Relative Efficiencies of the Maximum-Likelihood, Neighbor-joining, and Maximum-Parsimony Methods When Substitution Rate Varies with Site

Yoshio Tateno, Naoko Takezaki,† and Masatoshi Nei†*

*National Institute of Genetics, Mishima; and †Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University

The relative efficiencies of the maximum-likelihood (ML), neighbor-joining (NJ), and maximum-parsimony (MP) methods in obtaining the correct topology and in estimating the branch lengths for the case of four DNA sequences were studied by computer simulation, under the assumption either that there is variation in substitution rate among different nucleotide sites or that there is no variation. For the NJ method, several different distance measures (Jukes-Cantor, Kimura two-parameter, and gamma distances) were used, whereas for the ML method three different transition/transversion ratios (R) were used. For the MP method, both the standard unweighted parsimony and the dynamically weighted parsimony methods were used. The results obtained are as follows: (1) When the R value is high, dynamically weighted parsimony is more efficient than unweighted parsimony in obtaining the correct topology. (2) However, both weighted and unweighted parsimony methods are generally less efficient than the NJ and ML methods even in the case where the MP method gives a consistent tree. (3) When all the assumptions of the ML method are satisfied, this method is slightly more efficient than the NJ method. However, when the assumptions are not satisfied, the NJ method with gamma distances is slightly better in obtaining the correct topology than is the ML method. In general, the two methods show more or less the same performance. The NJ method may give a correct topology even when the distance measures used are not unbiased estimators of nucleotide substitutions. (4) Branch length estimates of a tree with the correct topology are affected more easily than topology by violation of the assumptions of the mathematical model used, for both the ML and the NJ methods. Under certain conditions, branch lengths are seriously overestimated or underestimated. The MP method often gives serious underestimates for certain branches. (5) Distance measures that generate the correct topology, with high probability, do not necessarily give good estimates of branch lengths. (6) The likelihood-ratio test and the confidence-limit test, in Felsenstein's DNAML, for examining the statistical significance of branch length estimates are quite sensitive to violation of the assumptions and are generally too liberal to be used for actual data. Rzhetsky and Nei's branch length test is less sensitive to violation of the assumptions than is Felsenstein's test. (7) When the extent of sequence divergence is $\leq 5\%$ and when $\geq 1{,}000$ nucleotides are used, all three methods show essentially the same efficiency in obtaining the correct topology and in estimating branch lengths. Clearly, the simplest method, i.e., the NJ method, is preferable in this case.

Introduction

The maximum-likelihood (ML) method of phylogenetic inference (Felsenstein 1981) has nice statistical properties, compared with many other methods. In practice, however, it requires a number of simplifying assumptions that do not necessarily hold with actual data. It is therefore important to examine the effects of violation of these assumptions on the statistical efficiency of the method. Using simple model trees with a constant substitution rate, Fukami-Kobayashi and Tateno (1991) have shown that the probability of obtaining the correct tree for this method is relatively insensitive to different assumptions about the ratio (R) of the transition rate (s) to the transversion rate (v) and about the G+C content, though the branch length estimates are affected substantially. By contrast, using a special model of amino acid substitution and actual data, Reeves (1992) concluded that one of the most important factors that reduce the efficiency of the ML method is substitution-rate variation among different sites.

In the study of the efficiency of the ML method, it is also important to compare its efficiency with that of other methods of phylogenetic inference. Saitou and Imanishi (1989) studied the probability of obtaining the

Key words: maximum-likelihood method, neighbor-joining method, maximum-parsimony method, varying rate of substitution, phylogeny.

Address for correspondence and reprints: Masatoshi Nei, Institute of Molecular Evolutionary Genetics, Pennsylvania State University, 328 Mueller Laboratory, University Park, Pennsylvania 16802-5303.

Mol. Biol. Evol. 11(2):261–277. 1994.

0737-4038/94/1102-0010$02.00

correct topology for the ML, maximum-parsimony (MP), and a few other distance (e.g., neighbor-joining [NJ]) methods and showed that the ML method and the NJ method are nearly equally efficient and that these two methods are generally more efficient than the MP method (Eck and Dayhoff 1966; Fitch 1971) and than Fitch and Margoliash's (1967) method. However, they did not examine the effect of violation of the assumptions of the ML method on the efficiency of this method.

The purpose of this paper is to examine the effect of violation of the assumptions required for the ML method and to compare the statistical efficiency of that method with that of the MP and the NJ methods. The methods used are primarily computer simulations.

Mathematical Models and Methods

The method of our computer simulation was essentially the same as that of Jin and Nei (1990). We considered four DNA sequences of 1,000 nucleotides each and assumed that they evolve following the model trees given in figure 1. The reason we considered only four sequences is that the ML method requires an enormous amount of computer time, in this kind of study. We considered various types of nucleotide substitution to generate the four "extant" DNA sequences. The sequences thus obtained were used to reconstruct a tree, and the tree reconstructed was compared with the model tree. This process was repeated 1,000 times for each paramcter set, except in the case of the ML method, where only 100 replications were used, because of the limited computational time available. Note that the model trees in figure 1 are all unrooted trees and represent the cases where the rate of nucleotide substitution varies with evolutionary lineage.

The "extant" DNA sequences were generated by using Kimura's (1980) substitution model, under the assumption either that the substitution rate is the same for all sites or that the rate varies with site according to a gamma distribution. To simulate rate heterogeneity among different sites, we used three different gamma distributions with parameter $a = 0.5$, 1, and 2 (see Jin and Nei 1990). These distributions are shown in figure 2. Note that the gamma parameter $a = 0.5$ is close to the estimate (0.47) obtained for one of the two hypervariable parts of the control region of human mitochondrial DNA (mtDNA) (Wakeley 1993) and that many genes do not have such a low a value (Uzzell and Corbin 1971). The mtDNA hypervariable region is also known to have a high R value, ~ 15 (Vigilant et al. 1991). In other genes, however, it is $\ll 15$ (Nei 1987, p. 84). In the present study we considered the cases $R = 0.5$, 9, and 15.

The methods of phylogenetic inference used were the MP, NJ, and ML methods. In the case of the NJ method, five different distance measures—i.e., Jukes-Cantor (Jukes and Cantor 1969) distance, Kimura (1980) two-parameter distance, and Jin and Nei's (1990) gamma distances with $a = 0.5$, 1, and 2—were used to see the effects of distance estimation. The NJ method is a special case of the minimum evolution (ME) method (Rzhetsky and Nei 1992) and, to be efficient, supposedly requires an unbiased distance measure. Therefore, if we use different distance measures, we can see how these distance measures affect the efficiency of obtaining the correct tree. The NJ and ME methods give the same tree if the number of sequences used is four.

In the original paper of Felsenstein (1981), a simple model of nucleotide substitution with no s/v bias was used. In the recent versions (version 2.6 and later) of his program package PHYLIP (Felsenstein 1991), he modified the model to accommodate the s/v bias, and the transition matrix is given by equation (A3) in the Appendix. The biological justification of this model is not as clear-cut as Hasegawa et al.'s (1985) model but permits an analytical solution for the estimate of

