J Mol Evol (1996) 42:294–307

JOURNAL OF MOLECULAR EVOLUTION

Phylogenetic Analysis Using Parsimony and Likelihood Methods

Ziheng Yang

College of Animal Science and Technology, Beijing Agricultural University, Beijing 100094, China, Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, 328 Mueller Laboratory, University Park, PA 16802, USA

Received: 10 March 1995 / Accepted: 15 September 1995

Abstract. The assumptions underlying the maximum-parsimony (MP) method of phylogenetic tree reconstruction were intuitively examined by studying the way the method works. Computer simulations were performed to corroborate the intuitive examination. Parsimony appears to involve very stringent assumptions concerning the process of sequence evolution, such as constancy of substitution rates between nucleotides, constancy of rates across nucleotide sites, and equal branch lengths in the tree. For practical data analysis, the requirement of equal branch lengths means similar substitution rates among lineages (the existence of an approximate molecular clock), relatively long interior branches, and also few species in the data. However, a small amount of evolution is neither a necessary nor a sufficient requirement of the method. The difficulties involved in the application of current statistical estimation theory to tree reconstruction were discussed, and it was suggested that the approach proposed by Felsenstein (1981, *J. Mol. Evol.* 17: 368–376) for topology estimation, as well as its many variations and extensions, differs fundamentally from the maximum likelihood estimation of a conventional statistical parameter. Evidence was presented showing that the Felsenstein approach does not share the asymptotic efficiency of the maximum likelihood estimator of a statistical parameter. Computer simulations were performed to study the probability that MP recovers the true tree under a hierarchy of models of nucleotide substitution; its performance relative to the likelihood method was especially noted. The results appeared to support the intuitive examination of the assumptions underlying MP. When a simple model of nucleotide substitution was assumed to generate data, the probability that MP recovers the true topology could be as high as, or even higher than, that for the likelihood method. When the assumed model became more complex and realistic, e.g., when substitution rates were allowed to differ between nucleotides or across sites, the probability that MP recovers the true topology, and especially its performance relative to that of the likelihood method, generally deteriorates. As the complexity of the process of nucleotide substitution in real sequences is well recognized, the likelihood method appears preferable to parsimony. However, the development of a statistical methodology for the efficient estimation of the tree topology remains a difficult open problem.

Key words: Maximum likelihood — Maximum parsimony — Molecular evolution — Molecular systematics — Phylogeny — Computer simulation

Introduction

The maximum parsimony (MP) method of phylogenetic tree reconstruction using nucleotide sequence data performs a site-by-site analysis. For each tree topology, it calculates the minimum number of nucleotide changes (substitutions) that are required to explain the observed site pattern. The numbers of changes are summed over sites to give a parsimony score for each tree topology, and the topology having the smallest total number of changes is taken as the estimate of the phylogeny, which

Present address: Department of Integrative Biology, University of California, Berkeley CA 94720-3140, USA

is known as the most parsimonious tree. The method was proposed by Edwards and Cavalli-Sforza (1963; see also Cavalli-Sforza and Edwards 1967) to analyze gene-frequency data and by Camin and Sokal (1965) for morphological characters. The application of the method to molecular sequence data goes back to Pauling and Zuckerkandl (1963), Zuckerkandl (1964), and Eck and Dayhoff (1966), and Fitch (1971) and Hartigan (1973) devised an algorithm for its systematic implementation.

