Mol Evol (1995) 40:689–697

JOURNAL OF MOLECULAR EVOLUTION

Evaluation of Several Methods for Estimating Phylogenetic Trees When Substitution Rates Differ over Nucleotide Sites

Ziheng Yang*

Department of Zoology, The Natural History Museum, London SW7 5BD, United Kingdom

Received: 4 May 1994 / Accepted: 25 November 1994

Abstract. Several maximum likelihood and distance matrix methods for estimating phylogenetic trees from homologous DNA sequences were compared when substitution rates at sites were assumed to follow a gamma distribution. Computer simulations were performed to estimate the probabilities that various tree estimation methods recover the true tree topology. The case of four species was considered, and a few combinations of parameters were examined. Attention was applied to discriminating among different sources of error in tree reconstruction, i.e., the inconsistency of the tree estimation method, the sampling error in the estimated tree due to limited sequence length, and the sampling error in the estimated probability due to the number of simulations being limited. Compared to the least squares method based on pairwise distance estimates, the joint likelihood analysis is found to be more robust when rate variation over sites is present but ignored and an assumption is thus violated. With limited data, the likelihood method has a much higher probability of recovering the true tree and is therefore more efficient than the least squares method. The concept of statistical consistency of a tree estimation method and its implications were explored, and it is suggested that, while the efficiency (or sampling error) of a tree estimation method is a very important property, statistical consistency of the method over a wide range of, if not all, parameter values is prerequisite.

**Present address:* Z. Yang, Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, 328 Mueller Laboratory, University Park, PA 16802, USA

Key words: Phylogeny — Maximum likelihood — Least squares — Consistency — Sampling error — Rate variation at sites — Gamma distribution — Computer simulation

Introduction

For DNA sequences that code for biological products or are otherwise functional, rates of nucleotide substitution are expected to vary among sites due to different selectional constraints at different sites. Ignoring rate variation at sites when it exists can drastically affect some aspects of phylogenetic analysis, such as estimation of sequence divergence (e.g., Gillespie 1986; Takahata 1991) or estimation of evolutionary parameters such as the transition/transversion rate ratio (Wakeley 1994; Yang et al. 1994, in press). As likelihood values calculated for any tree topology are quite different depending on whether or not rate variation over sites is accounted for in the model (Yang et al. 1994), variable rates among sites may also affect the estimation of the tree topology.

Several distance measures between two sequences have been proposed that use a gamma distribution to model variable rates at sites (Jin and Nei 1990; Li et al. 1990; Tamura and Nei 1993). The gamma distribution involves a shape parameter α, which is inversely related to the extent of rate variation at sites (var = $1/\alpha$; the scale parameter of the distribution is chosen such that the mean is one). As there was no satisfactory approach to estimating this parameter, arbitrary values, such as 2, 1, or 0.5, were suggested (Jin and Nei 1990; Li et al. 1990).

Some recent analyses used estimates of the α parameter obtained from the parsimony analysis of the sequence data (e.g., Tamura and Nei 1993). This method is known to produce overestimated α (Wakeley 1993), and the bias can be very large; for example, in analyzing DNA sequences of about 3.5 kb from several primate mitochondrial genomes, which contain the first and second codon positions in six protein-coding regions and 11 tRNA-coding regions, Takezaki (personal communication) found that the estimate of α by the parsimony method (0.82) is over five times higher than that by the maximum likelihood estimation (0.16). Calculation of the pairwise distances appears to be very important to the performance of distance matrix methods (e.g., Jin and Nei 1990), and the use of arbitrary values or unreliable estimates of α is unsatisfactory.

The gamma distribution model for rates at sites has also been extended to a joint likelihood analysis (Yang 1993), which also provides reliable estimation of the α parameter for use in distance estimators. As this method is computationally very slow, two approximate methods were suggested by Yang (1994b). The approximations appeared to be reasonable when applied to real data, but their statistical properties have not been rigorously evaluated.

In this paper, several tree estimation methods will be examined in the presence of rate variation among sites. Two criteria were regarded as useful by Yang (1994a) and Yang et al. (in press) for evaluating the performance of a tree estimation method. The statistical consistency of a method can be examined by considering the limiting case of infinitely long sequences (Felsenstein 1978; see also Hendy and Penny 1989; Debry 1992; Zharkikh and Li 1993; Takezaki and Nei 1994). Yang (1994a) suggested a formal proof that the maximum likelihood method and the distance matrix method based on the (ordinary) least squares criterion are consistent when the true model is assumed in the analysis. When all assumptions underlying a method are not satisfied, the method is usually consistent for some, but not all, combinations of parameter values; that is, it is consistent within only a portion of the parameter space. This region was termed the "consistency domain" of the method by Yang (1994a), the size of which indicates the robustness of the method. Furthermore, the efficiency of a tree estimation method, or its sampling error, can be measured by the probability that an estimated tree is the true tree. This can be examined using computer simulations (e.g., Hasegawa and Yano 1984; Jin and Nei 1990; Fukami-Kobayashi and Tateno 1991; Hasegawa et al. 1991). The two criteria were used by Huelsenbeck and Hillis (1993) in comparing several parsimony and distance matrix methods in various combinations of branch lengths in a four-species tree. In this study, computer simulations will be performed to examine the sampling errors of several tree estimation methods, with attention paid to their statistical consistency.

A second objective of this study is to accumulate further results concerning the similarities and differences of phylogenetic tree estimation in comparison with the estimation of a conventional statistical parameter. This follows the previous attempts by Yang (1994a) and Yang et al. (in press) to understand the peculiar nature of the parameter space of the tree estimation problem.

