Efficiencies of Fast Algorithms of Phylogenetic Inference Under the Criteria of Maximum Parsimony, Minimum Evolution, and Maximum Likelihood When a Large Number of Sequences Are Used

Kei Takahashi and Masatoshi Nei

Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University

In phylogenetic inference by maximum-parsimony (MP), minimum-evolution (ME), and maximum-likelihood (ML) methods, it is customary to conduct extensive heuristic searches of MP, ME, and ML trees, examining a large number of different topologies. However, these extensive searches tend to give incorrect tree topologies. Here we show by extensive computer simulation that when the number of nucleotide sequences (*m*) is large and the number of nucleotides used (*n*) is relatively small, the simple MP or ML tree search algorithms such as the stepwise addition (SA) plus nearest neighbor interchange (NNI) search and the SA plus subtree pruning regrafting (SPR) search are as efficient as the extensive search algorithms such as the SA plus tree bisection-reconnection (TBR) search in inferring the true tree. In the case of ME methods, the simple neighbor-joining (NJ) algorithm is as efficient as or more efficient than the extensive NJ+TBR search. We show that when ME methods are used, the simple p distance generally gives better results in phylogenetic inference than more complicated distance measures such as the Hasegawa-Kishino-Yano (HKY) distance, even when nucleotide substitution follows the HKY model. When ML methods are used, the simple Jukes-Cantor (JC) model of phylogenetic inference generally shows a better performance than the HKY model even if the likelihood value for the HKY model is much higher than that for the JC model. This indicates that at least in the present case, selecting of a substitution model by using the likelihood ratio test or the AIC index is not appropriate. When *n* is small relative to *m* and the extent of sequence divergence is high, the NJ method with p distance often shows a better performance than ML methods with the JC model. However, when the level of sequence divergence is low, this is not the case.

Introduction

Most of the currently used methods of phylogenetic inference from molecular data are based on some form of optimization principle (Nei 1996; Swofford et al. 1996). Under this principle, the preferred tree is determined by assigning an optimality score to all possible topologies (or all potentially correct topologies) according to a certain procedure and choosing the topology that shows the highest or lowest optimal score. For example, in the case of maximum-parsimony (MP) methods (Eck and Dayhoff 1966; Fitch 1971), the minimum number of evolutionary changes that explains the entire sequence evolution (tree length [TL]) is computed for each topology, and the topology showing the smallest TL value is chosen as the preferred tree (MP tree). Similarly, in minimum-evolution (ME) methods (Edwards and Cavalli-Sforza 1963; Saitou and Imanishi 1989; Rzhetsky and Nei 1992), the total sum of branch length estimates (*S*) is used as the optimality score, and the topology showing the smallest *S* value is chosen as the preferred tree (ME tree). In contrast, maximum-likelihood (ML) methods use the log likelihood value (ln*L*) for each topology as the optimality score, and the tree showing the highest ln*L* is chosen as the ML tree (Felsenstein 1981).

One problem with the methods based on the optimization principle is that an enormous amount of computational time is required when the number of sequences is large, especially with ML methods. For this reason, a number of heuristic search algorithms that attempt to find the optimal tree are currently used (Swofford and Begle 1993). Yet, it is customary for many researchers to make heroic efforts to find the optimal tree, spending a large amount of computer time (Maddison 1991; Chase et al. 1993; Wainright et al. 1993; Rice, Donoghue, and Olmstead 1997). However, it should be noted that the tree with the optimal score is not necessarily the true tree. Nei, Kumar, and Takahashi (1998) showed that when the number of nucleotides examined (*n*) is small and the number of sequences (*m*) is large, the optimality scores of the MP and ME trees are always smaller than or equal to those of the true tree, whereas the optimality score of the ML tree is always greater than or equal to that of the true tree. This indicates that the optimization principle tends to give incorrect topologies when *n* is small relative to *m.* Nei, Kumar, and Takahashi (1998) also showed that for obtaining the true tree (not the optimal tree), simple topology search algorithms are often as efficient as extensive search algorithms. In the case of ME methods, a simple algorithm called neighbor joining (NJ; Saitou and Nei 1987) was shown to be as efficient as the standard ME method in almost all cases examined. However, the simple algorithms used for MP and ML methods were ad hoc and did not work well under certain conditions.

The main purpose of this paper is to examine the efficiencies of various search algorithms for MP, ME, and ML trees when the number of sequences used is large and to find simple algorithms for inferring the true tree with a reasonably high level of accuracy. We study this problem by computer simulation and by examining the relationships between the efficiencies of inferring the optimal and the true trees. We also examine the effi-

Key words: phylogenetic inference, maximum parsimony, minimum evolution, maximum likelihood, topological distance, tree-building algorithms.

Address for correspondence and reprints: Masatoshi Nei, 328 Mueller Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802. E-mail: nxm2@psu.edu.