There has been much controversy concerning the mechanism of the method (see, e.g., Felsenstein 1973, 1978; Farris, 1973, 1977; Felsenstein and Sober 1986; Sober 1988). The method does not make explicit assumptions concerning the process of sequence evolution. Indeed, there were suggestions that the method involved no assumptions at all (e.g., Wiley 1975:236), and there were further suggestions that reconstruction of phylogeny should ideally be free from any assumptions concerning the evolutionary process. Felsenstein (1978) used a simple model for the evolution of a two-state character along a four-species tree with highly variable rates of change among lineages and showed that parsimony can be statistically inconsistent; as the amount of data (the number of nucleotides in each sequence) increases without bound, the method will produce a wrong tree with the probability approaching one. According to Sober's (1988:166–172) interpretation, both Felsenstein (1973, 1978) and Farris (1973, 1977) agreed that parsimony does involve assumptions, but they disagreed on whether Felsenstein's overly simplified model is relevant to the performance of parsimony in real data analysis. Felsenstein (1973, 1978, 1988) suggested that when the amount of evolution is small and the rate of evolution is more or less constant among lineages, parsimony may be an acceptable approximation to likelihood (see also Cavalli-Sforza and Edwards 1967; Goldman 1990). More recent studies have shown that even the existence of a molecular clock combined with a small amount of evolution does not guarantee the statistical consistency of the parsimony method, and the situation seems to become worse when more sequences are in the data (Hendy and Penny 1989; Zharkikh and Li 1993; Takezaki and Nei 1994). Computer simulations that compared different tree reconstruction methods have tended to suggest that the performance of MP in finite data samples is often poorer than distance matrix and likelihood methods (Hasegawa and Yano 1984; Saitou and Imanishi 1989; Jin and Nei 1990; Hasegawa and Fujiwara 1993; Huelsenbeck and Hillis 1993; Kuhner and Felsenstein 1994; Tateno et al. 1994).

In this paper an attempt will be made to identify the assumptions underlying the parsimony method of tree reconstruction. I will do this initially by examining the way the method works, based on the understanding that, to justify a treatment of the data, certain assumptions have to be made concerning the process that has given rise to the data. Computer simulations will be performed to corroborate this intuitive examination, with various substitution models assumed. The model will be made more and more complex and realistic by adding components that are known to reflect characteristics of sequence evolution, and the probability that parsimony recovers the true tree will be estimated. The working hypothesis is that when the true model moves further and further away from the model underlying a method, the performance of the method may be expected to deteriorate.

The main objective of this study is to examine the differences of parsimony and likelihood methods of tree reconstruction and to identify important factors that account for their success or failure. Initially, the assumptions of the parsimony method will be intuitively examined. Then the simulation experiment will be described, and the results will be presented. The differences and difficulties of tree reconstruction in comparison with the estimation of a conventional statistical parameter will be discussed. This discussion also provides an interpretation of some ''counterintuitive'' results found in the simulation experiment.

An Intuitive Examination of Assumptions Underlying the Parsimony Method of Phylogenetic Tree Reconstruction

Consider the sequence data of the 895-bp segment of the mitochondrial genome (mtDNA) of human (H), chimpanzee (C), gorilla (G), and orangutan (O) (Brown et al. 1982). The segment codes for three tRNAs and parts of two proteins. The data can be ''summarized'' as the observed numbers of different site patterns in the form of Table 1. Sites with identical nucleotides in different species (patterns 1–4 in Table 1) and those at which only one species has a different nucleotide (e.g., patterns 6, 8, 10 in Table 1) are considered ''noninformative,'' as these two kinds of sites require no and one change, respectively, for any tree topology and so do not contribute to the discrimination of the trees. Site patterns represented by *xxyz, xyxz,* and *yzxx,* where *x, y,* and *z* are any different nucleotides (e.g., patterns 24, 31, and 33 in Table 1), are also ''noninformative'' by this criterion, as these patterns can be explained by two substitutions for any tree topology. The only ''informative'' patterns are *xxyy, xyxy,* and *xyyx,* which support the three bifurcating trees T_1 = ((H,C),G,O), T_2 = ((H,G),C,O), and T_3 = ((C,G),H,O), respectively. For four species, parsimony uses only 36 (= 3 × 12) out of 256 possible site patterns and does not use the data efficiently. It should be noted that the ''noninformative sites'' do contain phylogenetic information. Even the ''constant'' patterns provide information about the amount of evolution and the pattern of nucleotide substitution. For example, the higher frequency of *C* (0.311) than that of *G* (0.104) observed in the mtDNA data suggests that *G* probably has a higher rate of change