Methods

Model of Nucleotide Substitution and Monto Carlo Generation of Sequence Data. The model of nucleotide substitution implemented in J. Felsenstein's DNAML program in the PHYLIP package is assumed, by which the probability that nucleotide i will change into j in a small time interval Δt is given as

$$Q_{ij}\Delta t = \begin{cases} (1 + \kappa/\pi_{\#})\pi_j \Delta t, & \text{for transitions } (T \leftrightarrow C,\ A \leftrightarrow G) \\ \pi_j \Delta t, & \text{for transversions } (T, C \leftrightarrow A, G) \end{cases} \quad (1)$$

where π_j is the frequency of nucleotide j when the process is in equilibrium, with $\pi_{\#} = (\pi_T + \pi_C)$ if j is T or C, or $\pi_{\#} = (\pi_A + \pi_G$ if j is A or G (Hasegawa and Kishino 1989; Kishino and Hasegawa 1989). The diagonals of the rate matrix $\mathbf{Q} = \{Q_{ij}\}$ are specified by the requirement that row sums of $\mathbf{Q}$ are 0. The matrix is multiplied by a constant such that the average rate of substitution is one, i.e., $-\sum_i \pi_i Q_{ii} = 1$. This means that time t, or the branch length in a tree, is measured by the expected number of nucleotide substitutions per site that have occurred during the time period or along the branch. Parameter κ is the transition/transversion rate ratio; a κ larger than 0 will allow transitions to occur with higher rates than transversions. The model was designated "F84" by Yang (1994b), and formulae for estimating sequence distances under this model, either assuming a gamma distribution of rates at sites or a single rate for all sites, were given by Yang (1994b).

The overall rate of substitution for a site is assumed to be a random variable drawn from a gamma distribution with shape parameter α. The model is then referred to as "F84+Γ." Only the case of four species is considered. The existence of a molecular clock (i.e., rate constancy over lineages) is not assumed, and then the model does not permit identification of the root in a tree (Felsenstein 1981). Three (unrooted) bifurcating trees are then possible for four species. They are shown in Fig. 1, and T_1 is taken as the true tree used for generating data. Parameters in the model include $\theta = \{\pi_T, \pi_C, \pi_A, \kappa, \alpha\}$ and branch lengths $\mathbf{t}_1 = \{t_{11}, t_{12}, t_{13}, t_{14}, t_{15}\}$ in the true tree T_1. In this formulation, branch lengths in the true tree are regarded as parameters (together with θ), but those in the wrong trees are not.

Theoretically, given the true tree topology T_1, the values of parameters θ and $\mathbf{t}_1$, and the tree estimation method, there will exist a probability $P_i(n)$ that tree T_i will be the best estimate in a sample of sequences of length n, with $\sum_i P_i(n) = 1$. The most important of these probabilities is $P_1(n)$, which is the probability that the method recovers the true tree. It would be good if we could calculate $P_i(n)$ algebraically as a function of the parameters and sequence length. Unfortunately, this does not seem possible, at least for the likelihood method. The approach of computer simulation is therefore adopted, by which many samples of data are generated under the substitution model using given values of parameters and analyzed by the tree estimation method to see which tree is the best estimate, and the observed proportions are taken as estimates of the probabilities (e.g., Hasegawa and Yano 1984; Jin and Nei 1990; Fukami-Kobayashi and Tateno 1991; Hasegawa et al. 1991; Huelsenbeck and Hillis 1993).

Because of the computational burden of maximum likelihood parameter estimation by iteration, we are only able to study a few specific cases. Two values (0.2 and 0.8) were considered for α. Estimates of α from real data using the maximum likelihood approach have been in the

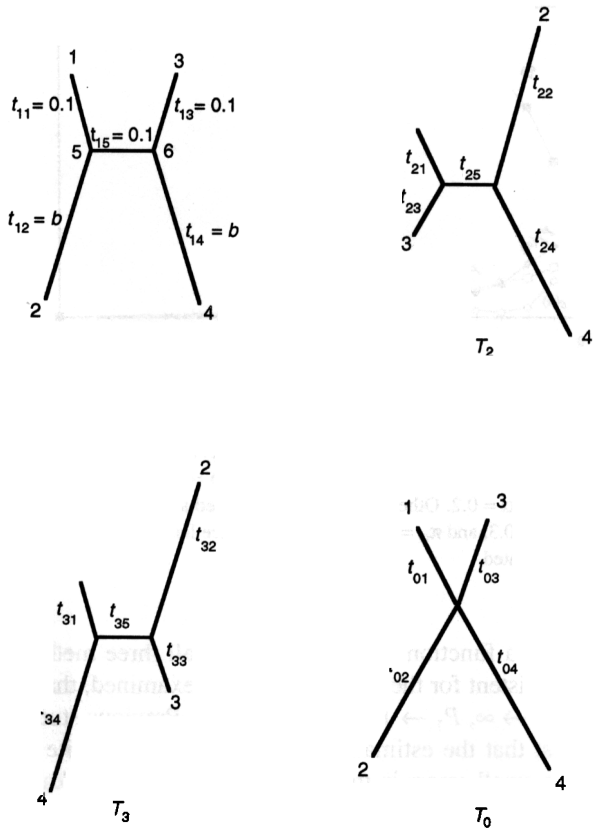

Fig. 1. The three (unrooted) bifurcating trees (T_1, T_2, T_3) and the star tree (T_0) for four species. T_1 was used to generate data, and two values (0.5 and 1.0) were used for branch length b. Other branch lengths in T_1 were fixed at the values shown. Branch lengths are measured by the average numbers of nucleotide substitutions per site accumulated along the branches. Branches in the wrong trees (T_2, T_3, T_0) are drawn to reflect their estimates from simulated data.

range 0.1–1.0 (Yang 1994b; Yang et al. 1994, and unpublished results), and 0.2 may represent severe rate variation while 0.8 may be an example of little rate variation. Two values (0.5 and 1.0) were used for branch length b in the true tree T_1 of Fig. 1; the former means the tree is "simpler" to estimate than the latter. Other parameters in the model were fixed at $\kappa = 5$, $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$, and $\pi_G = 0.4$. To study the behavior of $P_i(n)$ as functions of n, five sequence lengths (100, 200, 500, 1,000, 2,000) were examined for each of the four combinations of parameters b and α.