Mol. Biol. Evol. 17(8):1251–1258. 2000

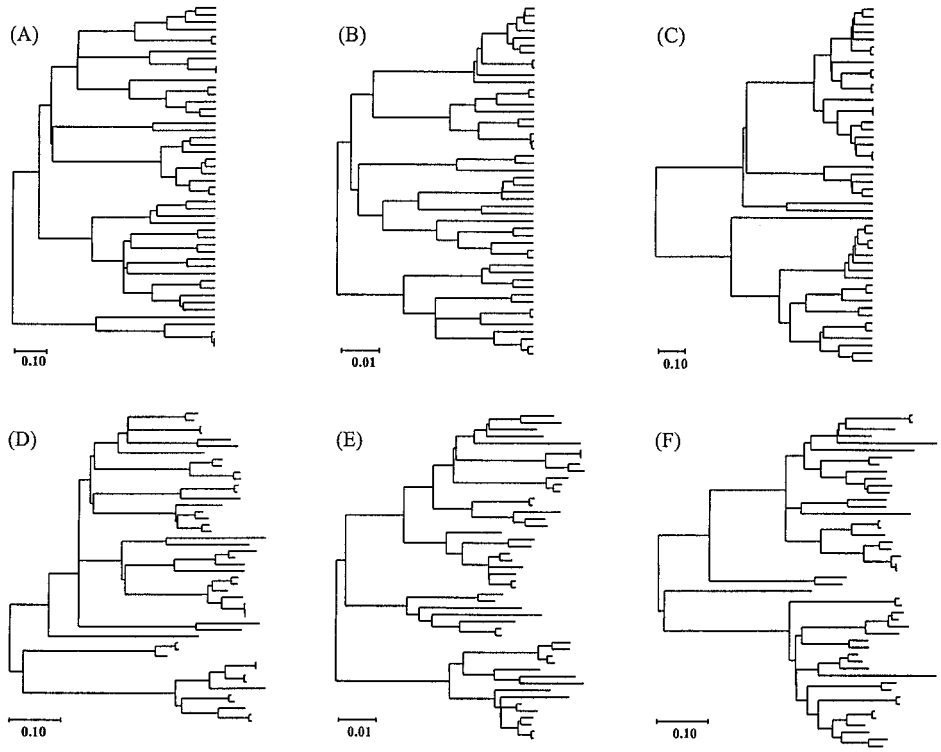

FIG. 1.—Six different model trees used for computer simulation. Trees A–C represent the cases of constant-rate evolution, and trees D–F represent the cases in which the substitution rate varies with evolutionary lineage. These trees were randomly generated by the methods described in the text.

ciencies of inferring the true tree when wrong nucleotide substitution models are used.

Methods of Computer Simulation

Model Trees

To study the efficiencies of various tree-building algorithms, we need model trees with which we can simulate the evolution of nucleotide sequences. Once the sequence data at the final stage of evolution are obtained, we can construct an MP, ME, or ML tree following a given tree-building algorithm and then compare the topology of the tree with that of the model tree (Nei 1991). We used six different model trees of 48 sequences, which are presented in figure 1. Model trees A, B, and C represent the cases of constant-rate (CR) evolution, and model trees D, E, and F represent the cases in which the evolutionary rate varies from branch to branch (varying rate [VR]). The model trees with CR evolution were randomly generated by the branching process method described by Kuhner and Felsenstein(1994). The model trees with VR evolution were produced by modifying the branch lengths of the CR trees obtained by Kuhner and Felsenstein's (1994) method. This modification was done by replacing each branch length by a random variable that followed a gamma distribution with the mean equal to the original branch length and the gamma shape parameter (a) equal to 10 (see Tateno, Nei, and Tajima [1982] for a similar method). The six model trees used here were generated independently. We also considered two levels of sequence divergence: high divergence (model trees A, C, D, and F) and low divergence (model trees B and E). High divergence refers to the case in which the expected number of nucleotide substitutions per site from the root to the tip of the tree is 0.5, and thus the expected number of substitutions between two most distantly related sequences (d_{max}) is 1.0. Low divergence refers to the case in which the expected number of substitutions from the root to the tip is 0.05 and d_{max} is 0.1.

Our simulations were not done for all combinations of sequence levels and tree topologies because of the prohibitive amount of computer time required. However, our preliminary study indicated that different model trees with a constant rate or a varying rate give essentially the same results for a given level of sequence divergence, apparently because all of the model trees were generated at random.

Models of Nucleotide Substitution

We used two different models of nucleotide substitution to generate sequence data that are commonly used for phylogenetic reconstruction: Jukes and Cantor's (1969) (JC) model and Hasegawa, Kishino, and Yano's (1985) model with a gamma distribution of substitution rate among different nucleotide sites (HKY+Γ). In the latter model, we used two different sets of substitution parameters. In the first set, we assumed that the transition/transversion rate ratio (k) was equal to 4, and the equilibrium frequencies of nucleotides A, T, C, and G were as follows: $g_A = 0.15$, $g_T = 0.15$, $g_C = 0.35$, and

$g_G = 0.35$, respectively. The gamma parameter (a) used in this case was 1.0. In the second set, we assumed $k = 10$, $g_A = 0.10$, $g_T = 0.10$, $g_C = 0.40$, $g_G = 0.40$, and $a = 0.5$. Note that in the present model, $k = 2R$, where R is the standard transition/transversion ratio (Kumar, Tamura, and Nei 1993; Nei and Kumar 2000). The numbers of nucleotides used (n) were 300 for high sequence divergence and 1,000 for low sequence divergence. For each model tree, we generated 100 random sets of sequence data and constructed MP and ME trees for each data set. However, for ML methods, we used only the first 50 data sets, because the computational time was prohibitively long.