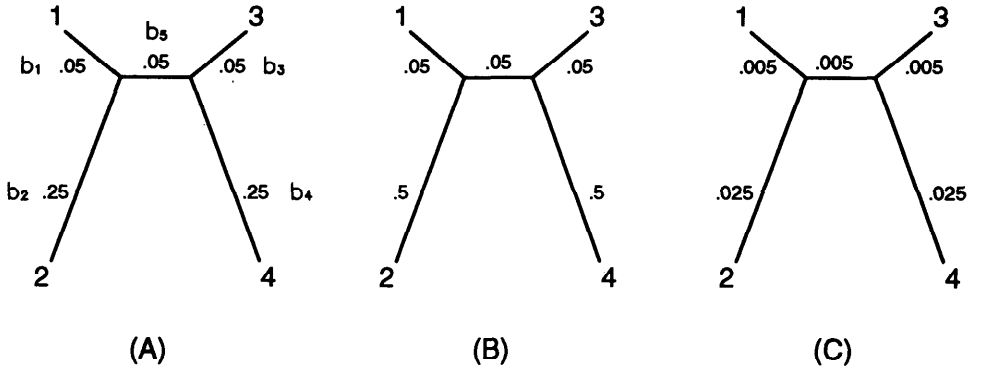

FIG. 1.—Three model trees used for computer simulation

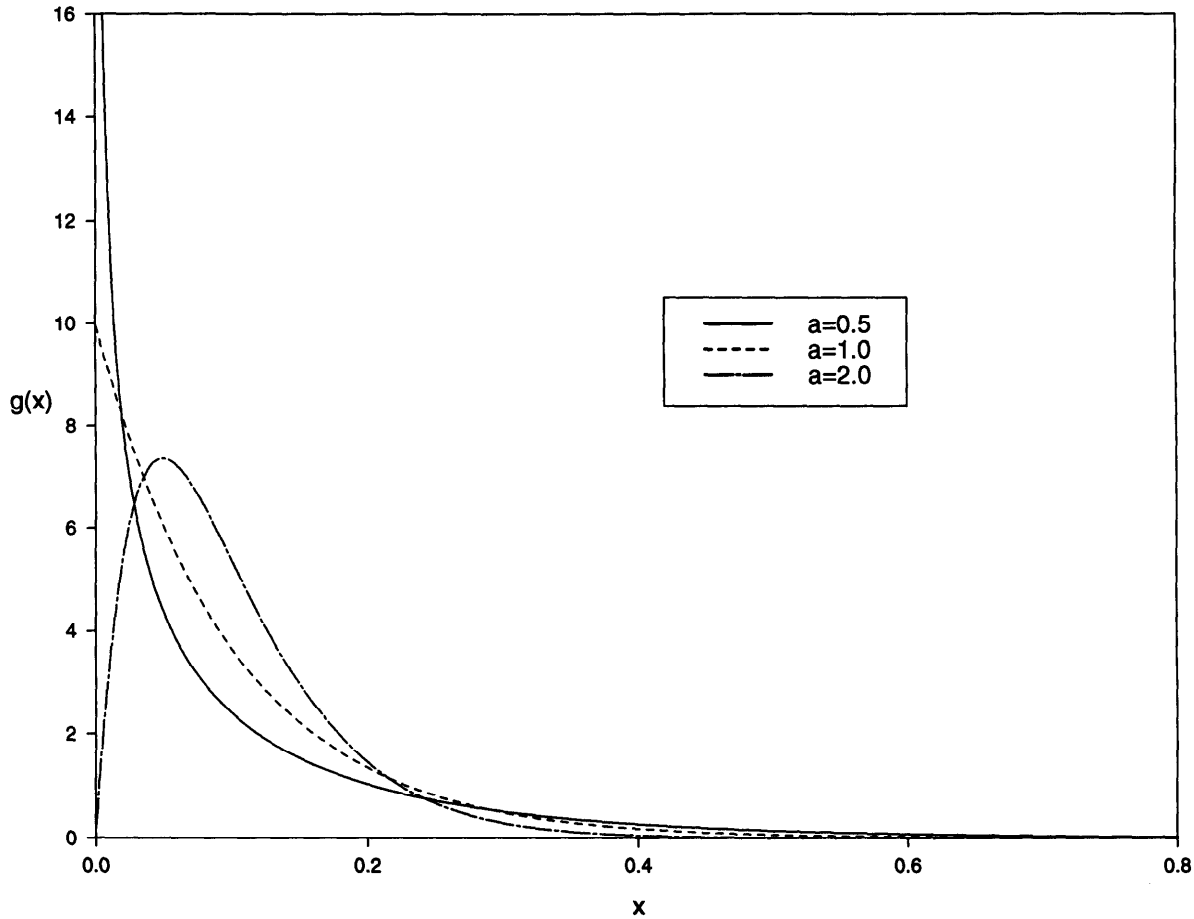

FIG. 2.—Gamma distributions with $a = 0.5$, 1, and 2. $g(x)$ is the probability density. a is the inverse of the square of the coefficient of variation.

the number of nucleotide substitutions per site (d) (eq. [A7]).

In this study, we used the program DNAML of PHYLIP version 3.4 to construct an ML tree. This program has several options, and we used the U (user) option, in which the tree topology to be tested is specified by the user. In this option, the number of iterations for searching for the ML value is greater than that of other options, such as the G (global search) option (J. Felsenstein, personal communication), so a better tree is expected to be obtained. Even in the U option, however, branch length estimates were not always reliable. We therefore eliminated the maximum limit of iterations, which was set to 10 in the source code of DNAML. This elimination resulted in termination of iterations, when the parameter ε in the program is $<10^{-6}$, and enhanced the accuracy of branch length estimates tremendously. The parameters that determine R in DNAML are specified by the R value specified in the T option of the program and by the base frequencies (see eq. [A5] in the Appendix). The default option of DNAML uses the base frequencies in the data. To give the exact parameter values that correspond to the Jukes-Cantor and Kimura two-parameter models, however, we used 0.25 for all the base frequencies with the F option. Version 2.6 and later versions of DNAML do not accept $R = 0.5$. We therefore modified the program slightly to accommodate this case.

DNAML has an option for a substitution model with several different classes of substitution rates, but this model cannot be applied to the case of continuous variation of substitution rate, which we consider in this paper. Therefore, we have not used this model.

As mentioned above, the R value in DNAML is preassigned rather than estimated from the data. This increases the efficiency of the ML method compared with the case where R is estimated. To make the NJ method comparable with the ML method with this procedure, we preassigned the R value in the computation

of Kimura's two-parameter distance (modified Kimura distance), as is done in PHYLIP.

MP trees are usually produced by disregarding the s/v bias (Eck and Dayhoff 1966; Fitch 1971). However, it is possible to take care of this bias as well as the variation in substitution rate among different sites if we use the dynamically weighted parsimony (DWP) method of Sankoff and Cedergren (1983) and Williams and Fitch (1990). We have therefore used this DWP method as well as the standard MP method. Williams and Fitch's algorithm of DWP has nine different ways of character weighting (three initial weightings times three position weightings). Our study showed that among all these options the substitution method of initial weighting plus the quadratic position weighting was the best in our simulation. We therefore present only the results from this option, in this paper. The effect of violation of the assumptions required for each tree-making method was evaluated by computing the frequency of obtaining trees whose topology is correct among all the replications and by computing how much the average estimates of branch lengths deviate from the true values.

Results

Topologies

Case of Equal Rate

The proportions (P) of trees with the correct topology obtained among all replicate simulations (1,000 replications for MP and NJ and 100 replications for ML), for the case of equal rate of nucleotide substitution for all sites, are given in table 1. In this table, NJD and NJK represent the NJ method with Jukes-Cantor one-parameter distance and Kimura two-parameter distance, respectively. NJG represents the NJ method with two-parameter gamma distances (Jin and Nei 1990). The a value listed below NJG represents the gamma parameter for computing a gamma distance. ML0.5, ML9, and ML15 denote the ML method with the assumption of R = 0.5, 9, and 15, respectively.

Table 1 shows that when the rate of nucleotide substitution is the same for all sites with R = 0.5 and when tree A in figure 1 is used, the correct tree topology was obtained for all tree-making methods irrespective of the a and R values used for tree estimation. However, when R = 9 was used for generating DNA sequences, P = 88, for the standard MP method. The P value for the MP method further declines as R increases to 15. This is due to the fact that, when R = 9 and 15, multiple and parallel transitional changes can occur at the same nucleotide sites. In this case, however, if we use DWP, the P value increases significantly. By contrast, the NJ and ML methods generated the correct topology in all or nearly all replications, irrespective of the a and R values used for distance estimation or phylogenetic inference. However, this does not mean that these methods are perfect for phylogenetic inference, because in some cases the estimates of branch lengths are poor, as will be mentioned later.