The accuracy of the estimate of the probability $P_i(n)$ for given values of parameters depends on the number of repetitions of simulations, r. The standard error of the estimate is largest when the real $P_i(n)$ is near 0.5, where $\sigma = (0.5 \times 0.5/r)^{1/2} = 0.05$, at $r = 100$, which seems too crude. We have used $r = 200$, for which $\sigma = 0.035$ at $P_i(n) = 0.5$, a slight improvement. In sum, $2 \times 2 \times 5 \times 200 = 4{,}000$ data sets were simulated and each was analyzed using various tree estimation methods.

Specifically, n independent random variables were drawn from a gamma distribution with parameter α and these were used as rates for sites. A random sequence of n nucleotides was generated, say the sequence for node 5 in T_1 of Fig. 1, using the equilibrium nucleotide frequencies. Each site in the sequence was then independently "evolved" under the F84 model of substitution along the tree according to the specified branch lengths and the rate for the site, producing sequences for nodes 1 and 2, and a sequence for node 6, from which sequences for nodes 3 and 4 were generated. Sequences for the end nodes in T_1, i.e., those for nodes 1, 2, 3, and 4, constitute the data, to be used for tree reconstruction.

Methods for Phylogenetic Tree Estimation Examined in This Paper. Five methods (models), based either on a joint likelihood analysis of all sequences or on the least squares method applied to the matrix of estimated pairwise distances, were examined:

1. *MLdG.* The "discrete gamma" model (Yang 1994b) uses K equal-probability categories of rates to approximate the (continuous) gamma distribution. The mean of each category is used to represent all rates in the category. We use $K = 4$, as suggested by Yang (1994b). Parameters κ and α and branch lengths are estimated from the data for each tree topology, and in the meantime, the (maximum) likelihood for the tree is obtained. The tree with the highest (maximum) likelihood is the best estimate. Due to computational reason, the method of Yang (1993) which assumes a continuous gamma model for rates at sites is not used in this study, and *MLdG* is closest to the true model assumed to generate data. Results obtained from analyzing several real data sets suggest that the performance of *MLdG* is very close to that using the true model, F84+Γ (Yang 1994b).
2. *MLfr.* The "fixed rates" approach (Yang 1994b) involves two steps. Parameter κ and α are estimated from the star tree assuming the F84+Γ model (Yang 1993), and rates for sites are "predicted" using the star tree by the method of Yang and Wang (in press). In the second step, sites are combined into $K = 4$ classes according to their predicted rates, and those fixed rates are assumed in evaluating the bifurcating trees. Only branch lengths are estimated during the second step. (See Yang 1994b for the details.)
3. *LSG.* The least squares additive tree method (Cavalli-Sforza and Edwards 1967) also involves two steps. Pairwise distances are calculated using a formula based on the F84+Γ model (Yang 1994b), the α parameter being estimated assuming the star tree and the F84+Γ model (Yang 1993). The second step involves evaluation of tree topologies by the (ordinary) least squares criterion based on the estimated pairwise distances.
4. *ML0.* A single rate is assumed for all sites in this likelihood analysis, and this assumption is violated. The κ parameter is estimated from data for each tree topology, together with the branch lengths.
5. *LS0.* The least squares method based on pairwise distances estimated assuming a single rate for all sites differs from *ML0* only in that it performs a pairwise comparison while *ML0* performs a joint analysis of all sequences.

In the likelihood analyses (*MLdG, MLfr,* and *ML0*), the frequency parameters—π_T, π_C, π_A, and π_G—are estimated by the averages of the observed frequencies in the four sequences, while, in estimating pairwise distances for the *LSG* and *LS0* methods, they are estimated by the averages of frequencies in the two compared sequences.

Results

Comparison of Tree Estimation Methods

Estimates of $P_1(n)$, the probability that the estimated tree is the true tree, are listed in Table 1 for parameter combinations $b = 0.5$, $\alpha = 0.2$ and $b = 0.5$, $\alpha = 0.8$, and in Table 2 for $b = 1$, $\alpha = 0.2$ and $b = 1$, $\alpha = 0.8$. The *MLdG, MLfr,* and *LSG* methods all allow for rate variation over sites to some extent. However, unlike a method assuming the F84+Γ model, statistical consistency is not guaranteed for any of them as all three are only approximations to the F84+Γ model: *MLdG* uses four categories to approximate the gamma distribution; *MLfr* uses the star tree to predict rates and also applies an approximation to

Table 1. The number of cases out of 200 simulations in which the true tree is recovered by different tree estimation methods when $b = 0.5$ in the true tree T_1 of Fig. 1[a]

		100	200	500	1,000	2,000
$\alpha = 0.2$						
	MLdG	104	139	166	192	199
	MLfr	104	115	162	183	196
	LSG	96	102	135	175	186
	ML0	106	115	131	160	176
	LS0	85	89	93	109	96
$\alpha = 0.8$						
	MLdG	144	169	195	199	200
	MLfr	146	177	196	198	200
	LSG	100	139	175	187	199
	ML0	136	165	191	198	200
	LS0	97	118	140	149	165
	ML0	152	186	200	200	200
	LS0	100	146	177	194	200

