Syst. Biol. 43(3):329-342, 1994

STATISTICAL PROPERTIES OF THE MAXIMUM LIKELIHOOD METHOD OF PHYLOGENETIC ESTIMATION AND COMPARISON WITH DISTANCE MATRIX METHODS

ZIHENG YANG[1]

Department of Zoology, Downing Street, Cambridge CB2 3EJ, England

Abstract.—A proof is presented that the maximum likelihood (ML) method of phylogenetic estimation from DNA sequences (Felsenstein, 1981, J. Mol. Evol. 17:368-376) is statistically consistent despite the irregularity of the parameter space of the estimation problem. The distance matrix method using the least squares (LS) criterion is also consistent, but disconnection of two steps in the method, i.e., estimation of sequence divergence and construction of the tree topology, appears to lead to both theoretical contradictions and practical problems. Comparison of the ML and LS methods shows that the ML method is much more, indeed extremely, tolerant to violation of its assumptions and also has smaller sampling errors caused by limited data. This conclusion should be general, independent of particular models, tree topologies, and number of species in the data set. The problem of evaluating the reliability of the estimated tree topology was examined. The test of positivity of the interior branch length in the estimated tree is not a test of the significance of the ML tree but may be taken as a test of the significance of the LS tree. [Phylogenetic estimation; maximum likelihood; least squares; consistency; sampling error; robustness; parameter space; molecular systematics.]

Under regularity conditions, maximum likelihood (ML) estimators are well known to be consistent, asymptotically efficient, and normally distributed (e.g., Kendall and Stuart, 1979:39-64), that is, all the information in the data is taken into account and properly handled by the method. The application of the ML methodology to estimation of the phylogenetic tree from DNA sequences as suggested by Felsenstein (1981), however, has some apparent difficulties. For example, Nei (1987:325) wrote that

> the likelihood computed in this method is conditional for each topology, so that it is not clear whether or not the topology showing the highest likelihood has the highest probability of being the true topology when a relatively small number of nucleotides are examined.

Also "the likelihood function to be used varies from topology to topology, so that the ML values for different topologies are conditional and cannot be compared in the usual statistical sense" (Saitou, 1988:261; see also Li and Gouy, 1991).

Two problems are involved here. The first is essentially that of statistical consistency. By the likelihood method, the event that the ML tree is the true tree is equivalent to the event that the true tree has the highest likelihood, using the jargon of probability theory. If the method is consistent, that is, the ML tree is the true tree with probability 1 when the amount of data approaches infinity, we would expect that in a finite sample the ML tree will also have the highest probability of being the true tree. The second problem is whether the difference in likelihood between different trees can be used as a statistic in comparing trees.

The difficulties arise from the fact that in Felsenstein's (1981) formulation, the tree topology is treated as if it were a statistical parameter. In Goldman's (1990) model, the tree topology has been explicitly formulated as a parameter. Yet, the forms of the likelihood functions for different tree topologies are different. Branch lengths, which are themselves parameters to be estimated, depend on a specific tree topology, and branch lengths in one topology are meaningless in another. Some parameters are thus functions of others, and even the

[1] Present address: College of Animal Science and Technology, Beijing Agricultural University, Beijing 100094, China.

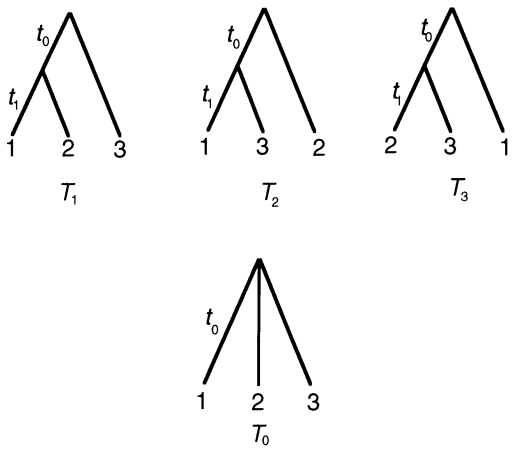

FIGURE 1. Possible (rooted) tree topologies with three species (1–3). T_0 is the star tree. Time or branch length (t_i) is measured by the expected number of nucleotide substitutions per site accumulated along the branch.

number of branch lengths can be different for different values of the tree parameter if we are comparing other than strictly bifurcating trees. The parameter space of the tree estimation problem is therefore very irregular.

Saitou (1988) performed theoretical studies and computer simulations to examine the efficiency of the ML method as compared with parsimony methods and distance matrix methods (see also Saitou and Imanishi, 1989). The ML method performed poorly in those simulations, but this result was later found to be due to the inefficiency of the computer program for the ML method used by those authors (Hasegawa et al., 1991). This interpretation is confirmed by our reanalysis of the primate mitochondrial DNA (mtDNA) sequence data analyzed by Saitou (1988), which shows that Saitou's program did not reach the maximum points. Nevertheless, the theoretical criticisms remain unanswered.

In this paper, I provide an explicit proof that the ML method of phylogenetic tree estimation is indeed consistent, despite the irregularity of the parameter space. Also, likelihoods of different trees are proper random variables and can be compared in the usual statistical sense. The ML method and a distance matrix method using the least squares (LS) criterion are compared in terms of their robustness to violation of their assumptions and their sampling errors.

PROTOCOLS OF THE ML AND LS METHODS

Suppose that the data consist of s homologous (gapless) DNA sequences, each of n nucleotides. The base compositions in different species at one site are called a site pattern, and there are 4^s possible site patterns (e.g., Goldman, 1993). Let the observed number of occurrences of the ith site pattern be $n_i = nf_i$. The likelihood function, which is proportional to the probability of a multinomial distribution, will have the following general form (e.g., Goldman, 1993):

$$L = \prod_{i=1}^{4^s} p_i^{n_i}, \qquad (1)$$

where p_i is the probability of observing the ith site pattern under the assumed substitution model and tree topology.

In this paper, the analysis makes use of $\ell = \log(L)/n$, which is the support per nucleotide site and can still be referred to as the (log-) likelihood. Tree comparisons are not influenced by this scaling.