When tree B is used as the model tree, the efficiency of the standard MP method declines dramatically, and for R = 9 and 15, P = 0. This is of course due to the fact that there are many multiple and parallel transitional changes, which are neglected in parsimony analysis. This effect is known to be serious when the lengths of branches b_2 and b_4 in figure 1 are long relative to those of b_1 and b_3 and tends to give an incorrect topology (inconsistency) as the number of nucleotides increases to infinity (Felsenstein 1978). A theoretical study (N. Takezaki, unpublished data) indicates that when branch lengths b_1, b_3, and b_5 = 0.05, the inconsistency of a parsimony tree occurs whenever b_2 and $b_4 \geq 0.394$ for R = 0.5; ≥ 0.302 for R = 9; and ≥ 0.292 for R = 15. Our results are consistent with these theoretical expectations. By contrast, the MP method gives a consistent tree for the

Table 1
***P* in the Case of Equal Nucleotide Substitution Rate among Different Sites**

					NJG					
R	MP	DWP	NJD	NJK	a = 0.5	a = 1	a = 2	ML0.5	ML9	ML15
Model tree A:										
0.5	100	100	100	100	100	100	100	100	100	100
9	88	99	99	100	100	100	100	100	100	100
15	83	97	99	100	100	100	100	100	100	100
Model tree B:										
0.5	2	3	98	98	100	100	100	100	96	94
9	0	90	21	92	99	99	99	72	100	100
15	0	89	16	90	99	99	98	50	98	99

case of model tree A. However, to obtain the correct tree with a 100% probability for this model tree with $R = 9$ or 15, a number of nucleotides $\geqslant$1,000 must be used. The behavior of DWP is interesting. When $R = 0.5$, it is as bad as the standard MP method. However, when $R = 9$ or 15, the P value for DWP is higher than that for NJD and is nearly as high as that for NJK. This indicates that DWP is effective when R is high but not when $R = 0.5$.

The NJ method is known to give consistent trees whenever unbiased distance measures are used (Saitou and Nei 1987; DeBry 1992). Thus, the P value for NJD is high when $R = 0.5$ but declines as R increases. This is because the Jukes-Cantor distance does not take into account the s/v bias. This bias is considered in the Kimura distance, so NJK gives high P values even for $R = 9$ and 15. However, to obtain the correct topology with a 100% probability, the number of nucleotides used must be >1,000. Note that NJK gives biased estimates of nucleotide substitutions when the number of nucleotides examined is small (Tajima 1993).

It is interesting to see that, when the gamma distances are used, NJG gives $P = 100$ for $R = 0.5$ and gives $P \geqslant 98$ for $R = 9$ and 15. This result is counterintuitive, because in the present case the rate of nucleotide substitution is the same for all sites. However, this paradox can be resolved if we note the condition for obtaining the correct topology for the NJ method. Saitou and Nei (1987) have shown that this condition for the case of four sequences is

$$\begin{aligned} d_{12} + d_{34} &< d_{13} + d_{24}\,, \\ d_{12} + d_{34} &< d_{14} + d_{23}\,, \end{aligned} \qquad (1)$$

where d_{ij} is the distance between sequences i and j. When there is no rate heterogeneity among different sites, the gamma distance is known to give an overestimate of nucleotide substitutions, and the extent of overestimation increases as the true distance (number of nucleotide substitutions) increases and as the a value decreases. Thus, d_{14} and d_{23} are more overestimated than d_{12} and d_{34}, and d_{24}, which is the largest distance, is most seriously overestimated. Therefore, relation (1) holds more easily for gamma distances than for the Jukes-Cantor or the Kimura distance. This explains why NJG shows a higher P value than NJD or NJK.

The ML method shows a high P value when the correct R value is used for computing the ML value. Thus, when $R = 0.5$ is used for generating DNA sequences, the ML method with the assumption of $R = 0.5$ (ML0.5) gives $P = 100$. However, P declines as the R value assumed for computing the ML value increases (e.g., as in ML9 and ML15). In the case of $R = 9$, however, P is 100 for ML9 and ML15 but is 72 for ML0.5. When $R = 15$, P is nearly 100 for ML9 and ML15 but is 50 for ML0.5. The same tendency was observed in Fukami-Kobayashi and Tateno's (1991) study, though these authors considered the case where the molecular clock works. This indicates that the ML method is slightly more sensitive than the NJG method to violation of the assumptions made in the estimation of topology.

In actual data analysis, of course, it is possible to compute the ML value for various R values and then to choose the R value that gives the highest ML value. The ML method with this chosen R value is expected to give a better topology. However, the purpose of this study is to examine the effect of violation of the assumptions of the ML model, and the s/v bias was used merely as an example. The evolution of actual DNA sequences usually deviates from any model of nucleotide substitution currently available for the ML method (see, e.g., Tamura 1994), and it is not always easy to take all the deviations into account mathematically. Therefore, the results of this simulation raise some doubts about the robustness of the ML method.

Case of Varying Rate

In most tree-making methods, when the rate of nucleotide substitution varies according to a gamma distribution, the P value for model tree A is generally lower than when the rate is equal, as expected (table 2). Exceptions are the cases of NJG in which a gamma distance with a proper a value is used. Thus, NJG with $a = 0.5$, 1, or 2 gives a high P value for the case of the gamma distance generated with $a = 0.5$, 1, or 2, respectively. However, it is interesting that NJG with $a = 0.5$ tends to show a high P value even when the a value used for generating sequence data is >0.5. This is caused by overestimation of long pairwise distances when a smaller a value is used for estimating distances, as discussed earlier. When a is large, the P values are close to those for the case of equal rates ($a = \infty$). NJD and NJK usually show a P value smaller than does NJG, as expected.

The P value for the ML method is generally higher when a correct R value is used than when an incorrect R value is used. However, ML15 tends to show a high P value even when the R value used for sequence generation is 0.5 or 9. Generally speaking, the ML method is as good as the NJG method in obtaining the correct topology. By contrast, the MP method generally shows a smaller P value. DWP is generally better than MP and is nearly as good as NJK.

Table 3 shows the P values for model tree B for the case of rate heterogeneity among sites. In this case, the

Table 2
***P* Obtained for Model Tree A in the Case of Varying Nucleotide Substitution Rate**

					NJG					
R	MP	DWP	NJD	NJK	$a = 0.5$	$a = 1$	$a = 2$	ML0.5	ML9	ML15
$a = 0.5$:										
0.5	88	90	90	90	100	98	95	98	95	91
9	62	93	77	87	95	91	86	88	99	99
15	55	89	72	84	95	88	84	84	97	97
$a = 1$:										
0.5	95	96	98	98	100	100	100	100	99	99
9	69	97	87	96	99	99	97	97	100	100
15	63	92	83	93	100	98	95	95	100	100
$a = 2$:										
0.5	98	98	100	100	100	100	100	100	100	99
9	80	98	96	99	100	100	99	99	99	99
15	72	95	93	99	100	99	99	98	100	100

MP method rarely chooses the correct topology, though DWP increases *P* substantially when $R = 9$ and 15. NJD and NJK are also very poor at choosing the correct topology, except for the case of $R = 0.5$ and $a = 2$. (NJK shows a relatively high *P* value when $R = 9$ or 15 and $a = 2$.) This occurs because in this case the Jukes-Cantor and Kimura distances seriously underestimate the true distances when these true distances are large. NJG shows a high *P* value when $R = 0.5$ and the *a* value used for sequence generation is the same as that used for distance estimation. For NJG, *P* declines as *R* increases, for a given value of *a*, as expected. NJG again shows a higher *P* value when the *a* value used for distance estimation is smaller than that used for sequence generation.

The *P* value for the ML method is considerably lower when the substitution rate varies with nucleotide site than when it is the same for all sites (table 3). This is particularly so for the case of $a = 0.5$. In this case, even if the same *R* value is used for generating sequence data and for estimating topology, *P* is only 48 for ML0.5. When the *R* value used for generating sequence data is greater than that used for estimating topology, the *P* value is quite low in all cases. By contrast, when the former is smaller than the latter, *P* tends to be as high as the *P* for the case in which both *R*s are the same, except for the case of $R = 0.5$. At any rate, these results show that ML is slightly more sensitive to violation of the assumptions in estimating topology than is NJG.

In model tree C, the expected length of each branch is one-tenth the length of the corresponding branch in tree A. We included this case because sequence data of this magnitude of differentiation are often used in phylogenetic analysis (Hedges et al. 1990; Vigilant et al. 1991). The *P* values for this tree are given in table 4. In this case, the *P* values are slightly smaller than those in table 1, except for MP with $R = 9$ and 15. This has occurred because in the present case the true length of the interior branch (b_5) was so small that there were cases in which the three possible topologies were indistinguishable. (The observed value of this branch length sometimes becomes zero.) At any rate, table 4 shows that the *P* values are virtually the same for all three tree making methods, whether there is an s/v bias or rate heterogeneity among different sites. This result indicates that when the extent of sequence difference is small, it is sufficient to use a simple statistical method such as NJ for inferring phylogenetic trees.