[a] The F84+Γ model was used to generate data, where α is the shape parameter of the gamma distribution. Other parameters were fixed at $\kappa = 5$, $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$, $\pi_G = 0.4$. The sequence length is n. See the Methods section for details of the tree estimation methods (*MLdG, MLfr, LSG, ML0,* and *LS0*)

Table 2. The number of cases out of 200 simulations in which the true tree is recovered by different tree estimation methods when $b = 1$ in the true tree T_1 of Fig. 1[a]

		100	200	500	1,000	2,000
$\alpha = 0.2$						
	MLdG	87	118	132	155	185
	MLfr	75	101	124	149	183
	LSG	76	88	92	120	149
	ML0	71	70	44	26	14
	LS0	46	30	16	3	2
$\alpha = 0.8$						
	MLdG	117	153	164	190	199
	MLfr	117	133	172	190	199
	LSG	87	74	104	156	164
	ML0	102	107	116	144	147
	LS0	58	43	31	15	9
$\alpha = \infty$[b]						
	ML0	124	145	189	199	200

[a] The F84+Γ model was used to generate data where α is the shape parameter of the gamma distribution. Other parameters were fixed at $\kappa = 5$, $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$, $\pi_G = 0.4$. The sequence length is n. See the Methods section for details of the tree estimation methods (*MLdG, MLfr, LSG, ML0* and *LS0*)

[b] There are many cases where the distance measures (Yang 1994b) are inapplicable and results for *LS0* are not presented

the likelihood function (Yang 1994b); the α parameter used in *LSG* for estimating pairwise distances involves systematic errors as it is obtained from a wrong tree topology (the star tree). In principle, the approach of Felsenstein (1978) (see also Debry 1992 and Yang 1994a) can be used to examine whether these methods are statistically consistent for given values of parameters. This approach has not been taken here. The trend of

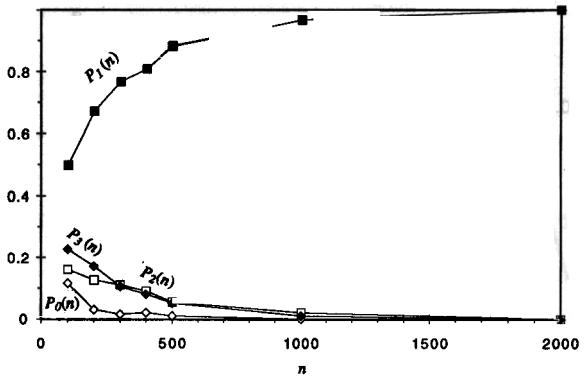

Fig. 2. Estimates of probabilities $P_1(n)$ (■), $P_2(n)$ (□), $P_3(n)$ (♦), and $P_0(n)$ (◇) that tree topologies T_1, T_2, T_3, and T_0 (Fig. 1) will be the estimated tree by the *MLdG* method. The F84+Γ model was used for generating data, with parameter combination $b = 0.5$ (in the tree T_1 of Fig. 1) and $\alpha = 0.2$. Other parameters were fixed at $\kappa = 5$, $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$, and $\pi_G = 0.4$. For each sequence length n, 200 data sets were simulated.

$P_1(n)$ as a function of n suggests that all three methods are consistent for the parameter values examined, that is, when $n \to \infty$, $P_1 \to 1$ (Tables 1 and 2). Previous studies suggest that the estimation of tree topology is quite robust to small errors in the assumed model (e.g., Fukami-Kobayashi and Tateno 1991; Yang 1994a); we may expect these three methods, especially *MLdG,* to have very large consistency domains, if they are not consistent over the whole parameter space of the F84+Γ model.

A typical pattern for a consistent method is shown in Fig. 2, for the case of *MLdG* for parameter combination $b = 0.5$, $\alpha = 0.2$. Note that the improvement in $P_1(n)$, the probability that the true tree is recovered by the method, increases very quickly with small n and then slowly when n is large. This is similar to estimation of a conventional statistical parameter. Results in Tables 1 and 2 suggest that in general *MLdG* performs better than *MLfr,* while both are more efficient than *LSG.* For $b = 0.5$, $\alpha = 0.2$, the estimated probability that the true tree will be recovered with $n = 1{,}000$ nucleotides is 0.96, 0.92, and 0.88 for *MLdG, MLfr,* and *LSG,* respectively (Table 1).

The *ML0* and *LS0* methods assume a single rate for all sites (F84), and this assumption is violated as the data are generated under F84+Γ. Results in Tables 1 and 2 suggest that both *ML0* and *LS0* are consistent for parameter combination $b = 0.5$, $\alpha = 0.8$, while for $b = 1$, $\alpha = 0.2$, neither *ML0* nor *LS0* is consistent. For $b = 1$, $\alpha = 0.8$, *ML0* is consistent while *LS0* is not (Table 2). For the combination $b = 0.5$, $\alpha = 0.2$, *ML0* is clearly consistent while the pattern for *LS0* is not clear. Large-scale simulations are performed, and *LS0* is found to be consistent (Fig. 3); this will be discussed in the next section. In general the results conform with the analysis of Yang (1994a) in that the least squares pairwise comparison has a smaller consistency domain than a joint likelihood analysis.