Table 1. The observed site-pattern frequencies (n_i) in the 895-bp mtDNA sequences of human (H), chimpanzee (C), gorilla (G), and orangutan (O) (Brown et al. 1982)

	2 21										
	2716	2 1 1	1 1	1 1							
n_i	21772	38133	65717	63264	67719	63243	11124	11112	11311	1	
Pattern *i*	AGCTG	TCATC	ACCCA	CCTTT	TGCAG	AGAAT	TATAC	GGTGC	ACTGC	C	Human
	AGCTA	TTACC	ACCCA	TCCTC	TGCCA	GGTAC	CGTAT	AAAGC	GTTAA	C	Chimpanzee
	AGCTG	TTGTT	ATCAA	CACCT	CGCAA	AAAAC	TGCCC	AGAGT	ATTAA	T	Gorilla
	AGCTA	CCACC	GTTCC	CACCT	TAATA	AAATT	AAAAT	GGGCG	CTACA	A	Orangutan
Supported tree	2	3 2	1	1 1		1 3	3 2	3			

[a] The 46 observed patterns are arranged columnwise in the order of occurrence in the data, for example, the first two patterns, AAAA and GGGG, are observed at 222 and 71 sites, respectively. The three tree topologies supported by the ''informative'' patterns are T_1 = ((H,C),G,O), T_2 = ((H,G),C,O), T_3 = ((C,G),H,O); other site patterns are ''noninformative'' by the parsimony analysis. So T_1, T_2, and T_3 are supported by 17 (= 5 + 3 + 6 + 3), 9(= 2 + 3 + 4), and 13(= 8 + 3 + 1 + 1) sites, respectively, and T_1 is the most parsimonious tree

Table 2. Log-likelihood values and parameter estimates obtained by fitting the HKY85 + C + dG model[a]

Tree		$\hat{t}_0$	$\hat{t}_H$	$\hat{t}_C$	$\hat{t}_G$	$\hat{t}_O$	$\hat{c}_2$	$\hat{c}_3$	$\hat{c}_4$	$\hat{\kappa}$	$\hat{\alpha}$
(HCGO)	−2,100.42		0.038	0.043	0.066	0.238	0.275	5.963	0.631	26.957	0.611
((HC)GO)	−2,097.26	0.016	0.038	0.043	0.051	0.219	0.277	5.347	0.587	25.337	0.764
((HG)CO)	−2,100.42	0.000	0.038	0.043	0.066	0.238	0.275	5.962	0.631	26.951	0.612
((HO)CG)	−2,099.86	0.009	0.029	0.041	0.064	0.222	0.279	5.780	0.632	25.063	0.691

[a] The 895-bp mtDNA sequences of human (H), chimpanzee (C), gorilla (G), and orangutan (O) (Brown et al. 1982) were analyzed by assuming the HKY85 model of nucleotide substitution (Hasegawa et al. 1985). Different rate parameters ($c_1 = 1$, c_2, c_3, c_4) were assumed for the three codon positions and for sites from the tRNA-coding region, and a discrete-gamma model (dG, with eight rate categories) was assumed to account for the remaining rate variation. Branch lengths in each of the tree topologies were also estimated by maximum likelihood, where t_H, t_C, t_G, and t_O are the lengths of branches leading to human, chimpanzee, gorilla, and orangutan, respectively, while t_0 is the interior branch length. The base frequency parameters are estimated using the averages of observed frequencies; they are 0.2542, 0.3313, 0.3106, and 0.1039 for *T, C, A,* and *G,* respectively. The maximum-likelihood tree topology under the model is ((HC)GO). This table is used to explain the protocol of phylogenetic tree estimation by maximum likelihood and the assumptions to be made to justify the practice of summarizing the sequence data in the form of Table 1

than does *C*, so that in the long run more *C*s are observed than *G*s. Such information could be employed in the estimation of the phylogeny.