Branch Lengths

Case of Equal Rate

In phylogenetic inference, it is important not only to determine correct topologies but also to obtain good estimates of branch lengths. Of course, estimation of branch lengths is meaningless unless the topology of a tree is correctly inferred. In the following, we therefore consider estimates of branch lengths only for the cases where the correct topology is obtained. To make the estimates for different tree-building methods comparable, we also consider the same 100 replications in which ML trees were estimated.

Table 5 shows estimates of the lengths of branches b_1 ($=b_3$), b_2 ($=b_4$), and b_5 of tree A (fig. 1) for the case of equal rates. These are average estimates of branch lengths for all the replications in which the correct topology was obtained. Here the values for b_1 and b_2 are the averages of the estimates for b_1 and b_3 and of the estimates for b_2 and b_4, respectively. The branch lengths for an MP tree were estimated by using Fitch's (1971

Table 3
***P* Obtained for Model Tree B in the Case of Varying Nucleotide Substitution Rate**

					NJG					
R	MP	DWP	NJD	NJK	$a = 0.5$	$a = 1$	$a = 2$	ML0.5	ML9	ML15
$a = 0.5$:										
0.5	1	2	2	2	89	43	15	48	23	18
9	0	69	2	29	75	43	22	20	70	71
15	0	74	3	23	70	41	20	20	76	77
$a = 1$:										
0.5	3	4	17	17	100	94	67	80	43	44
9	0	84	5	60	95	82	56	41	91	92
15	0	81	3	48	94	78	52	31	90	94
$a = 2$:										
0.5	2	3	51	51	100	100	95	97	80	80
9	0	86	9	77	97	94	81	53	94	94
15	0	86	5	68	96	94	80	34	94	95

method, whereas those for DWP were not computed, because they do not represent the number of nucleotide substitutions. As is well known, the MP method is expected to give underestimates of branch lengths when branches are long and multiple substitutions occur at the same sites. This is exactly the case for branch b_2 (and b_4), which is longest among the three branches examined. The branch length of b_5, of which the expected value is 0.05, is also slightly underestimated. The extent of underestimation of these branch lengths increases as the *R* value increases. However, the length of the shorter exterior branch b_1 (as well as b_3) tends to be overestimated, and the extent of overestimation slightly increases as *R* increases.

The branch length estimates for the NJ method depend on the distance measure used. (The branch length estimates for the NJ method are identical with those for the ME method, for the case of four sequences.) When the rate of nucleotide substitution is the same for all sites, NJK is supposed to give good estimates of

Table 4
***P* Obtained for Model Tree C**

					NJG					
R	MP	DWP	NJD	NJK	$a = 0.5$	$a = 1$	$a = 2$	ML0.5	ML9	ML15
Equal nucleotide substitution rate among different sites:										
0.5	98	98	98	98	99	99	99	98	98	96
9	96	96	98	98	99	99	98	98	99	99
15	96	95	98	98	99	98	98	96	98	98
$a = 0.5$:										
0.5	95	97	98	98	98	98	98	97	95	95
9	87	88	93	93	95	94	93	96	98	98
15	87	87	92	93	94	93	93	93	95	95
$a = 1$:										
0.5	97	97	98	98	98	98	98	99	99	99
9	93	90	97	97	97	97	97	97	97	97
15	93	92	96	96	97	97	97	98	97	97
$a = 2$:										
0.5	97	98	98	98	98	98	98	100	98	98
9	93	94	97	97	98	98	98	99	99	99
15	95	94	97	97	98	97	97	98	98	98

NOTE.—In the MP method only the case where the correct tree only was obtained was included.

Table 5
Average Branch Length Estimates ($\times 10^2$) for the Trees with Correct Topology in Model Tree A with Equal Rate

				NJG					
BRANCH	MP	NJD	NJK	$a = 0.5$	$a = 1$	$a = 2$	ML0.5	ML9	ML15
$R = 0.5$:									
b_1	5.6	5.0	5.0	−3.7	2.3	3.9	5.0	5.3	5.4
b_2	19.4	24.8	24.8	49.3	34.3	29.0	24.8	31.9	33.5
b_5	4.8	5.0	5.0	18.9	9.9	7.1	5.0	5.3	5.4
$R = 9$:									
b_1	5.9	5.7	5.1	−12.7	0.6	3.5	5.0	5.1	5.1
b_2	17.1	21.9	24.8	67.6	39.3	30.9	22.2	24.8	25.6
b_5	4.6	3.7	5.0	31.3	12.7	8.0	4.5	5.0	5.1
$R = 15$:									
b_1	6.0	5.8	5.1	−13.7	0.5	3.5	4.9	5.1	5.1
b_2	16.8	21.5	24.8	69.7	39.8	31.1	21.9	24.2	24.9
b_5	4.5	3.5	5.0	32.8	13.0	8.1	4.5	5.0	5.1

NOTE.—The true lengths of branches b_1, b_2, and b_5 are 0.05, 0.25, and 0.05, respectively. The standard errors of average branch length estimates were generally very small, so they are not presented.

branch lengths for all values of R, whereas NJD is expected to give good results only for $R = 0.5$. Indeed, these methods give good estimates of branch lengths for all of these cases. When $R > 0.5$, the Jukes-Cantor distance tends to give underestimates of nucleotide substitutions. Therefore, NJD gives underestimates of branch lengths for b_2 and b_5 but overestimates for b_1. This tendency is somewhat similar to that of the MP method.

As mentioned earlier, gamma distances give overestimates of the number of nucleotide substitutions when there is no variation in substitution rate among different sites. In the case of the NJ method, the lengths (l_1, l_2, and l_5, respectively) of branches b_1, b_2, and b_5 are given by

$$l_1 = d_{12}/2 - (d_{24}-d_{13})/4\,,$$
$$l_2 = d_{12}/2 + (d_{24}-d_{13})/4\,, \qquad (2)$$
$$l_5 = (d_{13}+2d_{14}+d_{24})/4 - d_{12}\,,$$

if $d_{14} = d_{23}$ and $d_{12} = d_{34}$. Therefore, if d_{24} is overestimated disproportionately compared with other distances, l_2 and l_5 are overestimated, whereas l_1 is underestimated. Table 5 clearly shows that this is the case. The extent of overestimation of l_2 and l_5 and the extent of underestimation of l_1 are most extreme when $a = 0.5$ and decline as a increases, as expected. In the case of $a = 0.5$, l_1 can be negative. However, despite the overestimation or underestimation of branch lengths, NJG gives the correct tree with a high probability, for the reason discussed earlier (table 1).

The ML method gives good estimates of branch lengths when the same R value as that used for sequence generation is used. Thus, ML0.5, ML9, and ML15 give good results for the cases of $R = 0.5$, 9, and 15, respectively. However, if ML9 and ML15 are used for the case of $R = 0.5$, all the branch lengths, especially b_2, are overestimated. By contrast, if ML0.5 and ML9 are used for the case of $R = 15$, the branch length for b_2 tends to be underestimated. The overestimation or underestimation of branch lengths when R is incorrectly assumed occurs because the incorrect assumption shifts the likelihood surface and the ML value is obtained for an incorrect set of branch length estimates (N. Takezaki, unpublished data).

The branch length estimates for model tree B are presented in table 6. With this model tree, MP produced the correct topology in only 4/100 replications for the case of $R = 0.5$. Yet, the average value for the branch length of b_5 for these four cases is not far from the true value (0.05). However, the branch length of b_2 is substantially underestimated, whereas that of b_1 is overestimated. When $R = 9$ or 15, no correct topologies were obtained (see table 1).