For all parameter values examined in this study, least

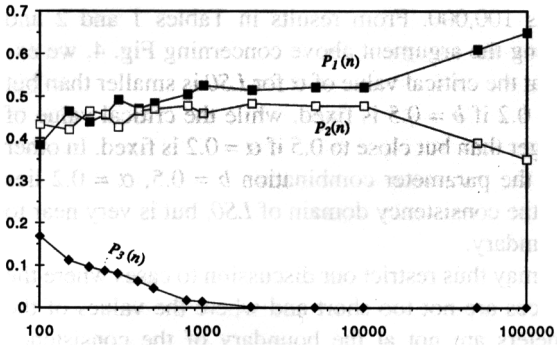

Fig. 3. Estimates of probabilities $P_1(n)$ (■), $P_2(n)$ (□), and $P_3(n)$ (♦) that tree topologies T_1, T_2, and T_3 (Fig. 1) will be chosen as the best estimate by the *LS0* method. The F84+Γ model was used for generating data, with $b = 0.5$ and $\alpha = 0.2$ (the same as in Fig. 2). For each sequence length n, $r = 1{,}000$ data sets were generated by simulation. There are 15 and two cases (out of 1,000) in which the star tree is best at $n = 100$ and at $n = 150$ respectively, and these are not shown. Note that $P_1(n)$ and $P_2(n)$ are both in the narrow range 0.4–0.6 for n as large as 10,000. Although the method is consistent [i.e., $P_1(n) \to 1$ as $n \to \infty$], it is very inefficient (cf. Fig. 2 for *MLdG* applied to the same data). Note that the n-axis is presented logarithmically.

squares methods perform more poorly than likelihood methods (*LSG* vs *MLdG* and *MLfr*, *LS0* vs *ML0*). If a least squares pairwise comparison is as efficient as a joint likelihood analysis, we would expect the performance of *LSG* to be somewhere between *MLdG* and *MLfr;* a likelihood implementation comparable to *LSG,* which would use the estimate of α from the star tree for evaluating the bifurcating trees, was not considered in this study. Nevertheless, *LSG* is noted to be much worse than either *MLdG* or *MLfr*. A least squares method using the true α to calculate pairwise distances was also used in the simulation, although the results are not presented as this option is not available to real data analysis. The results obtained from this method (not shown) are very similar to, or even slightly poorer than, those obtained from *LSG*. This appears to suggest that the poor performance of *LSG* for these parameter combinations may not be due to the inaccuracy of the estimate of α; use of the star tree was noted to give underestimates of α (Yang 1994b). Furthermore, for cases where *ML0* is consistent ($b = 0.5$, $\alpha = 0.2$; $b = 0.5$, $\alpha = 0.8$; $b = 1$, $\alpha = 0.8$), it is as efficient as, or even better than, *LSG,* especially for small samples ($n = 100$–500) (Tables 1 and 2). We expect *LSG* to have a larger consistency domain than *ML0,* and, for other values of parameters where *ML0* is inconsistent or only marginally consistent (see the next section), *LSG* will be more efficient than *ML0*. Nevertheless, results in Tables 1 and 2 clearly suggest the inefficiency of pairwise comparisons.

Statistical Consistency of Phylogenetic Tree Estimation and its Finite Sample Implications

It is noteworthy that the two properties for a tree estimation method, i.e., consistency and efficiency, are not exactly the same as the consistency and sampling error of an estimator of a conventional statistical parameter. In order to achieve a better understanding of the tree estimation problem and to assist practical interpretation of simulation results, the relationship between the two properties for a tree estimation method will be examined in this section. We make two speculations concerning the behavior of $P_i(n)$ as a function of the sequence length n, with parameters θ and $\mathbf{t}_1$ fixed. First, $P_i(n)$ might be expected to be a monotonic function of n. If the method is statistically consistent, that is, $P_1 \to 1$ as $n \to \infty$, we might expect P_1 to increase with n for any n, and probabilities for other trees to decrease with n. On the other hand, if the method is inconsistent, that is, $P_1 \to 0$ as $n \to \infty$, we might expect P_1 to decrease with any n, and we might expect the probability for one of the wrong trees to increase with n and approach 1 as $n \to \infty$. This speculation derives justification from the observation that the standard error of a conventional parameter is usually in proportion to $1/n^{1/2}$, whereas, in the current context, $P_i(n)$, especially $P_1(n)$, measures the sampling error of a tree estimation method.

Next, we speculate that, when the method is consistent, $P_1(n) > P_i(n)$ $(i \neq 1)$ holds for any n, i.e., the true tree will have higher probability of being the estimate in a finite sample than any of the wrong trees.

There are two immediate exceptions to the above speculations, which will be dealt with first. We note that both speculations are incorrect when n is very small. In typical DNA sequence data, there are a high proportion of "constant" site patterns, i.e., site patterns represented by identical nucleotides across species. When n is very small, say $n < 10$ or 50 depending on the overall amount of evolution as reflected in branch lengths $\mathbf{t}_1$, the most probable data sets will mainly consist of the four constant site patterns, and the most probable tree is the star tree. $P_0(n)$ will be near to 1 for very small n and then decrease and approach 0 as n increases, while $P_i(n)$ $(i \neq 0)$ will increase from values near to 0. Both speculations are then incorrect. Since there is no hope of recovering the true tree with such short sequences, we can restrict our discussion to relatively large n where $P_0(n)$ is negligible.