Phylogenetic Reconstruction

In this paper, we used five different algorithms for constructing MP trees that are available in the computer program PAUP* (Swofford 1998). The simplest algorithm used was stepwise addition with the closest option (SA). This algorithm is a rough search of MP trees, and it often fails to find the true optimal (MP) tree. Therefore, the tree obtained by this algorithm is usually subjected to various branch-swapping algorithms. The branch-swapping algorithms generally used are nearest-neighbor interchange (NNI), subtree pruning and regrafting (SPR), and tree bisection-reconnection (TBR) (see Swofford and Begle [1993] for the actual procedures). Among these algorithms, the simplest one is NNI, and the most extensive one is TBR. In this paper, these algorithms are denoted by SA+NNI, SA+SPR, and SA+TBR. To search for the most likely MP tree, however, researchers often repeat the TBR search many times starting with different SA trees obtained by rearranging the input order of sequences. We therefore used an algorithm in which the TBR search was repeated 100 times. This algorithm is denoted by 100SA+TBR.

For ME methods, we used the NJ method as the fast algorithm, because this method is known to be very fast and as efficient as or more efficient than the exhaustive ME method in obtaining the true tree when the number of sequences used is small (Saitou and Imanishi 1989; Kumar 1996; Gascuel 1997; Nei, Kumar, and Takahashi 1998). As the extensive search algorithm, we used the NJ+TBR algorithm, in which the NJ tree is subjected to the TBR search. We used this algorithm because the TBR search does not require much computational time for ME trees. For computing pairwise distances, we used p and JC distances when the JC model was used for generating sequence data and p, JC, and HKY+Γ distances when the HKY+Γ model was used for sequence generation. The HKY+Γ distance was estimated by the maximum-likelihood algorithm available in PAUP*.

For ML methods, we used primarily the SA (''as is'' option), SA+NNI, and SA+SPR search algorithms. The SA+TBR search was used only for a few cases because this algorithm required an enormous amount of computer time. When the JC model was used for generating sequence data, we used the same model for constructing trees. When sequence data were generated by the HKY+Γ model, we used both the JC and the HKY+Γ models for tree reconstruction. For the HKY+Γ model, we first constructed an ML tree using the SA algorithm with the JC model and obtained ML estimates of parameters k and a and nucleotide frequencies. Using these parameter estimates, we then constructed ML trees with the HKY+Γ model using the SA, SA+NNI, and SA+SPR algorithms. We used this simplified method to save computer time.

Accuracies of Estimated Tree Topologies

Since we used randomly generated model trees of 48 sequences and some of the interior branches of a realized tree (Nei and Kumar 2000) often had 0 nucleotide substitutions, the probability of obtaining the true topology was negligibly small. We therefore measured the accuracy of the estimated topology by the topological distance (d_T) between the estimated tree and the true tree (Robinson and Foulds 1981; Penny and Hendy 1985; Rzhetsky and Nei 1992). In practice, since our simulations were replicated 100 or 50 times, we used the average d_T value. This d_T value is roughly twice the number of branch interchanges that are required to obtain the true topology from the estimated tree when both trees are bifurcating. For MP methods, several equally parsimonious trees were often obtained for the same data set. In this case, we computed the average d_T value for all parsimonious trees for each replication and then obtained the average of this average d_T value for all replications. An alternative method is to construct the strict-consensus tree for all MP trees for a given data set and to compute d_T between this consensus tree and the true tree using Rzhetsky and Nei's (1992) formula. We can then take the average of this d_T for all replications. In practice, however, these two methods give very similar average d_T values, so we present only the former average here.

The main purpose of this paper is to find relatively simple algorithms that are as efficient as extensive search algorithms in inferring the true tree. However, to examine the efficiencies of various optimization algorithms, we also computed the average optimality scores of MP, ME, and ML trees obtained by the simple and extensive search algorithms.

Results

Table 1 shows the average topological distances from the true tree (d_T) and the average TLs of MP trees obtained by five different algorithms. When model tree A (CR model) is used and the sequence data are generated by the JC model, the d_T value for the simplest algorithm, SA, is 10.9, and it declines to 9.5 when the next-simplest algorithm, SA+NNI, is used. However, when the more extensive search algorithms SA+SPR, SA+TBR, and 100SA+TBR are used, d_T remains virtually the same. This means that in the present case, the SA+NNI search is quite efficient in inferring the true tree, and there is no need to conduct more extensive searches. However, the SA+NNI search is better than the SA search. The most extensive search,

Table 1
Average Topological Distances from the True Tree (d_T) (±SE) and Tree Lengths (TLs) of Inferred Trees Under the Maximum-Parsimony Criterion