Consider the case of three species. Assume the JC69 substitution model (Jukes and Cantor, 1969; Bishop and Friday, 1985), by which the rate of change between any two nucleotides is assumed to be the same, and assume a molecular clock, that is, the overall rate of change is the same along different lineages. There are then three bifurcating tree topologies, shown in Figure 1 as T_1, T_2, and T_3, together with the star tree T_0. This is the simplest case, but it already has all the complexities mentioned above. This case will be pursued in some detail to facilitate further analysis. The likelihood, after the scaling, is then

$$\ell = \log(L)/n = \sum_{i=1}^{64} f_i \log(p_i). \qquad (2)$$

With the JC69 model, probabilities of ob-

serving some of the site patterns, such as those for TTT, CCC, AAA, and GGG, are the same, and these patterns can be combined into one category. As a result, the 64 site patterns can be combined into only five categories (Saitou, 1988). These five categories can be labeled 0, 1, 2, 3, and 4. Category 0 consists of four site patterns with identical bases across species, i.e., TTT, CCC, AAA, and GGG, and category 4 consists of 24 site patterns with completely different bases, such as TCA, TCG, etc. Category 1 includes 12 site patterns, for which the bases in sequence 1 and 2 are the same and the base in sequence 3 is different, such as TTC and TTA. These patterns intuitively support tree topology T_1. Similarly, categories 2 and 3, each including 12 site patterns, support tree topologies T_2 and T_3, respectively.

Let p_i and f_i (i = 0, 1, 2, 3, 4) be the expected and observed frequencies of the ith category, respectively. The likelihood can thus be written

$$\ell = \sum_{i=0}^{4} f_i \log(p_i). \tag{3}$$

Equation 3 differs from Equation 2 by a constant. To be comparable with other substitution models, values obtained from Equation 2 are used, although Equation 3 is used to simplify calculation. The expected frequencies, p_i values, for tree topology T_1 as a function of t_0 and t_1 in T_1 (see Fig. 1) are given by the following equations:

$$\begin{aligned}
p_0(T_1, t_0, t_1) &= 4p_{(\mathrm{TTT})} \\
&= \frac{1}{16}(1 + 3b^2 + 6a^2b^2 + 6a^2b^3), \\
p_1(T_1, t_0, t_1) &= 12p_{(\mathrm{TTC})} \\
&= \frac{1}{16}(3 + 9b^2 - 6a^2b^2 - 6a^2b^3), \\
p_2(T_1, t_0, t_1) &= 12p_{(\mathrm{TCT})} \\
&= \frac{1}{16}(3 - 3b^2 + 6a^2b^2 - 6a^2b^3), \\
p_3(T_1, t_0, t_1) &= 12p_{(\mathrm{CTT})} \\
&= \frac{1}{16}(3 - 3b^2 + 6a^2b^2 - 6a^2b^3), \\
p_4(T_1, t_0, t_1) &= 24p_{(\mathrm{TCA})} \\
&= \frac{1}{16}(6 - 6b^2 - 12a^2b^2 \\
&\quad + 12a^2b^3),
\end{aligned} \tag{4}$$

where $a = \exp(-\frac{4}{3}t_0)$, $b = \exp(-\frac{4}{3}t_1)$. The factors 4, 12, 12, 12, and 24 sum to 64, indicating that the 64 site patterns are combined into five categories. Similar equations were given by Saitou (1988), but his formulae (4a–e) contained errors; they did not sum to 1. By symmetry, the p_i values for the other topologies, as functions of their own branch lengths, can be obtained by versions of equations of the same form.

Although the same notation is used for branch lengths in different trees, t_0 and t_1 in T_1 are different from those in T_2. In a more formal notation, different symbols should have been used for branch lengths in different trees. The p_i values and hence the likelihood functions are different from topology to topology. The role of the tree topology is to change the likelihood function and to specify branch lengths, and the role of the branch lengths is to change the values of the likelihood.

The ML method consists of maximization of ℓ for each of the tree topologies with respect to their branch lengths, thus leading to three likelihood values, for example, ℓ_1, ℓ_2, ℓ_3. The tree with the highest (maximum) likelihood is the ML tree.

The distance matrix method involves two steps. First, the distance, d_{ij}, between any two sequences i and j is estimated. With the JC69 model,

$$\hat{d}_{ij} = -\frac{3}{4}\log\left(1 - \frac{4}{3}p\right), \tag{5}$$

where p is the proportion of different sites between the two sequences. For $\hat{d}_{12}$, $p = f_2 + f_3 + f_4$.

The second step involves constructing the tree topology from the matrix of estimated pairwise distances. In this paper, an (unweighted) LS criterion was used to

compare different tree topologies. For example,

$$S_1 = (\hat{d}_{12} - 2\hat{t}_1)^2 + (\hat{d}_{13} - 2\hat{t}_0 - 2\hat{t}_1)^2 + (\hat{d}_{23} - 2\hat{t}_0 - 2\hat{t}_1)^2 \quad (6)$$

was minimized with respect to $\hat{t}_0$ and $\hat{t}_1$ in T_1 to calculate the sum of squares S_1 for tree topology T_1. The tree corresponding to the lowest S, the LS tree, was taken as the estimate of the true tree.

All assumptions made in the ML method concerning the process of nucleotide substitution were also made in the estimation of sequence distance. The second step of the LS method involves extra assumptions, some of which will be described later. In this study, the same substitution model was used in both methods so that they are directly comparable.

Justification of the ML Method of Phylogenetic Estimation

Consistency

Felsenstein (1978) stated that ML estimation of the phylogenetic tree is consistent but did not supply an explicit proof. However, the consistency of ML estimation of a tree topology does not follow directly from the consistency of ML estimation of a regular statistical parameter. The regularity conditions leading to consistency and other asymptotic properties of maximum likelihood estimators of regular statistical parameters, such as continuity and differentiability of the likelihood function with respect to parameters (e.g., Kendall and Stuart, 1979:42–45), are not satisfied in the tree estimation problem. Because different tree topologies have different likelihood functions and different sets of branch length parameters, the concepts of continuity and differentiability of the likelihood function with respect to the tree topology parameter do not seem to make sense. A formal proof that takes into account this complexity of the parameter space of the tree estimation problem is thus needed. Such a proof also provides insights into related problems, such as evaluation of models and trees.