In order to justify the practice of summarizing the sequence data in the form of Table 1, two assumptions concerning the substitution process are necessary: (1) The stochastic processes of nucleotide substitution are identical at different sites; (2) Substitutions occur independently among sites. These two assumptions ensure that data at different sites are identically and independently distributed (i.i.d.) in which case Table 1 is a legitimate summary of the original data. In fact, both assumptions are unrealistic for the mtDNA data. For example, Table 2 presents results obtained from fitting a model that uses independent rate parameters for the three codon positions and for sites in the tRNA region. Estimates of these rate parameters are in the proportion 1: 0.28:5.35:0.59. The model is still unrealistic as it assumes the same transition/transversion rate ratio and gamma parameter for different codon positions. Nevertheless, it should be noted that under this model, sites at different codon positions, even if their nucleotide compositions are identical, are treated differently. Without the assumption of among-site homogeneity, it is illegitimate to combine these sites into one category (as in Table 1).

Similar arguments may be used to examine other aspects of the parsimony analysis. For instance, the method treats all nucleotide differences in the same way—patterns like *TTCC, TTAA, TTGG* are further combined into one category, *xxyy*—which suggests that substitutions between different nucleotides, such as transitions and transversions, are assumed to occur with equal probability. This is similar to the model of Jukes and Cantor (1969), which assumes the same substitution rate between any two nucleotides. The major difference is that parsimony fails to account for different branch lengths, and short and long branches are treated in the same way when the numbers of changes are inferred and counted. This lack of a time structure in the method (Thompson 1975; Goldman 1990) is justifiable only if all branches are equally long. For practical data analysis, the requirement of equal branch lengths means similar substitution rates among lineages, relatively long interior branches, and, most often, few species in the data. Despite previous suggestions (e.g., Camin and Sokal 1965; Felsenstein 1973, 1978), a small amount of evolution does not appear to be a requirement for the method. Counting the

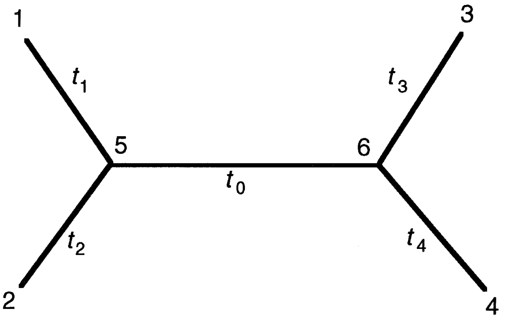

Fig. 1. A model tree of four species. The branch lengths, $t = (t_0, t_1, t_2, t_3, t_4)$, are measured by the expected numbers of nucleotide substitutions per site.

numbers of changes and calculation of the differences of such counts among trees, although not accurate in the presence of long branches, are justified even with a large amount of evolution.

Accounting for the rate heterogeneity across sites does not necessarily violate the i.i.d. assumption of data among sites (Yang 1993), and it is not intuitively clear whether or not parsimony assumes rate constancy among sites. The problem concerns the weighting of sites during the calculation of the score for tree selection. In a likelihood analysis, highly variable sites usually have low probabilities of occurrence and contribute more to the likelihood score (Yang et al. 1994). Parsimony assumes weight zero for ''noninformative'' sites and weight one for all remaining sites. (That ''equal weighting'' is not the same as ''no weighing'' was nicely discussed by Swofford and Olsen 1990.) The equal weighting adopted by parsimony is insufficient to handle information contained in the data even if rates are constant at sites, and appears worse when rates are variable (Yang et al. 1994). So parsimony may be expected to perform worse when rate variation among sites exists than when rates are constant.