	SA		SA+NNI		SA+SPR		SA+TBR		100SA+TBR	
	d_T	TL	d_T	TL	d_T	TL	d_T	TL	d_T	TL
JC model: high divergence (d_{max} = 1.0)										
CR (tree A) . .	10.9 ± 0.3	2,299.0	9.5 ± 0.3	2,292.8	9.4 ± 0.3	2,291.5	9.4 ± 0.2	2,291.5	9.6 ± 0.3	2,291.5
VR (tree D) . .	9.8 ± 0.4	1,457.0	10.2 ± 0.3	1,453.2	9.9 ± 0.3	1,452.7	9.7 ± 0.3	1,452.7	9.8 ± 0.3	1,452.6
JC model: low divergence (d_{max} = 0.1)										
CR (tree B) . .	5.3 ± 0.3	1,762.2	4.6 ± 0.2	1,761.4	4.6 ± 0.2	1,761.4	4.58 ± 0.2	1,761.4	4.58 ± 0.2	1,761.4
VR (tree E) . .	6.9 ± 0.3	1,742.8	5.9 ± 0.2	1,741.8	5.9 ± 0.2	1,741.8	5.9 ± 0.2	1,741.8	5.9 ± 0.2	1,741.8
HKY+Γ model: k = 4, g_A = 0.15, g_C = 0.35, g_G = 0.35, g_T = 0.15, a = 1, d_{max} = 1.0										
CR (tree C) . .	18.0 ± 0.5	2,011.4	16.6 ± 0.5	2,002.0	16.0 ± 0.4	2,000.0	16.0 ± 0.4	2,000.0	16.1 ± 0.4	1,999.9
VR (tree F). . .	16.6 ± 0.5	1,998.7	15.3 ± 0.4	1,990.6	14.7 ± 0.4	1,989.6	14.8 ± 0.4	1,989.5	14.6 ± 0.4	1,989.3

NOTE.—Results are from 100 replicate simulations. Average TLs are also shown. SA = stepwise addition; NNI = nearest-neighbor interchange; SPR = subtree pruning and regrafting; TBR = tree bisection-reconnection; CR = constant-rate model; VR = varying-rate model; JC = Jukes-Cantor; HKY = Hasegawa-Kishino-Yano.

100SA+TBR, gave a value of d_T = 9.6, which is slightly higher than that for less extensive searches (except SA). This happened because the MP trees identified by this extensive search algorithm were not necessarily closer to the true tree.

Note that in the present case, the minimum and maximum possible values of d_T are 0 and 90, respectively, because there are 45 interior branches for an unrooted tree of 48 sequences (Nei and Kumar 2000). Note also that d_T = 10.9 means an error in branching pattern (sequence partition) for 5.5 interior branches, whereas d_T = 9.5 means an error for about 5 interior branches. Therefore, the difference in d_T between 10.9 and 9.5 is rather small biologically, although the difference is statistically significant.

Table 1 also shows the average TLs for MP trees. This value is highest for SA and becomes smaller for SA+NNI and SA+SPR, but the TLs for SA+SPR, SA+TBR, and 100SA+TBR are essentially the same. In the present case, therefore, even TL does not decrease when extensive search algorithms are used.

When the high-divergence and VR model (tree D) is used, the d_T values are nearly the same as those for model tree A. In this case, however, there are no significant differences in d_T between different search algorithms. Actually, the SA+NNI search shows a slightly higher d_T value (10.2) than the SA search (9.8), although the difference is not statistically significant. This indicates that in the present case, the SA search is as efficient as the 100SA+TBR search. This is so despite the fact that the average TL is smaller for 100 SA+TBR than for SA.

For low divergence (d_{max} = 0.1), the d_T values are substantially smaller than those for high divergence, as expected. In this case, the d_T for SA is significantly greater than that for SA+NNI or other search algorithms. However, for inferring the true tree, SA+NNI is sufficient for both CR and VR cases, and there is no need to conduct the SA+TBR or 100SA+TBR search.

Table 1 also shows the d_T values for the CR and VR cases when sequence data were generated by the HKY+Γ model with a high divergence level (d_{max} = 1.0). The d_T values are considerably greater than those for sequence data generated by the JC model. This indicates that the MP method is less efficient when the pattern of nucleotide substitution is complex than when it is simple. In this case, SA+SPR is slightly better than SA+NNI, but the d_T values for SA+TBR and 100SA+TBR are nearly the same as the d_T for SA+SPR.

Table 2 shows the values of d_T and S for ME trees when the sequence data were generated by the JC model. In this case, two distance measures, p distance and JC distance, were used. For both distance measures, NJ, a simplified ME method, always shows a smaller d_T than the NJ+TBR search. Note also that the p distance gives a smaller d_T than the JC distance in all cases, although the sequence data were generated by the JC model. This is true for both CR and VR models, irrespective of the extent of sequence divergence.

Table 2 also shows the simulation results (d_T and negative log ML value [$-\ln L$]) for ML methods. In this case, SA is less efficient than SA+NNI and SA+SPR in inferring the true tree. The latter two algorithms show nearly the same d_T value. Therefore, the SA+NNI search appears to be sufficient for inferring the true tree by ML methods. For the first 10 data sets (model tree D), we conducted the SA+TBR search, but the d_T value was essentially the same as that for SA+NNI (data not shown). (In the present case, the SA+TBR search for an ML tree required a computational time of about 60 h for one replication with our 8500/132 PowerPC computer.) The $-\ln L$ value is also greater for SA than for SA+NNI and SA+SPR, but the latter two algorithms show essentially the same $-\ln L$ value.