The proof consists of two steps: (1) given the data, there is an upper limit on the likelihood which cannot be exceeded by any of the tree topologies; and (2) when the sequence length approaches infinity, the likelihood for the true tree will approach this limit, with its branch lengths approaching their true values, and thus the method will choose the true tree as the estimate.

Consider the simplest case described in the previous section; the proof does not depend on the model or tree topology. First, although the likelihood functions are different for different trees, they have the common form of Equation 2. The upper limit is obtained by taking all the expected frequencies, p_i's, as parameters, with the only restriction that $\Sigma p_i = 1$. Setting $\partial \ell / \partial \hat{p}_i = 0$ will give estimates of these parameters as $\hat{p}_i = f_i$, which means a perfect fit, and the likelihood for this unconstrained model (Navidi et al., 1991; Goldman, 1993) will be

$$\ell_{max} = \log(L_{max})/n = \sum_{i=0}^{4} f_i \log(f_i). \quad (7)$$

Second, suppose that the true tree is T_1 and the true branch lengths are t_0^* and t_1^*. When the sequence length approaches infinity, the observed site pattern frequencies will approach the expected ones, as given by Equation 4. That is, $f_i = p_i(T_1, t_0^*, t_1^*)$. When calculating ℓ_1, $\hat{t}_0 = t_0^*$, $\hat{t}_1 = t_1^*$ is the maximum point with $\ell_1 = \ell_{max}$.

Strictly speaking, it must be shown that ℓ_{max} cannot be reached by ℓ_2 or ℓ_3, i.e., the strict inequalities $\ell_2 < \ell_{max}$ and $\ell_3 < \ell_{max}$ hold, rather than $\ell_2 \leq \ell_{max}$ and $\ell_3 \leq \ell_{max}$. It is not clear whether it is possible to derive a rigorous proof of this, independent of the true tree topology and the substitution model. An intuitive argument holds that an exception to this assertion is highly unlikely. Because a wrong tree must provide perfect fit to data for its likelihood to reach ℓ_{max}, the problem is equivalent to that of existence or nonexistence of a root to a system of simultaneous equations that involves more equations than the number of variables. For the simplest case described above, it is easy to show that the wrong

trees can provide a perfect fit to data only if $t_0^* = 0$ or if $t_0^* = t_1^* = \infty$, which can be considered trivial and ignored. In this case, there are four independent equations (five observed frequencies for the categories with the sum to be 1) and two variables (two branch lengths in a wrong tree). When a more complex model is used or when more than three sequences are in the data set, the number of simultaneous equations increases much more rapidly than the number of variables, and a root to such a system is even more unlikely, that is, a wrong tree is not likely to give perfect fit to the data. For example, when the model of Hasegawa et al. (1985) instead of the JC69 model is assumed in the problem of Figure 1, there are six variables for a wrong tree (two branch lengths plus four free parameters in the substitution model) but $64 - 1 = 63$ independent equations.

The LS method as defined in the previous section is also consistent, because with infinitely long sequences the distances are accurately estimated and the sum of squares for the true tree will approach zero at the point $\hat{t}_0 = t_0^*$, $\hat{t}_1 = t_1^*$.

Comparison of Likelihood Values Across Tree Topologies

The above proof suggests two results concerning finite data. Suppose that T_1 is the ML tree. First, $\ell_{max} - \ell_1$, calculated from the real data, will be a natural statistic for evaluating the adequacy of the assumed substitution model. Without sampling errors, we should get perfect fit of the model to the data if the model is correct; with finite data the lack of fit caused by sampling errors can be tested statistically, as has been suggested by Navidi et al. (1991), Reeves (1992), and Goldman (1993). Second, $\ell_1 - \max(\ell_2, \ell_3)$ should be a natural statistic for evaluating the reliability of the ML tree. Without sampling errors, the likelihood of the true tree is greater than that of any of the others; with finite data, likelihood values for other topologies can be larger simply because of sampling errors.

Li and Gouy (1991) pointed out that the likelihood ratio test (χ^2 test) cannot be used to test likelihood differences among trees. This is the case even with the test of $\ell_{max} - \ell_1$. Apart from the difficulties in counting the number of degrees of freedom, there is also the problem of very few counts in many categories (Goldman, 1993). Contrary to Li and Gouy (1991), however, inapplicability of the χ^2 approximation is not evidence that likelihood differences cannot be examined by statistical tests at all; there are many known and unknown distributions other than the χ^2. Although the parameter space in the tree estimation problem is more complicated than in many statistical problems, it is nevertheless well defined. The number of tree topologies is fixed for a given data set, and the ranges of branch lengths are well specified. Suppose that T_1 in Figure 1 is the true tree. With given sample size, the frequencies of different site patterns follow the multinomial distribution. From such frequencies, the likelihood and branch lengths for each of the possible trees are determined: they are functions of these frequencies and thus are proper random variables with specific, if unknown, distributions. An example of the sampling distribution of the likelihood difference is given later (see Fig. 7).

Another question is whether a parametric test of the estimated tree topology can be constructed and, when it can, whether it can have higher power than nonparametric methods such as bootstrapping (Felsenstein, 1985), which appears to have low power (Zharkikh and Li, 1992; Hillis and Bull, 1993). Two tactics seem possible in this context. The first tactic is to take the tree topology as a statistical parameter with unknown dimension. The test now amounts to calculation of the variance of the estimated tree and construction of the confidence interval. If the confidence interval covers only the estimated topology, the result of the test is significant; otherwise it is not. This is exactly what the bootstrapping method is doing. In this formulation we may not expect a parametric method, if feasible, to be more powerful. The second tactic is to take different topologies as different hypotheses and regard the problem as one of hypothesis test-

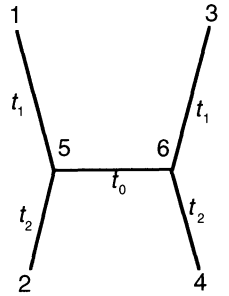

FIGURE 2. A model tree topology used in studying the robustness of tree estimation methods to violation of their assumptions and in examining the sampling errors of tree topology estimation. The branch length (t_1, t_2) is measured by the expected number of nucleotide substitutions per site accumulated along the branch.

ing. The null hypothesis should be that the estimated tree is not the true tree or one of the other topologies is the true tree, which seems very difficult to formulate if more than two trees are possibly true.