The fact that a method assuming a wrong model can still provide consistent estimation of the tree topology has led to another complexity concerning the behavior of $P_i(n)$ as $n \to \infty$. Take *ML0* as an example, which, assuming $\alpha = \infty$, is likely to be consistent for large α and to be inconsistent for small α. Table 2 shows that *ML0* is consistent when $\alpha = 0.8$ while not when $\alpha = 0.2$, b being fixed at 1. Consider values of α in the interval (0.2, 0.8). For this purpose, we follow the approach of Yang (1994a) and plot the limiting values $\ell_i/n = \ln\{L_i\}/n$, $n \to \infty$, as functions of α with other parameters fixed (Fig. 4), where L_i is the likelihood for tree T_i. T_2 (Fig. 1) is generally chosen as the estimate when *ML0* is inconsistent. There exists a critical value α^* between 0.2 and 0.8 such

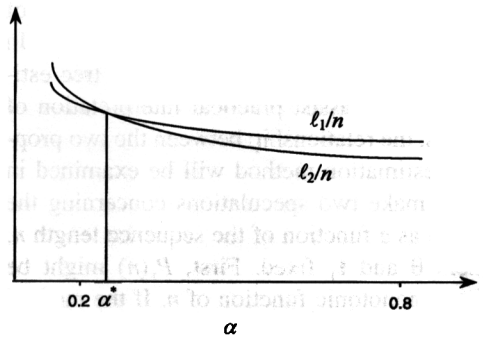

Fig. 4. A schematic representation of the consistency domain of *MLO* when the true model is F84+Γ. All parameters in the model except α were fixed at $\kappa = 5$, $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$, $\pi_G = 0.4$, and $b = 1$ (in the tree T_1 of Fig. 1). The (log) likelihood values ℓ_1/n and ℓ_2/n for tree topologies T_1 and T_2 are limiting values when $n \to \infty$, maximized over branch lengths in each tree and the κ parameter in the F84 model. Likelihood values for the tree T_3 are smaller than those for T_1 and T_2 and are not shown. There exists a critical value α^* between 0.2 and 0.8 (Table 2): if $\alpha > \alpha^*$, then $\ell_1/n > \ell_2/n$ ($n \to \infty$) and *MLO* is consistent, while if $\alpha < \alpha^*$, then $\ell_1/n < \ell_2/n$ ($n \to \infty$) and the method is inconsistent.

that $\ell_1/n > \ell_2/n$ ($n \to \infty$) and *MLO* is consistent when $\alpha > \alpha^*$ (*e.g.*, $\alpha = 0.8$), while $\ell_1/n < \ell_2/n$ ($n \to \infty$) and *MLO* is inconsistent when $\alpha < \alpha^*$ (*e.g.*, $\alpha = 0.2$). Exactly at $\alpha = \alpha^*$, we have $\ell_1/n = \ell_2/n$ ($n = \infty$), that is, tree topologies T_1 and T_2 will have identical (highest) likelihood values when $n \to \infty$, and the method will not converge to a single tree topology. Many such critical values can be identified from results obtained by Debry (1992) and Huelsenbeck and Hillis (1993) for several parsimony and distance matrix methods and by Yang (1994a) for the likelihood method.

Intuitively, information in the data, when properly handled, suggests that T_1 is the true tree. However, use of a wrong model (*MLO*) to analyze the data tends to suggest T_2 as the estimate. When these two tendencies are balanced, as at $\alpha = \alpha^*$ in Fig. 4, the (maximum) likelihoods for the two trees approach identical values as $n \to \infty$, while in finite samples, $P_1(n)$ and $P_2(n)$ will be very similar. With more than four sequences in the data, it seems possible that three or more trees, either including the true tree or not, may share the probability of being the best estimate—that is, as $n \to \infty$, the scores for several best topologies (likelihood values or sums of squares) may approach identical values.

One implication of this observation is that, if the values of parameters are at or near to the boundary of the consistency domain of the tree estimation method, the method will be very inefficient in finite samples ($n < \infty$). One such case is the *LS0* method applied to the parameter combination $b = 0.5$, $\alpha = 0.2$ (Table 1). To obtain a clear picture of the pattern, the number of simulations is increased to $r = 1{,}000$, and also more values of n are examined. The results, shown in Fig. 3, suggest that *LS0* is consistent for those values of parameters, i.e., $P_1(n) \to 1$ as $n \to \infty$. However, *LS0* is very inefficient for n as large as 100,000. From results in Tables 1 and 2 and following the argument above concerning Fig. 4, we expect that the critical value of α for *LS0* is smaller than but near to 0.2 if $b = 0.5$ is fixed, while the critical value of b is larger than but close to 0.5 if $\alpha = 0.2$ is fixed. In other words, the parameter combination $b = 0.5$, $\alpha = 0.2$ lies within the consistency domain of *LS0*, but is very near to the boundary.

We may thus restrict our discussion to cases where the sequences are not too short and where the values of the parameters are not at the boundary of the consistency domain of the tree estimation method, which are cases most relevant for practical data analysis. From the estimates of $P_1(n)$ shown in Tables 1 and 2 and the corresponding estimates of $P_2(n)$ and $P_3(n)$ (not shown), the speculations are noted to be clearly correct for most of the methods and parameter combinations. Exceptions to these speculations appear likely, however, especially in light of the case explored in Fig. 3 where the pattern remains unclear. Nevertheless, we suggest that, when the speculations are incorrect, they may not be too wrong: for example, in cases where $P_1(n) < P_2(n)$ for some n even if $P_1(n) \to 1$ as $n \to \infty$, the differences between $P_1(n)$ and $P_2(n)$ are expected to be very small. We conclude that these two speculations, which are not always true but which are expected never to be too wrong, can be taken as practical guidelines.

For the estimation of a conventional parameter, statistical consistency is a very weak requirement, as there are infinitely many consistent estimators; if $\hat{\theta}$ is a consistent estimator of θ, then $(n\hat{\theta} + a)/(n\hat{\theta} + b)$ will be another consistent estimator of θ for any constants a and b. Without consistency, other properties such as efficiency are not really meaningful (Kendall and Stuart 1979, pp 1–37). The two speculations discussed above suggest that the consistency for a tree estimation method is a stronger requirement than the consistency for an estimator of a statistical parameter. The speculations also suggest that a similar point of view should be adopted concerning phylogenetic tree estimation; we should require a tree estimation method to be consistent for a very large portion of, if not the whole, parameter space; statistical consistency does not guarantee efficiency, but an inconsistent method is necessarily inefficient.

