From the data, we have $p^*_{III} = 50/2{,}413 = 0.0207$, $p^*_{II} = 0.0120$, and $p^*_I = 0.0124$. Using formula (20), we estimate that the sequence length required for 95% probability of obtaining tree III is $N_{0.95} = 1{,}813$. From formula (20) one can show that, for $N = 2{,}413$, the probability of obtaining tree III is $P_{III} = 0.969$. Bootstrap estimation of P_{III}, from $N_b = 1{,}000$ bootstrap replications, gives 0.977. From figure 10, the probability for a bifurcating tree to appear in ≥ 0.977 of the bootstrap replicates is <0.009 if a trifurcating tree is used as the model tree. Taken at face value, this small probability supports Graur et al.'s (1991) hypothesis that the guinea pig is not a rodent; that is, tree III is the true tree. However, we must note the assumptions involved. First, it assumes equal rates among the primate, myomorph, and guinea pig lineages, but there is evidence that the rate of amino acid substitution is considerably lower in the primate lineage, though approximately the same in the other two lineages (Li et al. 1992). Second, it assumes that all amino acid residue sites are variable and evolve at the same rates, but it is likely that some sites have evolved faster than the others and that some sites are invariable. Therefore, the probability that tree III is erroneous can be substantially larger than 0.009, and the hypothesis needs to be reexamined using more sequence data.

Acknowledgments

We thank Yun-Xin Fu, Michael Bulmer, J. Felsenstein, and J. Bull for comments. We also thank D. Hillis for sending us a copy of the Hillis and Bull (1992) article before its publication. This study was supported by NIH grant GM30998.

LITERATURE CITED

BRITTEN, R. J. 1986. Rates of DNA sequence evolution differ between taxonomic groups. Science **231**:1393–1398.

EFFRON, B. 1982. The jackknife, the bootstrap, and other resampling plans. CBMS-NSF Regional Conference Series in Applied Mathematics, no. 38. Society for Industrial and Applied Mathematics. Philadelphia.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**:401–410.

———. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

———. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22**:521–565.

FLEISS, J. L. 1981. Statistical methods for rates and proportions. John Wiley & Sons, New York.

GRAUR, D., W. A. HIDE, and W.-H. LI. 1991. Is the guinea pig a rodent? Nature **351**:649–652.

HILLIS, D. M., and J. J. BULL. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. (accepted).

JOHNSON, N. I., and S. KOTZ. 1969. Distribution in statistics: discrete distributions. John Wiley & Sons, New York.

———. 1970. Distribution in statistics: continuous univariate distributions—1. John Wiley & Sons, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

LI, W.-H. 1986. Evolutionary change of restriction cleavage sites and phylogenetic inference. Genetics **113**:187–213.

LI, W.-H., and M. GOUY. 1991. Statistical methods for testing molecular phylogenies. Pp. 249–277 *in* M. M. MIYAMOTO and J. CRACRAFT, eds. Phylogenetic analysis of DNA sequences. Oxford University Press, New York.

LI, W.-H., and D. GRAUR. 1991. Fundamentals of molecular evolution. Sinauer, Sunderland, Mass.

LI, W.-H., W. A. HIDE, A. ZHARKIKH, D.-P. MA, and D. GRAUR. 1992. The molecular taxonomy and evolution of the guinea pig. J. Hered. **83**:174–181.

LI, W.-H., K. H. WOLFE, J. SOURDIS, and P. M. SHARP. 1987. Reconstruction of phylogenetic trees and estimation of divergence times under nonconstant rates of evolution. Cold Spring Harb. Symp. Quant. Biol. **52**:847–856.

NEI, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

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

SAITOU, N., and M. NEI. 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J. Mol. Evol. **24**:189–204.

SEINO, S., G. I. BELL, and W.-H. LI. 1992. Sequences of primate insulin genes support the hypothesis of a slower rate of molecular evolution in human and apes than in monkeys. Mol. Biol. Evol. **9**:193–203.

WU, C.-I., and W.-H. LI. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA **82**:1741–1745.

ZHARKIKH, A., and W.-H. LI. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock. J. Mol. Evol. (accepted).

BRIAN CHARLESWORTH, reviewing editor

Received January 17, 1992; revision received May 20, 1992

Accepted May 29, 1992