ROBUSTNESS OF THE METHODS TO VIOLATION OF THEIR ASSUMPTIONS

When discussing consistency, the model is assumed to be correct. However, a good estimator should also be robust to minor violation of its assumptions. In fact, the irregular nature of the tree estimation problem has made it possible for a method to be consistent even if a wrong model is assumed, as far as only estimation of the tree topology is concerned. When the assumed model is wrong, the method is usually consistent for only certain values of the parameters in the true model where the assumed model is not too wrong but not for all possible values of parameters. In other words, the method is consistent in only part of the parameter space, and this region may be termed the "consistency domain" of the method, the magnitude of which really measures the robustness of the tree estimation method to violation of its assumptions. When the true model is used, the consistency domain of the ML and LS methods is the whole parameter space, as shown by the proof.

Debry (1992) examined the robustness of parsimony methods to departure from the assumption of a molecular clock. Because neither the ML method nor the LS method requires this assumption, I examined two other aspects of the assumptions, i.e., the effect of transition/transversion bias and the effect of unequal nucleotide frequencies. Following Felsenstein (1978) and Debry (1992), consider the limiting case of infinite sequence length. The data, that is, the expected site pattern frequencies, are calculated under a more complex model, and then a simpler model is used to perform the analysis to see whether the true tree is recovered.

The more complex model is HKY85 (Hasegawa et al., 1985), which assumes unequal nucleotide frequencies and also different rates for transitional and transversional substitutions. The ratio of these two rates is designated κ, which is equivalent to α/β in the notation of Hasegawa et al. (1985). The F81 model (Felsenstein, 1981), which assumes $\kappa = 1$, was used to analyze the data to study the robustness to the κ bias, and the K80 model (Kimura, 1980), which assumes equal base frequencies ($\pi_T = \pi_C = \pi_A = \pi_G = 1/4$), was used to study the robustness to base frequency variation. Both are special cases of HKY85. The model tree topology used is shown in Figure 2, and two sets of branch lengths were used: (1) $t_1 = 0.5$, $t_2 = 0.1$; (2) $t_1 = 1.0$, $t_2 = 0.1$.

Robustness to Transition/Transversion Rate Bias

The equilibrium nucleotide frequencies used in the HKY85 model were $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$, and $\pi_G = 0.4$. Consider the case of $\kappa \geq 1$ only. Because the F81 model assumes $\kappa = 1$, the methods were expected to be consistent for small κ in the HKY85 model and to be inconsistent for large κ.

To study the relationship between the interior branch length t_0 in the model tree of Figure 2 and the robustness of the ML and LS methods to the κ bias, let t_0 change from 0.01, to 0.02, . . . , to 0.20. For each t_0, κ increases from 1, to 2, . . . , to 200, and the expected (observed) frequencies are generated using HKY85 with these values of t_0 and κ. The F81 model is then used to

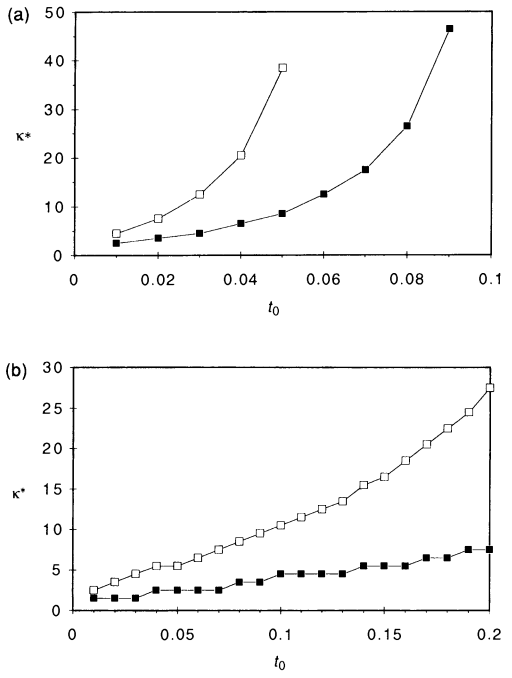

FIGURE 3. Robustness of the ML (□) and LS (■) methods of tree estimation to transition/transversion ratio (κ) bias. The HKY85 model is used to generate the data, with $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$, and $\pi_G = 0.4$. The F81 model is used to perform the analysis. The critical values (κ^*) are shown as a function of the interior branch length t_0 in the model tree of Figure 2. The method is consistent when $\kappa < \kappa^*$ but not when $\kappa > \kappa^*$. In other words, the region below the curve is the consistency domain of the tree estimation method. (a) The branch lengths are $t_1 = 0.5$, $t_2 = 0.1$. (b) The branch lengths are $t_1 = 1.0$, $t_2 = 0.1$.

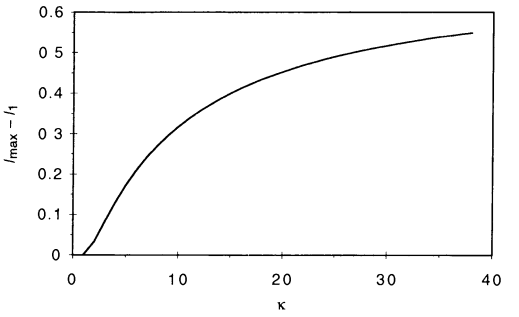

FIGURE 4. The lack of fit of the F81 model measured by $\ell_{max} - \ell_1$ as a function of κ in the HKY85 model. The $\ell_{max} - \ell_1$ value is the likelihood difference per site between the unconstrained model and the likelihood of the ML tree, calculated using the expected site pattern frequencies with infinitely long sequences. The model tree in Figure 2 is used with $t_0 = 0.05$, $t_1 = 0.5$, $t_2 = 0.1$ (see Fig. 3a). The true model is HKY85, with $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$, and $\pi_G = 0.4$.

estimate the best tree. The critical value κ^*, at which the method no longer gives the true tree, is then estimated as the average of the two neighboring κ values. The error of this estimate is thus no larger than 0.5.