Table 3 shows the results for ME and ML methods for the case in which the sequence data were generated by the HKY+Γ model. Here, we considered only the case of high divergence (d_{max} = 1.0) with the first and second parameter sets of the HKY+Γ model mentioned earlier (see also table 3). Because the HKY+Γ model is more complicated than the JC model, the d_T values for this model are higher than those for the JC model (table 2). For ME methods, the NJ algorithm again shows a

Table 2
Average Topological Distances from the True Tree (d_T) (±SE), Sums of Branch Lengths (S), and Negative Log Likelihood Values ($-\ln L$) of Inferred Trees Under the Minimum-Evolution (ME) and Maximum-Likelihood (ML) Criteria

	ME				ML					
	NJ		NJ+TBR		SA		SA+NNI		SA+SPR	
	d_T	S	d_T	S	d_T	$-\ln L$	d_T	$-\ln L$	d_T	$-\ln L$
JC model: high divergence (d_{max} = 1.0)										
CR (tree A)										
p.....	8.7 + 0.3	6.26	9.1 ± 0.3	6.25						
JC ...	11.6 ± 0.4	9.32	12.0 + 0.4	9.29	12.8 ± 0.5	9,763.8	9.0 ± 0.4	9,752.3	8.9 ± 0.4	9,752.2
VR (tree D)										
p.....	9.1 ± 0.4	4.18	9.5 ± 0.4	4.18						
JC ...	11.7 ± 0.4	5.28	12.7 ± 0.4	5.27	14.1 ± 0.5	6,792.0	9.3 ± 0.4	9,783.7	9.2 ± 0.4	9,783.3
JC model: low divergence (d_{max} = 0.1)										
CR (tree B)										
p.....	5.4 ± 0.2	1.66	5.7 ± 0.2	1.66						
JC ...	5.5 ± 0.2	1.83	6.2 ± 0.3	1.83	5.4 ± 0.4	11,309.0	4.2 ± 0.3	11,308.2	4.3 ± 0.3	11,308.2
VR (tree E)										
p.....	5.7 ± 0.3	1.65	6.5 ± 0.3	1.65						
JC ...	6.9 ± 0.3	1.81	7.6 ± 0.3	1.81	5.3 ± 0.3	11,204.0	5.2 ± 0.3	11,203.7	5.1 ± 0.3	11,203.7

NOTE.—Results are from 100 and 50 replications for ME and ML, respectively. NJ = neighbor-joining; TBR = tree bisection-reconnection; SA = stepwise addition; NNI = nearest-neighbor interchange; SPR = subtree pruning and regrafting; CR = constant-rate model; VR = varying-rate model; p = p distance; JC = Jukes-Cantor distance/model.

smaller d_T than NJ+TBR in all cases, irrespective of the distance measure. One of the surprising results is that the distance measure (HKY+Γ) based on the same model as the one used for generating sequence data shows a much poorer performance than p or JC distance. This again indicates that a simple distance measure often shows a better performance than complex distance measures in inferring the true tree.

For ML methods, we present the results for only two search algorithms, SA and SA+NNI, because an enormous amount of computer time was required for SA+SPR, and the simulation results in table 2 show that SA+SPR gives essentially the same results as those obtained by SA+NNI. (Actually, we constructed ML trees for SA, SA+NNI, and SA+SPR for the first 10 data sets for model tree F with the first parameter set of the

Table 3
Average Topological Distances from the True Tree (d_T) (±SE), Sums of Branch Lengths (S), and Negative Log Likelihood Values ($-\ln L$) of Inferred Trees Under the Minimum-Evaluation (ME) and Maximum-Likelihood (ML) Criteria with Different Nucleotide Substitution Models

	ME				ML			
	NJ		NJ+TBR		SA		SA+NNI	
	d_T	S	d_T	S	d_T	$-\ln L$	d_T	$-\ln L$
HKY+Γ model: k = 4, g_A = 0.15, g_C = 0.35, g_G = 0.35, g_T = 0.15, a = 1, d_{max} = 1.0								
CR (tree C)								
p	12.9 ± 0.6	5.36	13.2 ± 0.5	5.35				
JC	16.0 ± 0.6	7.21	16.3 ± 0.6	7.18	19.9 ± 0.7	8,868.8	16.6 ± 0.7	8,857.1
HKY+Γ	27.4 ± 0.7	12.63	32.6 ± 0.9	12.21	20.8 ± 0.7	8,590.6	16.6 ± 0.7	8,581.1
VR (tree F)								
p	9.8 ± 0.6	5.34	9.9 ± 0.4	5.34				
JC	10.6 ± 0.6	6.99	11.5 ± 0.5	6.98	14.3 ± 0.5	7,331.8	10.7 ± 0.6	7,322.2
HKY+Γ	18.4 ± 0.7	12.53	20.0 ± 0.7	12.21	15.7 ± 0.7	7,097.1	11.6 ± 0.5	7,088.8
HKY+Γ model: k = 10, g_A = 0.10, g_C = 0.40, g_G = 0.40, g_T = 0.10, a = 0.5, d_{max} = 1.0								
CR (tree C)								
p	15.2 ± 0.5	5.25	15.6 ± 0.5	5.24				
JC	16.3 ± 0.5	6.87	16.9 ± 0.5	6.85	21.2 ± 0.7	8,677.6	17.6 ± 0.7	8,662.6
HKY+Γ	28.9 ± 0.8	15.42	36.3 ± 1.1	14.87	23.8 ± 0.9	7,534.4	19.1 ± 0.9	7,526.1
VR (tree F)								
p	16.7 ± 0.6	5.04	17.3 ± 0.5	5.03				
JC	18.0 ± 0.6	6.45	17.8 ± 0.6	6.42	21.5 ± 0.7	8,533.9	18.5 ± 0.7	8,517.8
HKY+Γ	28.7 ± 0.8	18.46	30.8 ± 0.8	18.00	24.1 ± 0.7	7,409.9	18.7 ± 0.9	7,401.3