For $R = 0.5$, NJD and NJK give essentially the same branch length estimates, and the estimates are all close to the true values. For $R = 9$ and 15, NJD gives underestimates for branches b_2 and b_5 but overestimates for b_1. This pattern is the same as that for model tree A. By contrast, NJK gives fairly good estimates of branch lengths, but the branch length of b_1 tends to be underestimated, whereas that of b_5 tends to be overestimated. This underestimation or overestimation apparently oc-

Table 6
Average Branch Length Estimates ($\times 10^2$) for the Trees with Correct Topology in Model Tree B with Equal Rate

				NJG					
BRANCH	MP	NJD	NJK	$a = 0.5$	$a = 1$	$a = 2$	ML0.5	ML9	ML15
$R = 0.5$:									
b_1	6.4	5.0	5.0	−79.1	−12.6	−.8	5.1	5.8	5.9
b_2	32.4	49.8	49.8	204.5	93.8	67.2	49.8	307.0	530.5
b_5	5.6	4.9	4.9	99.4	26.0	12.2	4.8	5.3	5.5
$R = 9$:									
b_1	...	6.8	4.5	−301.7	−34.3	−6.1	4.9	4.9	4.9
b_2	...	39.4	50.4	482.4	129.1	77.2	41.0	50.0	53.9
b_5	...	3.0	5.7	335.0	50.7	18.7	4.6	5.2	5.4
$R = 15$:									
b_1	...	6.7	4.5	−371.5	−38.4	−6.6	4.9	5.0	5.0
b_2	...	37.9	50.4	563.7	136.2	79.0	39.9	46.9	50.0
b_5	...	2.4	5.6	405.5	54.7	19.0	4.5	4.8	5.1

NOTE.—The true lengths of branches b_1, b_2, and b_5 are 0.05, 0.50, and 0.05, respectively.

curred because the Kimura distance is known to give overestimates of nucleotide substitutions when the distance is large and the number of nucleotides used is relatively small (Tajima 1993). If the number of nucleotides used increases infinitely, all branch lengths are expected to converge to the correct values with the correct topology.

The extent of overestimation of pairwise distances by gamma distances increases as the true distance increases, as mentioned earlier. Since the true branch length of b_2 is two times greater in model tree B than in model tree A, gamma distances are expected to give an even higher degree of overestimation of branch lengths for b_2 and b_5 and of underestimation of branch length for b_1 in tree B than in tree A. This is indeed the case, as will be seen from table 6. Particularly in the case of NJG with $a = 0.5$, the extent of overestimation and underestimation is very serious. Nevertheless, the tree topology is estimated correctly with a high probability. This indicates that reconstruction of topology does not necessarily require good estimates of branch lengths.

The ML method gives good estimates of branch lengths when the correct R value is assigned in phylogenetic inference. However, when the assigned R value is higher than the true R value, the branch length of b_2 is overestimated. This overestimation is very serious when the true R value is 0.5. By contrast, when the assigned R value is smaller than the true value, branch lengths tend to be underestimated.

Case of Varying Rate

Table 7 shows the average branch length estimates for tree A, for the case of varying rate with $a = 0.5$. The MP method in this case gives even more underestimates than in the case of equal rate, except for b_1 (see table 5). This is reasonable because at sites with high substitution rate the extent of underestimation is serious, whereas at sites with low substitution rate the number of nucleotide substitutions is smaller than the average. NJD and NJK also give underestimates of branch lengths, except for b_1. Therefore, these methods have a property similar to that of MP, in branch length estimation. However, the extent of underestimation for b_2 is smaller in NJ than in MP, but that for b_5 is smaller in MP than in NJ.

Since the sequence data were generated with the assumption of varying rate with $a = 0.5$, NJG with $a = 0.5$ gives best estimates of branch lengths for all values of R. However, if we use $a = 1$ or 2 for estimating gamma distances, the branch length for b_5 is underestimated. The ML method also gives underestimates of all branch lengths.

Since the results for the case where sequence data were generated with the assumption of $a = 1$ were intermediate between those for the cases of $a = 0.5$ and $a = 2$, we shall not present them here. Instead, we shall discuss the results for the case of $a = 2$. In this case, NJG with $a = 2$ is expected to give good estimates of branch lengths. Table 8 shows that this is indeed the case. However, if $a = 0.5$ or 1 is used incorrectly for estimating gamma distances, the branch lengths for b_2 and b_5 tend to be overestimated, whereas the branch length of b_1 is underestimated. This occurs for the same reason as that mentioned in the case of equal rate. As noted earlier, however, this has an effect to produce the correct topology with a high probability. MP, NJD, NJK,

Table 7
Average Branch Length Estimates ($\times 10^2$) for the Trees with Correct Topology in Model Tree A with $a = 0.5$

				NJG					
BRANCH	MP	NJD	NJK	$a = 0.5$	$a = 1$	$a = 2$	ML0.5	ML9	ML15
$R = 0.5$:									
b_1	5.3	5.6	5.6	4.9	5.4	5.5	4.7	4.8	4.8
b_2	14.1	16.2	16.2	25.1	20.1	18.0	17.7	20.9	21.4
b_5	4.1	2.1	2.1	5.1	3.3	2.6	3.7	4.2	4.3
$R = 9$:									
b_1	5.1	5.3	5.2	4.8	5.3	5.4	4.2	4.4	4.4
b_2	11.9	13.8	15.3	25.3	19.1	16.7	15.0	16.5	16.8
b_5	3.8	1.9	2.2	5.4	3.3	2.6	3.6	3.6	3.6
$R = 15$:									
b_1	5.1	5.3	5.2	4.8	5.3	5.4	4.2	4.3	4.3
b_2	11.5	13.4	14.9	25.2	18.9	16.5	14.5	15.8	16.1
b_5	3.7	1.8	2.2	5.4	3.2	2.6	3.5	3.5	3.5

and ML all tend to give underestimates of branch lengths, except for b_1. In the case of $R = 0.5$, however, ML9 and ML15 tend to give overestimates of b_2, for the reason mentioned earlier. Because of this, ML9 tends to give rather good estimates of branch lengths, though the model of the ML method does not satisfy the condition for generating sequence data.

The branch length estimates for model tree B with $a = 0.5$ are given in table 9. As expected, NJG with $a = 0.5$ gives good estimates of branch lengths, though the estimate for b_1 tends to be an underestimate, whereas that for b_5 tends to be an overestimate. This underestimation or overestimation apparently occurs because the gamma distance is known to give biased estimates of nucleotide substitutions when the distance is long and the number of nucleotides examined is relatively small (Rzhetsky and Nei, accepted). To obtain good estimates of branch lengths for this case, we must use either a number of nucleotides $\gg 1{,}000$ or an unbiased estimator given by Rzhetsky and Nei. When NJG with $a = 1$ or 2 is used for estimating branch lengths, the branch lengths of b_2 and b_5 are underestimated, whereas that of b_1 is overestimated. MP, NJD, NJK, and ML all give underestimates of branch lengths, except for b_1, for the first three methods.

Table 10 shows the results for the case of model tree B with $a = 2$. In this case, NJG with $a = 2$ obviously gives good estimates of branch lengths, but NJG with a

Table 8
Average Branch Length Estimates ($\times 10^2$) for the Trees with Correct Topology in Model Tree A with $a = 2$

				NJG					
BRANCH	MP	NJD	NJK	$a = 0.5$	$a = 1$	$a = 2$	ML0.5	ML9	ML15
$R = 0.5$:									
b_1	5.6	5.6	5.6	1.5	4.4	5.1	5.0	5.2	5.2
b_2	17.6	21.6	21.6	38.7	28.5	24.8	22.4	28.0	29.1
b_5	4.5	3.5	3.5	11.6	6.5	4.8	4.5	4.9	5.0
$R = 9$:									
b_1	5.7	5.8	5.4	−0.1	3.8	4.9	4.7	4.8	4.8
b_2	15.3	18.9	21.2	46.0	30.2	24.9	19.7	21.8	22.3
b_5	4.2	2.8	3.7	16.2	7.6	5.1	4.2	4.5	4.6
$R = 15$:									
b_1	5.7	5.8	5.4	−0.2	3.7	4.9	4.6	4.8	4.8
b_2	14.9	18.5	21.0	46.9	30.4	24.9	19.3	21.2	21.7
b_5	4.2	2.7	3.7	16.7	7.7	5.2	4.1	4.4	4.5

Table 9
Average Branch Length Estimates ($\times 10^2$) for the Trees with Correct Topology in Model Tree B with $a = 0.5$

				NJG					
BRANCH	MP	NJD	NJK	$a = 0.5$	$a = 1$	$a = 2$	ML0.5	ML9	ML15
$R = 0.5$:									
b_1	6.4	6.2	6.2	4.2	6.1	6.4	4.7	4.7	4.7
b_2	21.0	27.6	27.6	50.6	36.1	30.7	29.1	39.3	41.5
b_5	4.8	1.5	1.5	5.9	2.7	2.1	3.7	4.6	4.9
$R = 9$:									
b_1	5.6	5.7	5.7	3.8	5.6	5.9	4.3	4.4	4.4
b_2	17.6	21.8	25.6	50.7	33.9	27.9	23.5	27.4	28.3
b_5	4.6	1.8	2.1	6.8	3.3	2.4	3.9	3.4	3.5
$R = 15$:									
b_1	...	6.0	6.0	3.9	5.8	6.1	4.3	4.4	4.4
b_2	...	21.0	24.3	50.4	33.2	27.0	22.2	25.8	26.6
b_5	...	1.6	1.8	6.9	3.1	2.2	3.9	3.3	3.3

= 0.5 or 1 gives overestimates of the branch lengths for b_2 and b_5 but underestimates for b_1, as in the case of model tree A. MP gives a serious underestimate for b_2 but a slight overestimate for b_1. NJD and NJK also give underestimates for b_2 and b_5. ML gives underestimates of branch lengths for b_2, except when $R = 0.5$ was used for ML9 and ML15. In the latter case, ML gives serious overestimates of the branch length.