The results are shown in Figure 3 for the two sets of branch lengths. Obviously ML is much more robust than LS. As expected, both methods are more robust to the ratio bias when the tree to be recovered is simpler, that is, when the first set of branch lengths were used. There are no cases where either of the two methods chooses the true tree when a κ larger than the critical value was used to generate the data. When the two methods chose the wrong tree, they both chose the tree with the two long branches together and the two short branches together, that is, the topology ((1, 3)2, 4) (see Fig. 2).

Figure 4 shows $\ell_{max} - \ell_1$, the likelihood difference between that of the unconstrained model and that of the best tree, as a function of the κ used for generating the data. The first set of branch lengths was used, with $t_1 = 0.5$, $t_2 = 0.1$, and $t_0 = 0.05$. The κ^* value for the ML method is 38.5 (Fig 3a). As the ratio κ becomes larger, the fit of the F81 model, as measured by the likelihood difference, becomes worse. Suppose that the observed values of $\ell_{max} - \ell_1$ in a finite sample are similar to the limiting values and the χ^2 approximation is used to test the significance (see Goldman, 1993, for criticisms). The unconstrained model has $4^4 - 1 = 255$ parameters. Counting the tree topology as one parameter, the F81 model has $5 + 3 + 1 = 9$ parameters. Using a normal approximation, $\chi^2_{0.01} \cong 306.0$, with df = 246. Therefore a sample size of $n = 306/(2 \times 0.172) = 890$ is enough for rejecting the F81 model when the real $\kappa = 5$. When there are more sequences in the data or the real κ is larger, the sample size needed for rejecting the F81 model is even smaller. The above calculations are very speculative. However, the qualitative conclusion is expected to be correct: our data

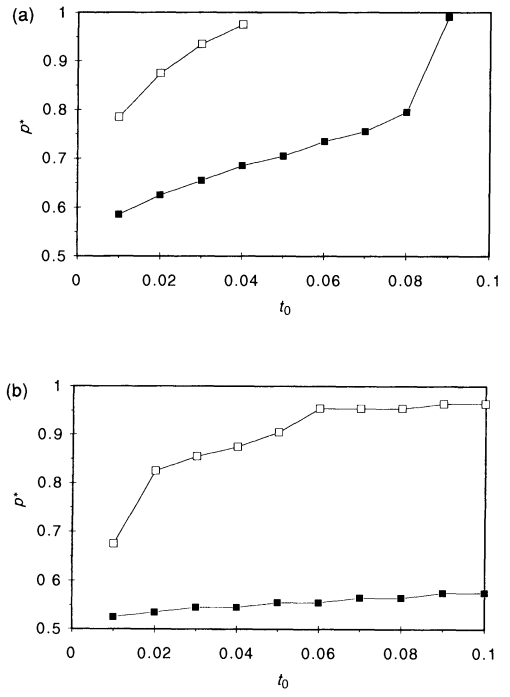

FIGURE 5. Robustness of the ML (□) and LS (■) methods of tree estimation to base frequency bias as reflected by p. The HKY85 model is used to generate data, with $\kappa = 10$ and nucleotide frequencies $\pi_T = \pi_A = p/2$, $\pi_C = \pi_G = (1 - p)/2$, and K80 is used to perform the analysis. The critical values (p^*) are shown as a function of the interior branch length t_0 in the model tree of Figure 2. The method is consistent when $p < p^*$. (a) The branch lengths are $t_1 = 0.5$, $t_2 = 0.1$. (b) The branch lengths are $t_1 = 1.0$, $t_2 = 0.1$.

set is typically large enough for rejecting a wrong model.

Robustness to Unequal Nucleotide Frequencies

Similarly, the HKY85 model was used to generate the data, with $\kappa = 10$. A variable p was introduced, such that the nucleotide frequencies in the HKY85 model were specified as $\pi_T = \pi_A = p/2$, $\pi_C = \pi_G = (1 - p)/2$. The K80 model used for analyzing the data was expected to be consistent when p is near to 0.5, and only the case of $p \geq 0.5$ was considered.

As before, t_0 changes from 0.01, to 0.02, . . . , to 0.20. For each t_0, p increases from 0.5 to 0.51, . . . , to 0.95. Site pattern frequencies are calculated using the HKY85 model with these values of branch lengths and nucleotide frequencies. Then the K80 model is used to estimate the best tree. The critical value p^* is determined in a way similar to that described above.

The results are shown in Figure 5 for the two sets of branch lengths. Once again, ML is much more robust than LS. For both sets of branch lengths, there are cases where LS chose the true tree when a p even larger than p^* was used to generate the data. This appears to mean that the consistency domain of a tree estimation method can consist of several disconnected regions in the parameter space. Figure 6 shows that the fit of the K80 model very rapidly becomes worse as p gets larger, as indicated by $\ell_{max} - \ell_1$.

SAMPLING ERRORS

The two methods were compared with respect to their sampling errors in the estimation of tree topology and branch lengths. Another important problem is how to evaluate the sampling error of the estimated tree topology by appropriate statistical tests.

Saitou (1988) attempted to derive the conditions under which the likelihood method will produce the true tree as the estimate. For the simplest case described before, he gave a proof that when T_1 is assumed to be the true tree, this condition is $n_1 > \max(n_2, n_3)$, or $f_1 > \max(f_2, f_3)$ (see Equations 3 and 4). This proof is invalid because it is based on the assumption that estimates of branch lengths are the same for the three trees, which can be met only if $f_1 = f_2 = f_3$, when the best tree is T_0 and the condition does not hold at all. Branch lengths in different tree topologies must be solved to maximize their own likelihood functions. My attempt to derive the single branch length in the star tree led to a seventh-order polynomial equation, for which an analytic solution is not possible. This probably means that analysis of the likelihood method must rely on intensive computation, as in this study.