NOTE.—Results are from 100 and 50 replications for ME and ML, respectively. NJ = neighbor-joining; TBR = tree bisection-reconnection; SA = stepwise addition; NNI = nearest-neighbor interchange; HKY = Hasegawa-Kishino-Yano; CR = constant-rate model; VR = varying-rate model; p = p distance; JC = Jukes-Cantor distance. p, JC, and HKY+Γ distances and JC and HKY+Γ models were used for ME and ML, respectively.

HKY+Γ model, but the latter two algorithms gave essentially the same d_T value.) Here, however, we used the JC model as well as the HKY+Γ model to construct ML trees. The results show that the SA+NNI search gives a significantly smaller d_T than does SA in all cases. Therefore, at least the SA+NNI search should be performed for inferring the true tree.

One of the most paradoxical results in this study is that the use of the correct model, HKY+Γ, does not necessarily give a better topology (a smaller d_T value) than the use of an incorrect model, JC. Actually, the JC model generally gives a better performance than the HKY+Γ model, particularly when the SA algorithm is used. Interestingly, however, the average ln*L* value is much higher for HKY than for JC, with the difference being more than 1,000 for the second set of HKY+Γ parameters. Obviously, if we conduct the likelihood ratio test for individual data sets, the difference will become highly significant. In other words, if we choose a substitution model by applying the likelihood ratio test to this type of data set (Swofford et al. 1996), we will certainly choose the HKY+Γ model, yet this model does not give a better performance for topology reconstruction.

Discussion

Fast Heuristic Algorithms

The primary purpose of this paper was to examine the efficiencies of several fast heuristic algorithms of phylogenetic inference under the MP, ME, and ML criteria when a large number of sequences are used. For the ME criterion, we reached the conclusion that the simple NJ algorithm is almost always better than other sophisticated algorithms in inferring the true tree. This is consistent with the results obtained by Kumar (1996), Gascuel (1997), and Nei, Kumar, and Takahashi (1998).

Currently, the most popular methods for obtaining MP and ML trees use extensive search algorithms such as SA+TBR or 100SA+TBR. This is based on the general belief that the MP or ML tree is likely to be the true tree or close to it. However, when the number of sequences (*m*) is large and the number of nucleotides examined (*n*) is relatively small, this is incorrect, and the MP and ML trees tend to give incorrect topologies, as was shown by Nei, Kumar, and Takahashi (1998). In this study, we have shown that relatively simple search algorithms such as SA+NNI or SA+SPR are sufficient for inferring the true tree.

Of course, when *m* is large and *n* is relatively small, it would be very difficult to find the true tree by any method. In this case, the important thing is to identify the major branching patterns of the tree and leave the other branching patterns unresolved. This can be done easily if we use the bootstrap test (Felsenstein 1985). For this purpose, a simple search algorithm has an added bonus, because it does not take much computer time.

Nevertheless, it should be noted that the results obtained by this type of simulation study depend on the model trees used. Although we used several model trees that were generated at random and our conclusion was largely independent of the model tree used, these model trees still represent only a small proportion of many possible trees. It is therefore desirable to examine more different topologies and to find general properties of different algorithms in the future. Note also that the pattern of nucleotide substitution in real data is much more complicated than the models used in this paper.

Models of Nucleotide Substitution

We have shown that for the ME or NJ method, the simple p distance is more efficient than the JC or HKY+Γ distance in inferring the true tree. The superiority of the p distance over the JC distance in phylogenetic reconstruction has been noted by Saitou and Nei (1987), Sourdis and Krimbas (1987), and Tajima and Takezaki (1994), but our results indicate that it is also superior to the HKY+Γ distance even when the sequence data are generated by the HKY+Γ model. This superiority is apparently caused by the small variance of this distance compared with the variances of other distance measures. In recent years, a number of investigators have used ML estimates of the HKY or HKY+Γ distance to construct an NJ or ME tree (e.g., Berntson, France, and Mullineaux 1999; Silberman et al. 1999; Zardoya, Economidis, and Doadrio 1999). Our study suggests that this practice should be discouraged. It appears that the HKY or HKY+Γ distance is useful only when the evolutionary rate varies extensively with evolutionary lineage and the number of nucleotides examined is very large (>10,000) (Takezaki and Gojobori 1999).