Trees and Models Assumed in the Computer Simulation of This Study

The Shape of the Model Tree as Reflected in the Lengths of Its External Branches

Only four-species trees are considered in this study, and the model tree topology is shown in Fig. 1. Ten sets of branch lengths (Table 3) are used to represent different shapes of trees as reflected in the external branch lengths: $(t_1, t_2) - (t_3, t_4)$. The probability of recovering the true tree by any method is expected to rise monotonically with the interior branch length (t_0), and so t_0 is not used in the representation of the three shapes. The ''B'' trees have branches that are 0.2 times as long as those of the ''A'' trees. Trees 1A and 1B have equal external branch lengths. A molecular clock holds in tree 2A and 2B: $t_1 = t_2, t_3 = t_1 + t_2, t_4 > t_3$. However, it is the shape of the tree rather than the clock assumption that is important. Trees 3A and 3B have two long branches and two short branches separated, which is the shape Felsenstein (1978) used to show the inconsistency of the parsimony method and which appears to be the most commonly used tree shape in previous simulation studies. Trees 4A and 4B have three short branches and one long branch, while trees 5A and 5B have three long branches and one short branch.

The Model of Nucleotide Substitution Assumed for Generating the Data

The substitution model of Hasegawa et al. (1985) and its special forms were assumed. According to this model, the rate of substitution from nucleotide i to j ($i \neq j$) is

$$Q_{ij} = \begin{cases} \kappa\pi_j, & \text{for transitions: } T \leftrightarrow C, A \leftrightarrow G \\ \pi_j, & \text{for transversions: } T,C \leftrightarrow A,G \end{cases} \tag{1}$$

where κ is the transition/transversion rate ratio and π_j is the equilibrium frequency of nucleotide j. The model is designated ''HKY85.'' Simpler models that were used include that of Kimura (1980), represented ''K80,'' which is equivalent to HKY85 by setting $\pi_T = \pi_C = \pi_A = \pi_G = 1/4$, and that of Jukes and Cantor (1969), referred to as ''JC69,'' which is a special case of K80 with $\kappa = 1$. Three values were used for κ in the K80 or HKY85 models: 1, 5, and 20. The frequency parameters in the HKY85 model were fixed at $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$, and $\pi_G = 0.4$.

The gamma distribution was used to describe variable substitution rates among sites. The shape parameter α of the distribution is inversely related to the extent of rate variation (Yang 1993). The ''discrete-gamma'' (dG) model of Yang (1994c) was used in this paper both for generating and for analyzing the data, with four equal-probability categories used to approximate the continuous gamma. Three values of α were used: ∞ (no rate variation), 0.8 (slight rate variation), and 0.2 (severe rate variation).

Generation of the Simulated Data

Two sequence lengths (N) were used: 250 and 500. For each combination of parameter values and sequence length, r simulated samples were generated with r determined by calculating the standard error of the estimated probability (P) of recovering the correct tree: $\sigma_P = [P(1 - P)/r]^{1/2}$. For every 200 replicates, σ_P was calculated using the current estimate of P. For the parsimony method, r was determined such that $\sigma_P \leqslant 0.005$ with the restriction that $r \geqslant 500$, while for the likelihood method, $\sigma_P \leqslant 0.01$ was used with the restriction $200 \leqslant r \leqslant 1{,}000$.

The probabilities of observing all site patterns were calculated by the approach of Felsenstein (1981) for models assuming a constant rate for all sites ($\alpha = \infty$) or

Table 3. Branch lengths and tree shapes examined in the computer simulation

Tree	t_0	t_1	t_2	t_3	t_4	Tree shape
1A	0.1	0.5	0.5	0.5	0.5	(equal external branches)
1B	0.02	0.1	0.1	0.1	0.1	
2A	0.1	0.5	0.5	0.6	1.4	(short, short)–(long, long)
2B	0.02	0.1	0.1	0.12	0.28	
3A	0.1	0.1	0.5	0.2	1	(short, long)–(short, long0
3B	0.02	0.02	0.1	0.04	0.2	
4A	0.05	0.05	0.05	0.05	0.5	(short, short)–(short, long)
4B	0.01	0.01	0.01	0.01	0.1	
5A	0.05	0.5	0.5	0.5	0.05	(long, long)–(long, short)
5B	0.01	0.1	0.1	0.1	0.01	