Nevertheless, categories 1, 2, and 3, with observed frequencies f_1, f_2, and f_3, support the trees T_1, T_2, and T_3, respectively, and

any reasonable method should choose the tree corresponding to the largest f. Therefore, the condition $f_1 > \max(f_2, f_3)$ should be correct, as was confirmed by numerical calculations (results not shown).

With this condition, it is easy to calculate the probability that the ML tree will be the true tree given the length of the sequence. Suppose that T_1 in Figure 1 is the true tree with branch lengths $t_0 = 0.015$ and $t_1 = 0.05$ (values obtained from the 895-bp mtDNA sequences concerning human-chimpanzee-gorilla divergence). This probability is 0.95 when the sequence length is 650 bp, as calculated by the multivariate normal approximation of Zharkikh and Li (1992). Figure 7 shows the theoretical distribution of $\ell_1 - \max(\ell_2, \ell_3)$ for these values of parameters, as obtained from computer simulation. The distribution of $\ell_1 - \ell_2$, the likelihood difference between T_1 and T_2 in Figure 1, was very similar to that shown in Figure 7 (results not shown). Kishino and Hasegawa (1989) suggested that the distribution of the observed likelihoods can be approximated by a multivariate normal distribution. The test derived from such an approximation was unsatisfactory and appears to be conservative when compared with some standard tests in cases where standard tests can apply (results not shown). The skewness of the distribution shown in Figure 7 might explain why a normal distribution is not satisfactory. For the values of parameters of Figure 7, the distributions of estimates of branch lengths in the true tree by the ML and LS methods were very similar, and both were symmetrical (results not shown).

It seems that comparison of the methods with respect to sampling errors of tree estimation must be done by computer simulation. Fukami-Kobayashi and Tateno (1991) and Hasegawa et al. (1991) performed simulation studies to compare the efficiency of the ML method and distance matrix methods in recovering tree topologies. My small-scale simulations produced results similar to those obtained by those authors, in that the ML method has a higher probability of recovering the true tree than does distance matrix methods and thus has smaller sampling errors. Figure 8 shows one example, where the F81 model was used to generate data and also used by both methods to perform analysis. (The HKY85 model was not used because no simple formula is available to calculate pairwise distances under this model.)

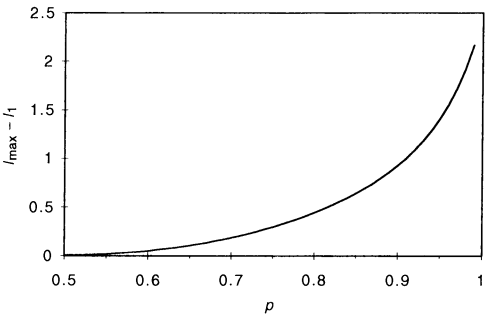

FIGURE 6. The lack of fit of the K80 model measured by $\ell_{max} - \ell_1$ as a function of p. The $\ell_{max} - \ell_1$ value is the likelihood difference per site between the unconstrained model and the likelihood of the ML tree in the limiting case of infinite sequence length. The true model is HKY85, with $\kappa = 10$ and $\pi_T = \pi_A = p/2$, $\pi_C = \pi_G = (1 - p)/2$. The model tree of Figure 2 is used, with $t_0 = 0.05$, $t_1 = 0.5$, $t_2 = 0.1$.

The cases in which estimates of branch lengths from the two methods are very similar are those where the tree is both the ML tree and the LS tree and where branch

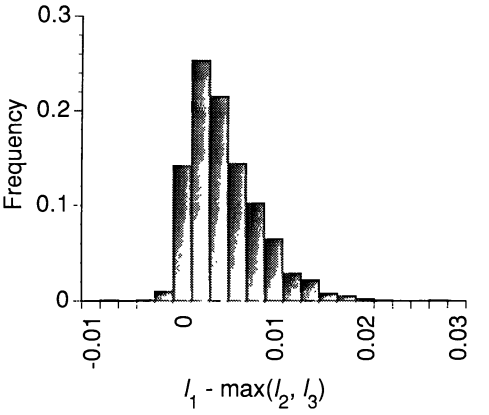

FIGURE 7. Sampling distribution of $\ell_1 - \max(\ell_2, \ell_3)$ estimated by computer simulation of 2,000 samples. The tree topology T_1 in Figure 1 is used to generate data, with branch lengths $t_0 = 0.015$ and $t_1 = 0.05$. The length of sequence is 650 bp, at which the probability that $\ell_1 - \max(\ell_2, \ell_3) > 0$ is 0.95.

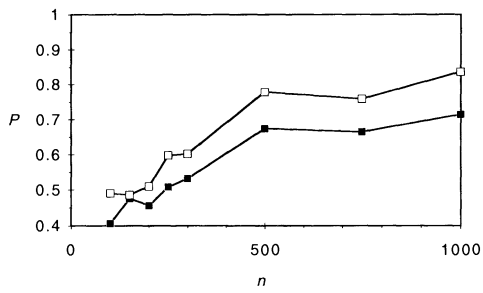

FIGURE 8. The probability that the true tree is recovered by the ML (□) and LS (■) methods of tree estimation as a function of the sequence length n, estimated by computer simulation of 500 samples. The true tree is shown in Figure 2 with $t_0 = 0.02$, $t_1 = 0.5$, and $t_2 = 0.1$. The F81 model is assumed to generate data, with $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$, and $\pi_G = 0.4$. The same model is used to perform analysis by both methods.

lengths are very small. These were the cases examined by Saitou (1988), who suggested that estimates of branch lengths by the two methods are then very similar. For other cases, estimates by the two methods are quite different. When the tree is neither the ML tree nor the LS tree, the LS method will very often produce zero (or negative) interior branch lengths, whereas the ML method may well give significantly positive values. The sampling properties of branch lengths in the wrong trees must be very different for the two methods, although it is not clear whether they are worth studying.

REANALYSIS OF THE 895-BP MTDNA SEQUENCES OF HUMAN AND APES

The 895-bp mtDNA sequences of human (H), chimpanzee (C), gorilla (G), orangutan (O), and gibbon (B) (Brown et al., 1982) were compared to reveal some differences between the two methods. The same data set was analyzed by Saitou (1988).