Estimations of Branch Lengths in the Tree and Parameters in the Model

Estimation of branch lengths in the true tree and parameters in the true model is a standard statistical estimation problem, and the asymptotic properties of maximum likelihood estimators are well established. In this section the estimates of branch lengths and other parameters by *MLO* will be examined, which ignores the rate variation among sites. The complexity in parameter estimation due to the uncertainty of the phylogeny will also be exam-

ined. Most of the results have been observed in previous analyses of real data, and here only qualitative results will be presented.

Ignoring rate variation at sites leads to underestimation of branch lengths. The underestimation, however, is not proportional and is much more serious for long branches than for short branches. Using the simple substitution model of Jukes and Cantor (1969), Gillespie (1986) showed that sequence divergence is underestimated if rate variation at sites exists and is ignored. Analyses using other substitution models lead to the same conclusion (e.g., Yang et al. 1994), which has also been confirmed by simulation studies (e.g., Jin and Nei 1990). The reason for the underestimation is that ignoring rate variation at sites tends to overlook some of the multiple substitutions that have occurred at the fast-changing sites.

The transition/transversion rate bias, as reflected in parameter κ, is underestimated when rate variation at sites is ignored. Yang et al. (1994, in press) observed negative correlations between estimates of κ and α when analyzing real data sets by the maximum likelihood approach; when the transition/transversion rate bias is ignored, i.e., if κ is fixed at 0 for the F84 model, α will be overestimated, while if rate variation at sites is ignored, i.e., if α is fixed at infinity, κ will be underestimated. Wakeley (1994) pointed out that the underestimation of κ is due to the fact that assuming rate constancy among sites tends to ignore some of the transitional substitutions that have occurred at the fast-changing sites.

Estimates of the κ parameter by *ML0* are very similar for different tree topologies. Estimates of parameters κ and α by *MLdG* are also stable, but some patterns are noticeable: estimates of α obtained from the star tree are normally smaller than those obtained from other trees, especially if α is very small—that is, if there is severe rate variation. The negative correlation between estimates of κ and α appears also to apply for estimates from different tree topologies, and as a result of this, estimates of κ from the star tree (for *MLdG*) are normally larger than those from other trees. These patterns were observed and discussed with real data under the (continuous) gamma model by Yang (1994b) and Yang et al. (1994). Using the star tree to estimate the α parameter is therefore not a good practice, although it is adopted in this study for convenience, and more reasonable trees should be used for real data analysis.

Estimates of interior branch lengths in the nonbest trees are often 0 by the least squares methods (*LSG* and *LS0*), while, with the likelihood analysis (*MLdG* and *ML0*), they are often strictly positive. Noticeably, cases in which all three bifurcating trees are better than the star tree are rare by the least squares methods, while such cases are commonplace by the likelihood methods. This is compatible with the theoretical analysis of Yang (1994a), which suggests that the test for positivity of interior branch lengths in the estimated tree may be taken as an evaluation of the reliability of the least squares tree, but not of the maximum likelihood tree. The *MLfr* method appears similar to the least squares methods in this respect, as noted by Yang (1994b).

Discussion

Sequencing the Right Genes

As we expect that the order among the $P_i(n)$'s for different tree topologies T_i's normally does not change with n, it is feasible to study the probability $P_1(n)$ for fixed n as a function of parameters θ and $\mathbf{t}_1$. Such information is useful from the point of view of experimental design; as different genes have different evolutionary dynamics which correspond to different values of parameters θ and $\mathbf{t}_1$, choosing genes that correspond to highest $P_1(n)$ as targets of sequencing can be expected to be most economical. In a similar vein, Saitou and Nei (1986) considered the sequence length required for the parsimony method to recover the true tree with a prespecified probability P.

The most important factor appears to be the branch lengths $\mathbf{t}_1$, which reflect the overall amount of evolution; with too little change the sequences are very similar and contain little evolutionary information, while with too much evolution the sequences will be too different and too noisy. Biologists have understood the economics of choosing sequences with the right amount of evolution, which, given the species, is largely determined by the importance or conservedness of the gene. The effects of other parameters may be confounded with the effects of branch lengths. We take *MLdG* as a close approximation to the true model, F84+Γ, and examine the effect of α on $P_1(n)$. Comparison of results for $\alpha = 0.2$ with those for $\alpha = 0.8$ suggests that, other parameters being equal, sequences with slight rate variation over sites provide more information about the phylogenetic relationship than sequences with severe rate variation. For example, with $n = 500$ and $b = 0.5$, there is about a 97.5% chance of recovering the true tree by *MLdG* if $\alpha = 0.8$, while the chance is only about 83% if $\alpha = 0.2$. When $\alpha = \infty$ (no rate variation over sites), this chance (*ML0*) is about 100% (Table 1). It is noted that the number of different site patterns observed in a data set is normally much smaller if the data are generated assuming a gamma distribution of rates over sites than if a single rate is assumed for all sites (results not shown). Particularly under the gamma distribution model, there are far more constant site patterns in the data, which provide less information concerning the phylogeny than variable site patterns. Similar arguments might be applicable to the effects of κ and the frequency parameters π_T, π_C, π_A, and π_G; sequences with little transition/transversion rate bias ($\kappa = 0$) or with nearly equal base frequencies ($\pi_T = \pi_C = \pi_A = \pi_G = 1/4$) might be expected to provide more information concern-

ing the evolutionary relationships than sequences with extreme rate bias or base frequencies.