by the approach of Yang (1994c) for the discrete-gamma models. These probabilities are then used to generate the observed numbers of site patterns, which constitute the data to be analyzed by each tree reconstruction method. The three methods examined in this study are maximum parsimony, maximum likelihood assuming the JC69 substitution model, and maximum likelihood assuming the true model used to generate the data. These are referred to as the MP, JC, and TRUE methods. For the TRUE method, the κ parameter in the K80 and HKY85 models and the α parameter in the (discrete-) gamma model were estimated by iteration from the simulated data for each tree, while the frequency parameters in the HKY85 model (the πs) were estimated by using the averages of the observed nucleotide frequencies.

Results of the Computer Simulation Experiment

Factors Affecting the Probability of Recovering the True Tree

Before presenting results of the simulation experiment, we briefly discuss effects of some important factors that affect the performance of any tree reconstruction method. Tree 1A of Table 3 and the parsimony method are used as an example, and the results are shown in Figs. 2–4.

Let p_1, p_2, p_3 be the probabilities of observing the three site patterns *xxyy, xyxy,* and *xyyx,* respectively. ($1 - p_1 - p_2 - p_3$ will be the probability of observing all other patterns.) With tree 1A, we have $p_1 > p_2 = p_3$, for which case Zharkikh and Li (1992:1129) derived an approximate formula for calculating the sample size required by parsimony to recover the true tree with a prespecified probability (P)

$$N_P = \left[\frac{\sqrt{p_2/\pi} + z_P\sqrt{p_1 + [1-(1/\pi)]p_2}}{\Delta p}\right]^2 + \frac{1}{\Delta p} \quad (2)$$

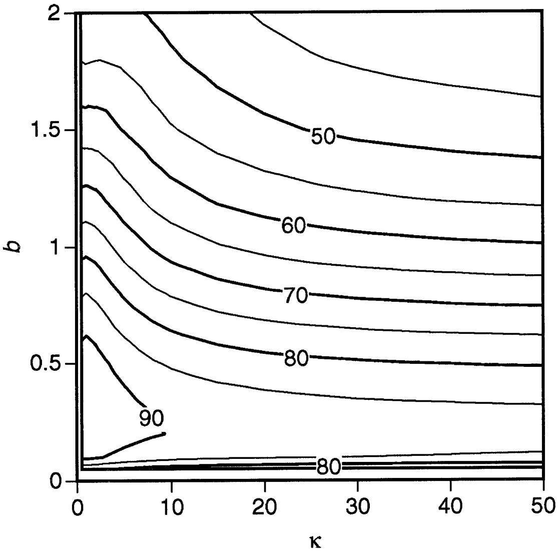

Fig. 2. The contour graph of the probability (P, ×100) that MP recovers the true tree as a function of the amount of evolution represented by b and the transistion/transversion rate ratio κ. The K80 model was assumed to generate data, while branch lengths are based on tree 1A of Table 1: $\mathbf{t}$ = (0.1b, 0.5b, 0.5b, 0.5b, 0.5b) (see Fig. 1). The sequence length is N = 500.

where $\Delta p = p_1 - p_2$, and z_P is the one-tail standard normal variate corresponding to probability P, e.g., $z_{0.95} = 1.65$. The probability of recovering the true tree given the sample size N is

$$P(N) = \Phi\left(\frac{\Delta p\sqrt{N - 1/\Delta p} - \sqrt{p_2/\pi}}{\sqrt{p_1 + (1 - 1/\pi)p_2}}\right) \quad (3)$$

where Φ(●) is the cumulative density function of the standard normal distribution, and can be calculated by using the method of Hill (1973). Eq. 3 was found to give results very similar to those obtained from simulations (results not shown), in accordance with the high accuracy of Eq. 2 reported by Zharkikh and Li (1992). Results for the contour graphs of Figs. 2–4 were calculated using Eq. 3.