To apply the LS method, the K80 model was used to estimate pairwise distances. This model leads to simple formulae for ML estimation of two parameters in pairwise comparisons, that is, the average number of nucleotide substitutions per site, d, and the transition/transversion rate ratio, κ. The κ value, although a part of the model assumptions, has normally been ignored. The results are provided in Table 1. Use of the model of Tamura and Nei (1993), which is more complex than HKY85 or K80, led to slightly higher estimates of these two parameters.

Table 1 reveals an assumption underlying distance matrix methods. Suppose that the assumed tree is the one in Figure 9. From the data in Table 1, it appears that different values for individual parameters are being assumed at different stages of one analysis of the same data. For example, the ratio κ along the branch leading to human is estimated as 32.4, 23.3, 7.1, and 6.0 when the human sequence is compared with that of chimpanzee, gorilla, orangutan, and gibbon, respectively. The problem is more serious when a more complex model is used. For example, when the model of Tamura and Nei (1993) is used, both nucleotide frequencies and the two transition/transversion rate ratios will be subject to such contradictions. In the same vein, the length of a single branch in the tree is also assumed to have different values in different comparisons; this problem exists for a model as simple as JC69 and is more serious when a molecular clock is assumed.

Although this internal inconsistency does not appear with infinitely long sequences, it does make distance matrix

TABLE 1. Number of nucleotide substitutions per site and transition/transversion rate ratio (in parentheses) estimated assuming the K80 model (895-bp mtDNA sequences from Brown et al., 1982).

	Human	Chimpanzee	Gorilla	Orangutan
Chimpanzee	0.097 (32.37)			
Gorilla	0.114 (23.28)	0.118 (21.24)		
Orangutan	0.185 (7.08)	0.201 (8.17)	0.195 (8.17)	
Gibbon	0.212 (5.99)	0.223 (6.63)	0.223 (6.42)	0.223 (6.42)

methods more sensitive to sampling errors. This may be the main reason for the paradox, found in computer simulations, that, even though the data are generated using the JC69 model, using the corrected sequence divergence estimates ($\hat{d}$ in Equation 5) in distance matrix methods can be worse than using the uncorrected proportion of different sites (p in Equation 5) when the overall distances among the sequences are small (Jin and Nei, 1990; Nei, 1991). This has been explained as p having a smaller sampling variance than $\hat{d}$. The interpretation is not justifiable, however, because comparison of sampling variances is meaningful only when the estimators are consistent and unbiased, and p is not a consistent or unbiased estimator of d, whereas $\hat{d}$ is (N. Goldman, pers. comm.).

Nevertheless, Table 1 suggests that κ might indeed be different along different branches in the tree. Such a model was fitted using the ML method. The HKY85 substitution model was adopted, but one rate ratio (κ) was assigned to each branch. The likelihood is $\log(L) = -2652.34$, whereas a model assuming the same ratio for all the branches gives $\log(L) = -2665.42$. The asymptotic χ^2 test involved comparison of $2 \times 13.08 = 26.16$, with $\chi^2_{0.01} = 16.81$ and df = 6. The difference was significant, and indeed the ratio was different along different branches. The three branches leading to human, chimpanzee, and gorilla have higher κ ratios, in the range of 33–36, whereas all the others are around 5–8. The ML estimates of branch lengths and the κ parameter are shown in Figure 9.

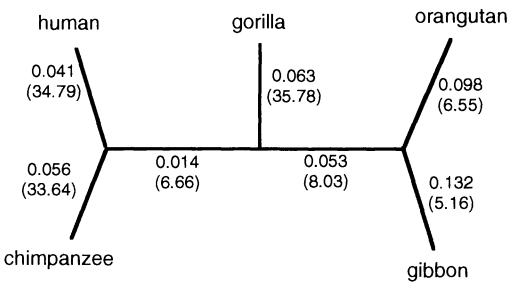

FIGURE 9. The maximum likelihood tree of the 895-bp mtDNA sequences of human, chimpanzee, gorilla, orangutan, and gibbon. The HKY85 model is used with one transition/transversion rate ratio (κ) assigned to each branch. The estimated branch lengths are given, with the rate ratios in parentheses. The likelihood value for this tree is $\log(L) - \log(L_{max}) = (-2652.34) - (-2476.97) = -175.37$.

DISCUSSION

Test of Phylogeny and Test of Branches

Two main methods have previously been suggested for testing the reliability of the estimated tree topology. One involves comparison of the best tree with all the other trees, and the other is a test of positivity of the interior branch lengths in the estimated tree (Nei et al., 1985; Li, 1989). Li and Gouy (1991) suggested that the first test might be too stringent, whereas Kishino and Hasegawa (1989) pointed out that the second is not stringent enough. The first test is indeed an evaluation of the estimated tree, i.e., a test of phylogeny. Therefore, I refer to the second test as a test of branches. It is worthwhile to consider whether a test of branches is equivalent to a test of phylogeny.

This question can be answered by considering the limiting case of infinite sequence length. The underlying assumption is that, if interior branch lengths in a wrong tree cannot be positive in the limiting case, they cannot be significantly positive in a finite sample, and thus a test of branches will be equivalent to a test of phylogeny. Otherwise, if a wrong tree can be better than the star tree in the limiting case, it can be significantly better than a star tree in a finite sample, and then a test of branches cannot be considered an evaluation of the estimated tree topology. In the following, this strategy was applied to the ML and LS methods of tree estimation. The true model was assumed so that both methods satisfy the requirement of consistency over the whole parameter space.

With the ML method, more often than not more than two bifurcating trees are better than the starlike tree(s) in the limiting case of infinitely long sequences, and thus the two tests are not equivalent.

With the LS method, sequence divergence is accurately estimated in the lim-

iting case and the analysis will be independent of the assumed substitution model. There is then a linear relationship between the true branch lengths in the true tree and branch lengths in any of the wrong trees, if negative branch lengths are allowed. By the LS criterion, negative (interior) branch lengths in a bifurcating tree mean that the tree cannot be better than the star tree if branch lengths are restricted to be positive. Therefore, estimates of branch lengths can be obtained in the wrong trees without restricting the branch lengths to be positive, and then the results can be converted into those with such a restriction. Accordingly, the following cases were examined: three species with a molecular clock, four species with a clock, and four species without a clock. In no case was a wrong tree better than the star tree in the limiting case. The generality of these results remains unknown, but it seems highly likely that this conclusion is general. This result suggests that the reliability of an LS tree may be examined by testing whether it is significantly better than the star tree, i.e., whether the interior branch length(s) are significantly positive. The reasons for this difference between the ML and LS methods are unknown.