Overall Evaluation of the MLdG and MLfr Methods

In this study, the method that assumes the true model (F84+Γ) is not examined, and the discrete gamma model is used as a close approximation. For real data, both models are approximate descriptions of rate variation in real sequences, and the discrete gamma model should be justifiable on its own, apart from the interpretation of the α parameter. The computational advantage of *MLdG* over F84+Γ suggests that *MLdG* is to be preferred.

The fixed rates approach (*MLfr*) appears to perform reasonably well. The performance obviously depends on the accuracy of prediction of rates at sites. The correlation between the real rate and the predicted rate, calculated by the method of Yang and Wang (in press), is $\rho = 0.623$ for parameter combination $b = 0.5$, $\alpha = 0.2$; $\rho = 0.532$ for $b = 0.5$, $\alpha = 0.8$ (Table 1); $\rho = 0.609$ for $b = 1$, $\alpha = 0.2$; and $\rho = 0.530$ for $b = 1$, $\alpha = 0.8$ (Table 2); in all cases, the true tree is assumed in the calculation. These values are expected to be slightly too high because we have used the star tree instead of the true tree to predict the rates and because maximum likelihood estimates of θ and $\mathbf{t}_1$ from simulated data involve sampling errors. It can be expected that if the true rates, i.e., the gamma-distributed random variables used to generate the data, were used instead of the predicted rates, the performance of *MLfr* would be better, presumably even better than using F84+Γ. By the same argument, the performance of *MLfr* relative to that of F84+Γ or *MLdG* can be expected to improve when there are more species in the data, as the accuracy of rate prediction is mainly determined by the number of sequences (Yang and Wang, in press). The major shortcoming of *MLfr* is its ad hoc nature; for instance, it fails to provide a natural measure of the model's goodness of fit to data. As the computation required by *MLdG* is only a few times that for the single rate model (*ML0*), *MLfr* is not recommended for real data analysis.

Addendum

After submission of the manuscript for this paper, Tateno et al. (1994) published a simulation study that compared several tree reconstruction methods with the data generated assuming gamma-distributed rates for sites. The maximum likelihood program (ML) used by those authors assumes a single rate for all sites, while the distance matrix method (neighbor-joining, NJ; Saitou and Nei 1987) uses pairwise distances estimated assuming either a single rate or gamma-distributed rates at sites. The results obtained by those authors (Tables 1–4 in Tateno et al. 1994) appear fully compatible with those of the present study, although their conclusions somewhat differ from those of this study. Tateno et al. (1994) appeared to have performed the comparison from a point of view of practical data analysis, and did not make a clear distinction between a tree estimation method and a computer program. For example, they suggested that ML was slightly more sensitive to violation of the assumptions in estimating topology than was NJ with gamma distances. This comparison is not fair, since no assumption is violated for NJ with gamma distances while ML ignores rate variation at sites. If ML is compared with NJ without the gamma correction, in which case both methods assume the same wrong model, ML has much higher probability of recovering the true tree than NJ (Table 3 in Tateno et al. 1994). For other values of parameters examined by those authors, all methods recovered the true tree with probabilities near to 1, and they concluded that the methods in general showed more or less the same performance. This conclusion may be incorrect and may be due to the authors' use of a single (large) sequence length; if only one sequence length, $n = 2{,}000$, had been used in the present study, no difference would have been detected among *MLdg, MLfr, LSG,* and *ML0* for the parameter combination $\alpha = 0.8$, $b = 0.5$, although their performances are quite different (Table 1).

Acknowledgments. I wish to thank Clive Moncrieff for many helpful discussions, especially concerning the finite-sample implications of statistical consistency of a phylogenetic tree estimation method. I am also grateful for his comments on an earlier version of the manuscript. Dr. M. Nei provided facilities during the revision of the manuscript. This study was supported by a grant from Department of Zoology, The Natural History Museum (London).

References

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Evolution 32:550–570

Debry RW (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

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

Gillespie JH (1986) Rates of molecular evolution. Ann Rev Ecol Syst 17:637–65

Hasegawa M, Kishino H (1989) Confidence limits on the maximum likelihood estimation of the hominoid tree from mitochondrial DNA sequences. Evolution 43:672–677

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

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

Hasegawa M, Rienzo AD, Kocher TD, Wilson AC (1993) Toward a more accurate time scale for the human mitochondrial DNA tree. J Mol Evol 37:347–354

Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst Zool 38:297–309

Huelsenbeck JP, Hillis DM (1993) Success of phylogenetic methods in the four-taxon case. Syst Biol 42:247–264

Jin L, Nei M (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol Biol Evol 7:82–102

Kendall M, Stuart A (1979) Advanced theory of statistics, vol 2. Charles Griffin & Company, London

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

Li W-H, Gouy M, Sharp PM, O'hUigin C, Yang Y-W (1990) Molecular phylogeny of rodentia, lagomorpha, primates, artidactyla and carnivora and molecular clocks. Proc Natl Acad Sci USA 87:6703–6707

Saitou N, Nei M (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

Saitou N, Nei M (1987) The neighbour joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Takahata N (1991) Overdispersed molecular clock at the major histocompatibility complex loci. Proc R Soc Lond [Biol] 243:13–18

Takezaki N, Nei M (1994) Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. J Mol Evol 39:210–218

Tamura K, Nei M (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

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

Wakeley J (1993) Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J Mol Evol 37:613–623

Wakeley J (1994) Substitution rate variation among sites and the estimation of transition bias. Mol Biol Evol 11:436–442

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

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

Yang Z (1994b) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314

Yang Z, Wang T (in press) Mixed model analysis of DNA sequence evolution. Biometrics

Yang Z, Goldman N, Friday AE (1994) Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol Biol Evol 11:316–324

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

Zharkikh A, Li WH (1993) Inconsistency of the maximum-parsimony method: the case of five taxa with a molecular clock. Syst Biol 42:113–125