Similarly, we have shown that in phylogenetic inference by ML methods, the HKY+Γ model is less efficient than the JC model even if the HKY+Γ model is used for generating sequence data. Somewhat similar results were obtained by Yang (1997) and Bruno and Halpern (1999) when they compared the efficiencies of topology reconstruction by the JC and the JC+Γ models for various model trees of four to eight sequences. Complex relationships between substitution models and the efficiency of topology construction have also been noted by Tateno, Takezaki, and Nei (1994) and Gaut and Lewis (1995). It appears that these results are partly caused by a larger number of parameters to be estimated in complex models than in simple models and are partly due to the complex nature of phylogenetic inference by ML methods as discussed by Nei (1987, 1996), Yang, Goldman, and Friday (1995), and Yang (1997). Note that the mathematical basis of phylogenetic analysis by ML methods is quite different from the standard parameter estimation in a single probability space (Yang, Goldman, and Friday 1995; Nei 1996). The pattern of nucleotide substitution in real sequence data is also much more complicated than the HKY+Γ model, and for this reason some authors (e.g., McArthur and Koop 1999; Silberman et al. 1999; Waits et al. 1999) have used even more complicated models. It has also been suggested that selection of a model to be used for constructing ML trees should be determined by conducting the likelihood ratio test (Swofford et al. 1996; Huelsen-

beck and Crandall 1997) or by using the Akaike information criterion (AIC) (Hasegawa and Kishino 1989). The present study raises serious questions about these procedures. A more careful study of the theoretical foundation of phylogenetic inference by ML methods is necessary.

In this paper, we were concerned primarily with topology construction, and our conclusion does not apply to the estimation of branch lengths for a given topology. For estimating branch lengths, it would be better to use a model close to the real substitution pattern.

Efficiencies of Different Tree-Building Methods

Although it was not our primary purpose to compare the efficiencies of different tree-building methods in inferring the true tree, our study provides some information on this problem. First, our study indicates that the efficiencies of the NJ and the ML methods depend on the mathematical model (or distance measure) used and that the correct model does not necessarily give the highest efficiency. Therefore, comparison of different tree-building methods should be done using the model that is most appropriate for each method. Our results showed that in the case of high divergence, the NJ algorithm with p distance generally gives a smaller d_T than the MP+NNI and the ML+NNI algorithms, particularly when the HKY+Γ model is used for generating nucleotide sequences. In this case, even NJ with JC distance shows a better performance than MP and ML methods.

When the extent of sequence divergence is low and $d_{max} = 0.1$, however, NJ with p distance tends to give a slightly higher d_T than the MP+NNI and ML+NNI methods. This is understandable, because a phylogenetic tree for closely related sequences with a relatively small n is expected to have many multifurcating nodes, and for identifying multifurcating nodes, MP and ML methods are superior to NJ, since the latter is intended to construct a bifurcating tree (Nei and Kumar 2000). In practice, of course, an inferred tree is usually subjected to the bootstrap test, and these multifurcating nodes are judged as unresolved in the inference of bifurcating trees. Therefore, it is difficult to compare the efficiencies of different tree-building methods in terms of d_T alone when d_{max} is low and n is small.

As mentioned above, the relative efficiencies of different tree-building methods depend on the model tree, the substitution model, the number of nucleotides, etc., and it is very difficult to make any general conclusions from a limited number of computer simulations. In the past, it was customary to compare different tree-building methods by using the same substitution model for all of the methods and a small number of sequences. Our study shows that this is not the right method of comparison if we want to have conclusions that can be used for practical application. In any theoretical study of phylogenetics, it is important to find the best method of analysis for each tree-building method and then compare different methods. Also, the mathematical models currently used in phylogenetic analysis are crude approximations to reality. Therefore, we should not make unwarranted extrapolations of simulation results to real sequence data. Actually, many recent empirical studies with real data sets (Andersson et al. 1999; Honda, Yokota, and Sugiyama 1999; Marko and Vermeij 1999; Meireles et al. 1999; Tan et al. 1999) suggest that MP, NJ, and ML methods generally give essentially the same phylogenetic inference when the bootstrap test is applied. Phylogenetic inference is a biological problem rather than a mathematical one, and it is important to use biological information in the final judgment of the reliability of inferred trees.

Acknowledgments

We thank James Lyons-Weiler and Mike Miyamoto for their comments on an earlier version of this paper. This study was supported by grants from NIH (GM20293) and NASA (NCC2-1057).

LITERATURE CITED

ANDERSSON, S. G. E., D. R. STOTHARD, P. FUERST, and C. G. KURLAND. 1999. Molecular phylogeny and rearrangement of rRNA genes in *Rickettsia* species. Mol. Biol. Evol. **16**: 987–995.

BERNTSON, E. A., S. C. FRANCE, and L. S. MULLINEAUX. 1999. Phylogenetic relationships within the class Anthozoa (phylum Cnidaria) based on nuclear 18S rDNA sequences. Mol. Phylogenet. Evol. **13**:417–433.

BRUNO, W. J., and A. L. HALPERN. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol. Biol. Evol. **16**:564–566.

CHASE, M. W., D. E. SOLTIS, R. G. OLMSTEAD et al. (42 coauthors). 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *Rbc*L. Ann. Mo. Bot. Gard. **80**:528–580.

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

EDWARDS, A. W. F., and L. L. CAVALLI-SFORZA. 1963. The reconstruction of evolution. Heredity **18**:553.

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

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

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

GASCUEL, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**:685–695.

GAUT, B. S., and P. O. LEWIS. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. **12**:152–162.

HASEGAWA, M., and H. KISHINO. 1989. Confidence limits on the maximum likelihood estimation of the hominoid tree for mitochondrial DNA sequences. Evolution **43**:672–677.

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

HONDA, D., A. YOKOTA, and J. SUGIYAMA. 1999. Detection of seven major evolutionary lineages in Cyanobacteria based on 16S rRNA gene sequence analysis with new sequences

of five marine *Synechococcus* strains. J. Mol. Evol. **48**:723–739.