Evaluation of the LS Method

Compared with the LS method based on pairwise distances, the ML method is more robust to violation of its assumptions and has smaller sampling errors, in that it has higher probability of recovering the true tree when a relatively short sequence is examined. Also, the LS method involves some internal contradictions, which must be part of the reason for its poorer performance. Other problems involved with the LS method can be explored in a further attempt to understand how the two methods differ.

First, when the distance between two sequences is estimated, all the other sequences are ignored. However, the distance between two sequences is not only reflected in the two sequences compared but also in how closely each of them is related to other sequences. All the sequences in the data provide information concerning the evolutionary process and, if made use of, would produce more accurate estimation of the pairwise distance. When the sequence data are transformed into the distance matrix, some information is lost, as pointed out by Penny et al. (1992; cf. Nei, 1991).

Second, it is well known that the larger the true distance, the larger the sampling error of the estimate and the more sensitive the estimate to the model assumptions underlying the estimation method (e.g., Gojobori et al., 1982; Nei, 1991). By adding the branch lengths along the tree, pairwise distances can be very large even though all the branches in the tree are small. Sometimes pairwise distances can be so large that the estimation formulae are not applicable (e.g., Gojobori et al., 1982).

Third, the LS criterion applied to the estimated pairwise distances instead of the real data may not be reasonable. A (weighted) LS criterion applied to the real data, i.e., the observed site pattern frequencies, will have a form similar to the following:

$$S = \sum_{i=1}^{4^s} \frac{(nf_i - np_i)^2}{nf_i(1 - f_i)}. \quad (8)$$

The likelihood method uses the ML criterion to get the best fit to the data,

$$\log(L) - \log(L_{\max}) = \sum_{i=1}^{4^s} nf_i \log \frac{p_i}{f_i}, \quad (9)$$

which is very similar to the minimum X^2 criterion

$$X^2 = \sum_{i=1}^{4^s} \frac{(nf_i - np_i)^2}{np_i}. \quad (10)$$

Results from the three criteria should be very similar, although neither Equation 8 nor Equation 10 can lead to simpler calculations than can the ML method (Equation 9).

All the problems involved with the LS method are not specific to models, tree topologies, or the amount of data. All the problems are caused by the two-step strategy and thus are completely avoided by the ML method, which performs a joint comparison of all the sequences (Bishop

and Friday, 1985). Therefore, the ML method should be generally better than the LS method. It seems difficult to imagine a specific case where an LS method would perform better than the ML method, as long as the sequences were related by a tree structure, even if the assumptions of the two methods were violated.

ACKNOWLEDGMENTS

Adrian Friday and Nick Goldman contributed greatly to this paper through their stimulating discussions. Both of them read an earlier version of the manuscript and gave many constructive and critical comments, which led to improvement of the presentation of the material in this paper. I am also grateful for the critical comments of J. Felsenstein and J. Slowinski, who acted as referees.

REFERENCES

BISHOP, M. J., AND A. E. FRIDAY. 1985. Evolutionary trees from nucleic acid and protein sequences. Proc. R. Soc. Lond. B 226:271–302.

BROWN, W. M., E. M. PRAGER, A. WANG, AND A. C. WILSON. 1982. Mitochondrial DNA sequences of primates: Tempo and mode of evolution. J. Mol. Evol. 18:225–239.

DEBRY, R. W. 1992. The consistency of several phylogeny-inference methods under varying evolutionary rates. Mol. Biol. Evol. 9:537–551.

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

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

FUKAMI-KOBAYASHI, K., AND Y. TATENO. 1991. Robustness of maximum likelihood tree estimation against different patterns of base substitution. J. Mol. Evol. 32:79–91.

GOJOBORI, T., K. ISHII, AND M. NEI. 1982. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotides. J. Mol. Evol. 18:414–423.

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

GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

HASEGAWA, M., H. KISHINO, AND N. SAITOU. 1991. On the maximum likelihood method in molecular phylogenetics. J. Mol. Evol. 32:443–445.

HASEGAWA, M., H. KISHINO, AND T. YANO. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

HILLIS, D. M., AND J. J. BULL. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182–192.

JIN, L., AND M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogeny analysis. Mol. Biol Evol. 7:82–102.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–123 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

KENDALL, M., AND A. STUART. 1979. The advanced theory of statistics, 4th edition, Volume 2. Charles Griffin, London.

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.

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

LI, W.-H. 1989. A statistical test of phylogenies estimated from sequence data. Mol. Biol. Evol. 6:424–435.

LI, W.-H., AND M. GOUY. 1991. Statistical methods for testing molecular phylogenies. Pages 249–277 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

NAVIDI, W. C., G. A. CHURCHILL, AND A. VON HAESELER. 1991. Methods for inferring phylogenies from nucleotide acid sequence data by using maximum likelihood and linear invariants. Mol. Biol. Evol. 8:128–143.

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

NEI, M. 1991. Relative efficiencies of different tree-making methods for molecular data. Pages 90–128 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

NEI, M., J. C. STEPHENS, AND N. SAITOU. 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. 2:66–85.

PENNY, D., M. D. HENDY, AND M. A. STEEL. 1992. Progress with methods for constructing evolutionary trees. Trends Ecol. Evol. 7:73–79.

REEVES, J. H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J. Mol. Evol. 35:17–31.

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

SAITOU, N., AND T. IMANISHI. 1989. Relative efficiencies of the Fitch-Margoliash, maximum parsimony, maximum likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol. Biol. Evol. 6:514–525.

TAMURA, K., AND M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526.

ZHARKIKH, A., AND W.-H. LI. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Mol. Biol. Evol. 9:1119–1147.

Received 3 June 1993; accepted 2 February 1994