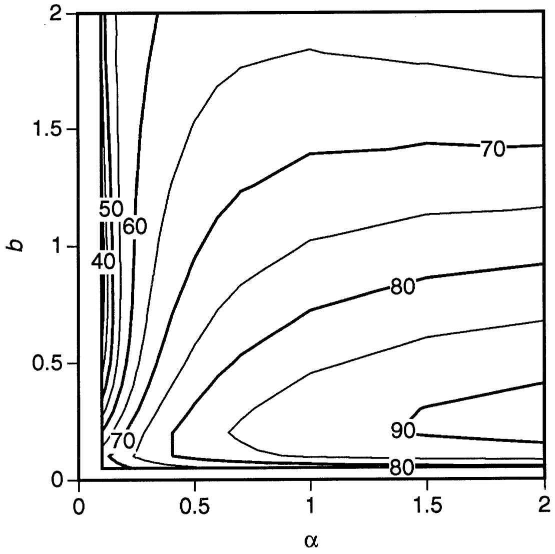

Fig. 3. The contour graph of the probability (P, ×100) that MP recovers the true tree as a function of the amount of evolution measured by b and the α parameter of the gamma distribution for variable rates at sites. The true model if JC69 + dG and branch lengths are $\mathbf{t} = (0.1b, 0.5b, 0.5b, 0.5b, 0.5b)$ (see Fig. 1). The sequence length is $N = 500$.

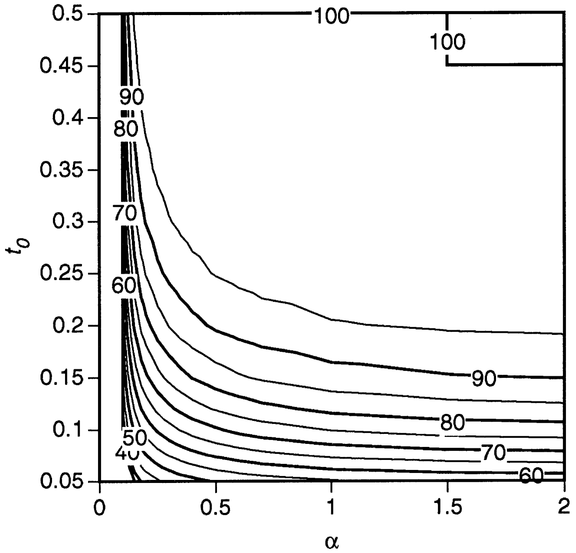

Fig. 4. The contour graph of the probability (P, ×100) that MP recovers the true tree as a function of the interior branch length (t_0) and the α parameter of the gamma distribution for variable rates among sites. The JC69 + dG model was assumed to generate data, with branch lengths $\mathbf{t} = (t_0, 0.5, 0.5, 0.5, 0.5)$ (see Fig. 1). The sequence length is $N = 500$.

The variable b in Figs. 2 and 3 measures the amount of evolution, and branch lengths are proportional to it. Clearly the best performance is achieved at intermediate values of b as both very similar and very different sequences contain little phylogenetic information. The probability of recovering the true tree improves dramatically with b when b is small and deteriorates slowly with b when b is larger than the optimum value. The effects of b and α are somewhat confounded, and the optimum values of b are smaller when substitution rates are highly variable, i.e., when α is very small (Fig. 3). The interior branch length (t_0) is of great importance to the success of tree reconstruction. The probability P steadily increases with t_0 (Fig. 4). Generally speaking, P increases with α (Figs. 3 and 4), which confirm the previous observation that sequences with no or little rate variation across sites provide more information concerning the phylogeny than sequences involving severe rate variation (Yang 1995b). The effect of the transition/transversion rate bias (κ) is minor in comparison with that of α and branch lengths (Fig. 2). Overall, extreme bias (i.e., large κ) leads to a decrease in the probability of recovering the true tree.