HUELSENBECK, J. P., and K. A. CRANDALL. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. **28**:437–466.

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

KUHNER, M. K., and J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11**:459–468.

KUMAR, S. 1996. A stepwise algorithm for finding minimum evolution trees. Mol. Biol. Evol. **13**:584–593.

KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetics analysis. Pennsylvania State University, University Park.

MCARTHUR, A. G., and B. F. KOOP. 1999. Partial 28S rDNA sequences and the antiquity of hydrothermal vent endemic gastropods. Mol. Phylogenet. Evol. **13**:255–274.

MADDISON, D. R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. Syst. Biol. **40**:315–328.

MARKO, P. B., and G. J. VERMEIJ. 1999. Molecular phylogenetics and the evolution of labral spines among Easter Pacific ocenebrine gastropods. Mol. Phylogenet. Evol. **13**: 275–288.

MEIRELES, C. M., J. CZELUSNIAK, M. P. C. SCHNEIDER, J. A. P. C. MUNIZ, M. C. BRIGIDO, H. S. FERREIRA, and M. GOODMAN. 1999. Molecular phylogeny of Ateline New World monkeys (*Platyrryhini, Atelinae*) based on γ-globin gene sequences: evidence that *Brachyteles* is the sister group of *Lagothrix.* Mol. Phylogenet. Evol. **12**:10–30.

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

———. 1991. Relative efficiencies of different tree making methods for molecular data. Pp. 133–147 *in* M. M. MIYAMOTO and J. L. CRACRAFT, eds. Recent advances in phylogenetic studies of DNA sequences. Oxford University Press, Oxford, England.

———. 1996. Phylogenetic analysis in molecular evolutionary genetics. Annu. Rev. Genet. **30**:371–403.

NEI, M., and S. KUMAR. 2000. Molecular evolution and phylogenetics. Oxford University Press, Oxford.

NEI, M., S. KUMAR, and K. TAKAHASHI. 1998. The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proc. Natl. Acad. Sci. USA **95**:12390–12397.

PENNY, D., and M. D. HENDY. 1985. The use of tree comparison metrics. Syst. Zool. **34**:75–82.

RICE, K. A., M. J. DONOGHUE, and R. G. OLMSTEAD. 1997. Analyzing large data sets: *Rbc*L revisited. Syst. Biol. **46**: 554–563.

ROBINSON, D. F., and L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–147.

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

SAITOU, N., and M. IMANISHI. 1989. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic reconstructions in obtaining the 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.

SILBERMAN, J. D., C. G. CLARK, L. S. DIAMOND, and M. L. SOGIN. 1999. Phylogeny of the genera *Entamoeba* and *Endolimax* as deduced from small-subunit ribosomal RNA sequences. Mol. Biol. Evol. **16**:1740–1751.

SOURDIS, J., and C. KRIMBAS. 1987. Accuracy of phylogenetic trees estimated from DNA sequence data. Mol. Biol. Evol. **4**:159–166.

SWOFFORD, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (and other methods). Sinauer, Sunderland, Mass.

SWOFFORD, D. L., and D. P. BEGLE. 1993. PAUP: phylogenetic analysis using parsimony. Version 3.1. User's manual. Laboratory of Molecular Systematics, Smithsonian Institution, Washington, D.C.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. 2nd edition. Sinauer, Sunderland, Mass.

TAJIMA, F., and N. TAKEZAKI. 1994. Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. Mol. Biol. Evol. **11**:278–286.

TAKEZAKI, N., and T. GOJOBORI. 1999. Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences. Mol. Biol. Evol. **16**:590–601.

TAN, I. H., J. BLOMSTER, G. HANSEN, E. LESKINEN, C. A. MAGGS, D. G. MANN, H. J. SLUIMAN, and M. J. STANHOPE. 1999. Molecular phylogenetic evidence for a reversible morphogenetic switch controlling the gross morphology of two common genera of green seaweeds, *Ulva* and *Enteromorpha.* Mol. Biol. Evol. **16**:1011–1018.

TATENO, Y., M. NEI, and F. TAJIMA. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J. Mol. Evol. **18**:387–404.

TATENO, Y., N. TAKEZAKI, and M. NEI. 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.

WAINRIGHT, P. O., G. HINKLE, M. L. SOGIN, and S. K. STICKEL. 1993. Monophyletic origins of the metazoa: an evolutionary link with fungi. Science **260**:340–342.

WAITS, L. P., J. SULLIVAN, S. J. O'BRIEN, and R. H. WARD. 1999. Rapid radiation events in the family Ursidae indicated by likelihood phylogenetic estimation from multiple fragments of mtDNA. Mol. Phylogenet. Evol. **13**:82–92.

YANG, Z. 1997. Phylogenetic analysis by maximum likelihood (PAML). University of California, Berkeley.

YANG, Z., N. GOLDMAN, and A. E. FRIDAY. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. **44**:384–399.

ZARDOYA, R., P. S. ECONOMIDIS, and I. DOADRIO. 1999. Phylogenetic relationships of Greek Cyprinidae: molecular evidence for at least two origins of the Greek cyprinid fauna. Mol. Phylogenet. Evol. **13**:122–131.

NARUYA SAITOU, reviewing editor

Accepted April 18, 2000