In the case of model tree C, all the methods gave branch length estimates that were generally close to the true values, though the branch length of b_2 was either slightly overestimated or slightly underestimated when the assumptions were violated (data not shown). This indicates that when the extent of sequence divergence is small, even branch lengths can be estimated by any of the three methods. (The MP method showed a general tendency to give slight underestimates for b_2 but slight overestimates for b_1.)

Statistical Tests

One of the advantages of the ML method over the MP method is that it provides statistical tests of topological differences and of branch length estimates (Felsenstein 1981), and these tests are incorporated into DNAML. However, some of the tests, such as the branch length test, are very crude, as is mentioned in the DNAML manual. Yet, several authors (e.g., Gaut and Clegg 1991; Ward et al. 1991; Cooper et al. 1992) have

Table 10
Average Branch Length Estimates ($\times 10^2$) for the Trees with Correct Topology in Model Tree B with $a = 2$

				NJG					
BRANCH	MP	NJD	NJK	$a = 0.5$	$a = 1$	$a = 2$	ML0.5	ML9	ML15
$R = 0.5$:									
b_1	6.5	6.5	6.5	−17.9	1.1	5.0	5.0	5.4	5.7
b_2	27.6	39.9	39.9	111.5	63.9	49.9	41.9	129.9	432.3
b_5	5.1	2.6	2.6	33.5	10.3	5.1	4.3	5.1	5.1
$R = 9$:									
b_1	...	6.7	4.6	−26.5	0.5	4.6	4.7	4.8	4.7
b_2	...	31.6	41.7	131.4	112.2	48.0	33.9	41.1	43.6
b_5	...	2.4	4.6	44.2	65.1	5.6	4.3	4.5	4.7
$R = 15$:									
b_1	...	6.9	5.6	−36.7	−0.9	4.3	4.7	4.9	4.8
b_2	...	29.5	39.3	151.7	70.0	50.2	32.7	38.3	40.1
b_5	...	2.5	3.6	56.6	13.4	6.2	4.5	4.2	4.3

used these tests, disregarding the warning in the manual. We therefore examined the reliability of the branch length test in DNAML. There are two methods of testing branch lengths in DNAML. One is the likelihood-ratio test (LRT), and the other is the confidence-limit test (CLT). Both tests are conducted by making a number of simplifying assumptions (Felsenstein 1981, 1991). We applied these two methods to every set of sequence data generated for model trees A and B, to study the accuracy of the methods (100 replications for each parameter set).

In the present case we have four sequences, so there are three different unrooted trees (see fig. 3). If a test is valid, it should choose one of them, which is likely to be correct, and reject (or possibly not reject) the other two. In the case of the branch length test, this means that if the length of the interior branch of one topology is statistically significant, the lengths of interior branches for two other topologies should not be significant. In other words, the lengths of interior branches of two or three topologies cannot be simultaneously significant for the same set of sequence data. In practice, however, the application of the LRT and the CLT often produces results in which either all three or two of the three possible topologies are statistically significant, as shown in figure 3.

Table 11 shows all the results of the LRT and the CLT, for the case of model tree A. When all the assumptions of a likelihood model are satisfied, the results of the LRT and the CLT are quite reasonable; the lengths of interior branches of two or three topologies almost never become simultaneously significant. However, when the assumptions are not satisfied, there are many cases in which the lengths of interior branches of two or three topologies become simultaneously significant in both the LRT and the CLT. This is so even for the case of $a = 2$, where the variation in substitution rate is relatively mild. These results raise serious doubts about the general utility of the LRT and the CLT in DNAML, since variation in substitution rate is quite common. Although the results for model tree B are not presented, they were more seriously flawed than those for tree A.

We have done a similar test of interior branches for NJ trees by using Rzhetsky and Nei's (1992) ME method. (In the case of four sequences, the NJ and ME trees are identical with each other.) In this case, two or three topologies rarely show a significant interior branch length simultaneously even when the assumptions for estimating distances are not satisfied.

Table 12 shows the number of replications in which an incorrect tree became statistically significant by LRT and CLT in DNAML for tree A, whether the correct one also became significant or not. The ML method rarely identifies an incorrect tree as a correct tree, if all the assumptions of the model are satisfied. However, when the assumptions are not satisfied, it may identify an incorrect tree as a correct one, with a high probability. By contrast, this probability is quite low in the case of NJ trees (table 13). In the case of model tree B, however, the probability was appreciably high, though not as high as for the ML method.

The above results indicate that the statistical tests in DNAML are quite sensitive to violation of the assumptions of the ML models, and thus they should not be used as a general test. By contrast, the statistical tests proposed by Rzhetsky and Nei (1992) are quite robust when the assumptions are violated. Yet, when the distances for some pairs of sequences are very large, this test also may lead to an erroneous conclusion.

Discussion

The MP method is commonly used in phylogenetic inference from molecular data. As shown by Sourdis

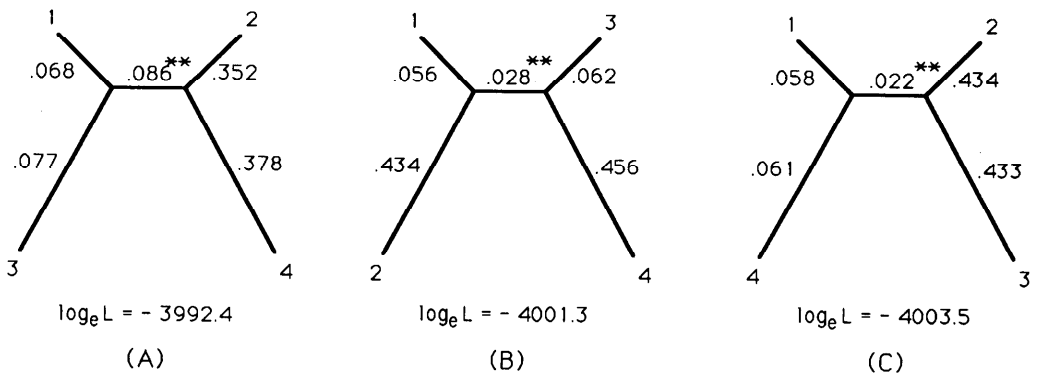

FIG. 3.—Three different trees obtained by the ML method for the same set of sequence data. The sequence data used here were generated under the assumption of equal rate with $R = 9$, but the trees were estimated by ML0.5. The true tree used is tree *B* in fig. 1. In this case, the true topology (*B*) has an ML value lower than that for the incorrect topology (*A*). The length of the interior branch was significant at the 1% level in all three topologies according to the LRT of DNAML. The lengths of all exterior branches were also statistically significant.

Table 11
Number of Replications in Which the Lengths of Interior Branches of Two or Three Different Topologies Became Simultaneously Significant by the Statistical Tests of DNAML for Model Tree A

	$R = 0.5$			$R = 9$			$R = 15$		
TEST[a]	ML0.5	ML9	ML15	ML0.5	ML9	ML15	ML0.5	ML9	ML15
Equal nucleotide substitution rate among different sites:									
LRT(3)	0	47	80	11	0	0	23	0	0
CLT(3)	0	27	48	5	0	0	11	0	0
LRT(2)	0	48	20	68	1	1	64	1	1
CLT(2)	0	63	51	66	1	1	71	0	0
$a = 0.5$:									
LRT(3)	2	92	98	81	11	15	88	9	9
CLT(3)	0	81	89	72	3	3	80	5	4
LRT(2)	73	8	2	19	55	50	12	59	55
CLT(2)	64	19	11	28	51	48	20	57	51
$a = 2$:									
LRT(3)	0	76	94	36	0	0	47	0	0
CLT(3)	0	51	72	18	0	0	35	0	0
LRT(2)	13	24	6	61	10	10	52	14	11
CLT(2)	8	48	28	77	3	4	62	8	6

NOTE.—The results for the case of $a = 1$ are not included, to save space. They were intermediate between those of the cases of $a = 0.5$ and $a = 2$.

[a] LRT(3) and LRT(2) represent the cases where the LRT showed a significant interior branch length for three and two different topologies, respectively, at the 5% level. CLT(3) and CLT(2) represent the cases where the CLT showed a significant interior branch length for three and two different topologies, respectively.

and Nei (1988) and Nei (1991), this method is useful when the number of nucleotide differences per site is small for all comparisons of sequences and when the number of nucleotides examined is very large. In other cases, however, this method is less powerful in obtaining the correct tree than other methods, such as the NJ and ML methods (see, e.g., Saitou and Imanishi 1989; Jin and Nei 1990). The present study shows that even when the MP method produces a consistent tree, its efficiency in obtaining the correct tree is lower than that of the NJ and ML methods (table 1).

The efficiency of obtaining the correct tree (P) in the NJ method depends on the distance measure used, if the extent of sequence divergence is high. Thus, with model tree B the NJ methods with different distance measures give different P values, particularly when the a value is small. It is therefore important to use appropriate distance measures if one wants to use the NJ method for phylogenetic inference. When there is an s/v bias and there is no other complication, this bias can easily be taken care of by using Kimura distance. If there is a sufficient amount of sequence data, it is also possible to estimate the a value (see, e.g., Uzzell and Corbin 1971; Kocker and Wilson 1991; Tamura and Nei 1993; Wakeley 1993). However, the actual pattern of nucleotide substitution is generally more complicated than the models used in this paper. It is therefore preferable to examine the suitability of various distance measures for the data set to be analyzed and to choose the best possible measure for distance estimation. Goldman (1993) and Tamura (1994) developed ML methods of model selection for this purpose, whereas A. Rzhetsky and M. Nei (unpublished data) developed another approach. If we use these methods, it is possible to choose a statistically efficient distance measure and use it for constructing an NJ tree.

It should be mentioned that application of these statistical methods to model selection is necessary only when the extent of sequence divergence is large. If the number of nucleotide substitutions per site for the most divergent sequences is relatively small (say ≤ 0.2), most distance measures give essentially the same estimate. In this case, it is sufficient to use Jukes-Cantor distance or Kimura distance (or even the proportion of nucleotide differences) for constructing NJ trees.

Some authors are concerned with the fact that the NJ method generates only one final tree and that this tree may not be the best one in terms of the criterion of ME. Actually, computer simulations (Saitou and Imanishi 1989; Rzhetsky and Nei 1992) have shown that in most cases the NJ tree has the same topology as that of the real (global) ME tree, unless the number of sequences used is very large. Therefore, in practice, one can regard the NJ tree as the ME tree. Of course, if one wants to find the real ME tree, one can use Rzhetsky and Nei's (1992) method. Rzhetsky and Nei (1993) have

Table 12
Number of Replications in Which the Length of the Interior Branch of a Wrong Tree was Statistically Significant by the LRT and the CLT of DNAML for Model Tree A

	$R = 0.5$			$R = 9$			$R = 15$		
TEST	ML0.5	ML9	ML15	ML0.5	ML9	ML15	ML0.5	ML9	ML15
Equal nucleotide substitution rate among different sites:									
LRT	0	95	100	79	1	1	87	1	1
CLT	0	90	99	71	1	1	82	0	0
$a = 0.5$:									
LRT	75	100	100	100	66	65	100	68	64
CLT	64	100	100	100	54	51	100	63	56
$a = 1$:									
LRT	23	100	100	99	26	28	99	33	26
CLT	13	100	100	99	13	15	99	17	11
$a = 2$:									
LRT	13	100	100	97	10	10	99	14	11
CLT	8	99	100	95	3	4	97	8	6

NOTE.—The significance level (type I error) used is 5%.

shown that the expected value of the sum of all branch lengths is smallest for the true tree, irrespective of the number of sequences and the topology of the tree, as long as unbiased estimates of evolutionary distances are used. Therefore, the ME (or NJ) method has a solid theoretical foundation.

For obtaining the correct tree, the ML method is slightly more sensitive to violation of the assumptions made than is the NJ method with gamma distances, when the evolutionary distance is long (table 3). The branch length estimates obtained by the ML method are also affected substantially by violation of the assumptions. Of course, as mentioned earlier, it is possible to compute the likelihood values under different mathematical models and choose the model that maximizes the likelihood. This feature has already been incorporated into Felsenstein's DNAML program, with respect to R. However, this adds more computational time to the ML method, which is already computation intensive. Furthermore, the enhancement of the P value by this procedure does not necessarily make the ML method better than the NJ method.

In the present study, we used a large number of iterations for obtaining the ML value, to increase the accuracies of the topology and branch length estimates of an ML tree, as mentioned earlier. We could do this because we used only four DNA sequences. When the number of sequences is large, however, this will increase the computational time tremendously. Therefore, in actual data analysis it will be necessary to limit the number of iterations to ~ 10, as in DNAML. Our preliminary study has shown that this procedure decreases the accuracy of the ML tree, particularly of the branch length estimates.

As mentioned earlier, table 4 shows that when the extent of sequence divergence is small, all tree-making methods give essentially the same results. In this case, therefore, there is no need to use a time-consuming tree-making method such as the MP or the ML method. Furthermore, the MP method is expected to produce many equally parsimonious trees for this case, unless a large number of nucleotides are examined, so that it is difficult to know the real splitting pattern of sequences (see, e.g., Hedges et al. 1992; Brown et al. 1993). The NJ or ME method usually does not have this problem and gives a unique final tree.

Table 13
Number of Replications in Which the Length of the Interior Branch of a Wrong NJ Tree was Stastically Significant by the Rzhetsky-Nei Test for Model Tree A

	$R = 0.5$		$R = 9$		$R = 15$	
	NJD	NJK	NJD	NJK	NJD	NJK
Equal rate	0	0	0	0	0	0
$a = 0.5$	3	3	4	1	3	2
$a = 1$	0	0	2	0	4	1
$a = 2$	0	0	2	2	3	2

NOTE.—NJG was used because the mathematical formula for the covariance of gamma distances was not available. The significance level (type I error) used is 5%.

The present study is based on a simple case of four sequences, so that some of our conclusions may not be applicable to the case of a larger number of sequences. It seems to be important to extend this type of simulation to these cases in the future.

Addendum

After this paper was submitted for publication, Yang (1993) developed an algorithm for obtaining an ML tree by taking into account continuous variation in substitution rate among different nucleotide sites. However, this method requires much more computational time than does DNAML, and it seems to be difficult to use when the number of sequences is moderately large (five or more sequences).

Acknowledgment

This work was supported by research grants from the National Institute of Health (GM-20293) and the National Science Foundation (DEB-9119802) to M.N.

APPENDIX

Mathematical Models Used in the ML Method of Phylogenetic Inference

The model of nucleotide substitution in Felsenstein's DNAML program takes into account the differences in rates of transition and transversion and the base compositions. This model is different from Hasegawa et al.'s (1985) model, in which the matrix of substitution rates is

	A	T	C	G
A		βg_T	βg_C	αg_G
T	βg_A		αg_C	βg_G
C	βg_A	αg_T		βg_G
G	αg_A	βg_T	βg_C	

(A1)

An element λ_{ij} of this matrix represents the rate of substitution from nucleotide i to j (i, j = A, T, C, G). All the elements in each row sum to one, so that a diagonal element $\lambda_{ii} = 1 - \sum_j \lambda_{ij}$ ($i \neq j$), though this is not presented. g_i is the proportion of the ith nucleotide. The transition rate and the transversion rate from nucleotide i to j are αg_j and βg_j, respectively.

In this model the proportions of transitional (P) and transversional (Q) nucleotide differences between a pair of sequences that diverged time t ago are given by

$$P = 2\left[B+(A-B)e^{-2\beta t} - \frac{g_T g_C}{g_Y} e^{-2(g_Y\alpha+g_R\beta)t} - \frac{g_A g_G}{g_R} e^{-2(g_R\alpha+g_Y\beta)t}\right], \quad (A2)$$

$$Q = 2C(1-e^{-2\beta t}),$$

where $g_Y = g_T + g_C$, $g_R = g_A + g_G$, $A = g_T g_C/g_Y + g_A g_G/g_R$, $B = g_T g_C + g_A g_G$, and, $C = g_Y g_R$ (Hasegawa et al. 1985).

The model of nucleotide substitution in Felsenstein's DNAML program is not given explicitly in the manual of the program but is described by Kishino and Hasegawa (1989). The substitution matrix in this model can be written in the following way:

	A	T	C	G
A		βg_T	βg_C	$(\delta/g_R+\beta)g_G$
T	βg_A		$(\delta/g_Y+\beta)g_C$	βg_G
C	βg_A	$(\delta/g_Y+\beta)g_T$		βg_G
G	$(\delta/g_R+\beta)g_A$	βg_T	βg_C	

(A3)

The diagonal element λ_{ii} in the matrix is again given by $1 - \sum_j \lambda_{ij}$ ($i \neq j$). The transversion rate βg_j is identical to that in Hasegawa et al.'s model. However, the transition rate, which is αg_j in Hasegawa et al.'s model, is assumed to be different for pyrimidines ($\alpha_Y g_j$) and for purines ($\alpha_R g_j$), where

$$\alpha_Y = \frac{\delta}{g_Y} + \beta, \qquad \alpha_R = \frac{\delta}{g_R} + \beta, \quad (A4)$$

$g_Y = g_T + g_C$, and $g_R = g_A + g_G$. Thus, α_Y and α_R are each a sum of β (a parameter that determines the transversion rate) and the amount (δ/g_Y, δ/g_R) of transitional change that exceeds β. When $g_Y = g_R$, we have $\alpha_Y = \alpha_R = \alpha$, and this model reduces to Hasegawa et al.'s.

Note that R in the DNAML model is given by

$$R = (A\delta/\beta+B)/C. \quad (A5)$$

In the case of the Jukes-Cantor model, where $g_i = 0.25$ and $\delta = 0$, R becomes 0.5. In the case of the Kimura two-parameter model, where $g_i = 0.25$, R is given by $\delta/\beta + 1/2 = \alpha/(2\beta)$.

With this model we can derive the following equations for P and Q:

$$P = 2[B+(A-B)e^{-2\beta t}-Ae^{-2(\delta+\beta)t}]; \qquad Q = 2C(1-e^{-2\beta t}). \quad (A6)$$

The simplicity of these equations compared with equation (A2) makes it possible to derive an analytical formula for estimating the number of nucleotide substitutions per site ($d = 2(1-\sum_i g_i\lambda_{ii})t$). The estimates ($\hat{d}$) of d and of its variance [$V(\hat{d})$] are given by

$$\hat{d} = -2A \log\left(1 - \frac{\hat{P}}{2A} - \frac{A-B}{2AC}\hat{Q}\right) + 2(A-B-C)\log\left(1 - \frac{\hat{Q}}{2C}\right), \quad \text{(A7)}$$

and

$$V(\hat{d}) = \frac{1}{n}[a^2\hat{P} + b^2\hat{Q} - (a\hat{P} + b\hat{Q})^2], \quad \text{(A8)}$$

where n is the number of nucleotides examined, and

$$a = \frac{AC}{AC - C\hat{P}/2 - (A-B)\hat{Q}/2},$$

$$b = \frac{A(A-B)}{AC - C\hat{P}/2 - (A-B)\hat{Q}/2} - \frac{A-B-C}{C - \hat{Q}/2}.$$

LITERATURE CITED

BROWN, J. R., A. T. BECKENBACH, and M. J. SMITH. 1993. Intraspecific DNA sequence variation of the mitochondrial control region of white sturgeon (*Acipenser transmontanus*). Mol. Biol. Evol. **10**:326–341.

COOPER, A., C. MOURER-CHUVIRÉ, G. K. CHAMBERS, A. VON HAESELER, A. C. WILSON, and S. PÄÄBO. 1992. Independent origins of New Zealand moas and kiwis. Proc. Natl. Acad. Sci. USA **89**:8741–8744.

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

ECK, R. V., and M. O. DAYHOFF. 1966. Atlas of protein sequence and structure. National Biomedical Research Foundation, Silver Springs, Md.

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

———. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

———. 1991. PHYLIP (phylogeny inference package), version 3.4. University of Washington, Seattle.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.

FITCH, W. M., and E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science **155**:279–284.

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

GAUT, B. S., and M. T. CLEGG. 1991. Molecular evolution of alcohol dehydrogenase 1 in members of the grass family. Proc. Natl. Acad. Sci. USA **88**:2060–2064.

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

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating o the human-ape splitting by a molecular clock of mitochon drial DNA. J. Mol. Evol. **22**:160–174.

HEDGES, S. B., S. KUMAR, K. TAMURA, and M. STONEKING 1992. Human origins and analysis of mitochondrial DN sequences. Science **255**:737–739.

HEDGES, S. B., K. D. MOBERG, and L. R. MAXSON. 1990 Tetrapod phylogeny inferred from 18S and 28S ribosoma RNA sequences and a review of the evidence for amniot relationships. Mol. Biol. Evol. 7:607–633.

JIN, L., and M. NEI. 1990. Limitations of the evolutionar parsimony method of phylogenetic analysis. Mol. Biol. Evol **7**:82–102.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protei molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalia protein metabolism. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolu tionary rate of base substitutions through comparativ studies of nucleotide sequences. J. Mol. Evol. **16**:111–120

KISHINO, H., and M. HASEGAWA. 1989. Evaluation of th maximum likelihood estimate of the evolutionary tree to pologies from DNA sequence data and the branching orde in hominoidea. J. Mol. Evol. **29**:170–179.

KOCHER, T. D., and A. C. WILSON. 1991. Sequence evolutio of mitochondrial DNA in humans and chimpanzees: con trol region and a protein-coding region. Pp. 391–413 *in* S OSAWA and T. HONJO, eds. Evolution of life. Springer Tokyo.

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

———. 1991. Relative efficiencies of different tree-makin methods for molecular data. Pp. 90–128 *in* M. M. MIYA MOTO and J. CRACRAFT, eds. Phylogenetic analysis of DN sequence. Oxford University Press, New York.

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

RZHETSKY, A., and M. NEI. 1992. A simple method for esti mating and testing minimum-evolution trees. Mol. Biol Evol. **9**:945–967.

———. 1993. Theoretical foundation of the minimum-evo lution method of phylogenetic inference. Mol. Biol. Evol **10**:1073–1095.

———. Unbiased estimates of the number of nucleotide sub stitutions when substitution rate varies among different site J. Mol. Evol. (accepted).

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

SAITOU, N., and M. NEI. 1987. The neighbor-joining method a new method for reconstructing phylogenetic trees. Mol Biol. Evol. **4**:406–425.

SANKOFF, D. D., and R. J. CEDERGREN. 1983. Simultaneou comparison of three or more sequences related by a tree Pp. 253–263 *in* D. SANKOFF and J. B. KRUSKAL, eds. Tim

warps, string edits, and macromolecules: the theory and practice of sequence comparison. Addison-Wesley, Reading, Mass.

SOURDIS, J., and M. NEI. 1988. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. Mol. Biol. Evol. **5:** 298–311.

TAJIMA, F. 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **10:**677–688.

TAMURA, K. 1994. Model selection in the estimation of the number of nucleotide substitutions. Mol. Biol. Evol. **11:** 146–149.

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.

UZZELL, T., and K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. Science **172:** 1089–1096.

VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES, and A. C. WILSON. 1991. African populations and the evolution of human mitochondrial DNA. Science **253:**1503–1507.

WAKELEY, J. 1993. Substitution rate variation among sites in hypervariable region I of human mitochondrial DNA. J. Mol. Evol. **37:**613–623.

WARD, R. H., B. L. FRAZIER, K. DEW-JAGER, and S. PÄÄBO. 1991. Extensive mitochondrial diversity within a single Amerindian tribe. Proc. Natl. Acad. Sci. USA **88:**8720–8724.

WILLIAMS, P. L., and W. M. FITCH. 1990. Phylogeny determination using dynamically weighted parsimony method. Pp. 615–626 *in* R. F. DOOLITTLE, ed. Methods in enzymology. Vol. **183:** Molecular evolution: computer analysis of protein and nucleic acid sequences. Academic Press, New York.

YANG, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10:**1396–1401.

MICHAEL BULMER, reviewing editor

Received May 13, 1993

Accepted September 27, 1993