The observations that an intermediate amount of evolution and long interior branches lead to high probability of recovering the true tree can be expected to apply irrespective of the tree reconstruction method, the shape of the tree, or the model of sequence evolution. The effects of α and κ, however, may be confounded with the effects of branch lengths. Indeed, complicated patterns arise in the simulations concerning the effects of these parameters, as will be described below.

Simulation Results When the Substitution Model of Kimura (1980) Is Assumed

Results obtained by assuming trees 1A, 1B, 2A, 2B, 3A, and 3B of Table 3 are presented in Table 4, and results obtained from trees 4A, 4B, 5A, and 5B are presented in Table 5. Although only nine combinations of κ and α were examined for each tree, it is useful to bear in mind that the estimated probabilities for any method represent a ‘‘performance surface’’ over the κ–α plane.

Trees 1A and 1B (Equal External Branch Lengths)

There is a general deterioration in the performance of MP and JC when rate variation across sites becomes serious or when the transition/transversion rate bias gets large. The TRUE method, however, appears to improve with the increase of κ when tree 1A is assumed although it also becomes worse with the decrease of α. For both trees 1A and 1B, MP is considerably poorer than the two likelihood methods. JC also has lower probabilities of recovering the true tree than TRUE. In both cases, the differences are larger when α = 0.2 or κ = 20, that is, under a more complex model.

For parameter combination κ = 1 and α = ∞, JC performs slightly better than TRUE. The TRUE method which assumes the K80 model estimates κ from the data, and is not so efficient as JC, which makes use of the fact that the real κ = 1. This observation is true for other trees and is expected to apply for other cases where a model more complicated than the true model is used.

All methods recover the true tree with higher probabilities when tree 1B is assumed instead of tree 1A; the long branches of tree 1A mean too much noise in the sequences. The performance of MP relative to the like-

Table 4. Estimated probabilities (×100) that the true tree is recovered by different methods when the K80 model is assumed[a]

		$\alpha = \infty$			$\alpha = 0.8$			$\alpha = 0.2$		
	N	MP	JC	TRUE	MP	JC	TRUE	MP	JC	TRUC
1A: equal branch lengths										
$\kappa = 1$	250	64.2	75.8	75.2	59.8	68.1	65.2	46.1	54.9	52.3
	500	78.5	88.3	87.8	73.3	79.5	79.7	57.4	61.2	61.1
$\kappa = 5$	250	60.5	70.5	77.2	58.1	65.4	69.2	47.3	54.6	53.6
	500	72.7	82.9	91.0	70.8	77.0	80.5	58.2	60.8	64.7
$\kappa = 20$	250	53.6	59.2	78.4	52.3	55.7	66.9	45.9	53.7	58.9
	500	64.5	70.6	89.6	64.0	68.3	84.4	54.6	60.6	69.7
1B: equal branch lengths										
$\kappa = 1$	250	82.8	88.6	88.5	72.3	77.5	75.8	55.5	65.3	63.9
	500	95.9	96.5	96.5	87.2	91.5	91.4	70.1	79.2	77.0
$\kappa = 5$	250	77.9	80.9	82.7	66.8	75.1	77.2	51.9	59.5	61.9
	500	92.5	93.0	94.0	82.2	88.7	89.9	65.4	70.5	76.3
$\kappa = 20$	250	72.7	80.5	81.9	61.8	68.5	70.0	49.0	57.5	57.2
	500	88.0	90.8	93.1	76.2	80.4	82.4	59.6	65.3	70.0
2A: (short, short)–(long, long)										
$\kappa = 1$	250	56.8	44.8	44.6	54.3	58.5	51.5	43.3	45.1	39.7
	500	68.3	52.7	52.5	66.4	70.2	62.4	52.5	57.5	50.1
$\kappa = 5$	250	54.0	48.3	53.5	53.6	56.8	52.1	44.4	50.6	45.1
	500	65.0	57.1	67.3	64.2	64.7	61.3	53.8	55.7	47.7
$\kappa = 20$	250	48.6	47.1	62.7	50.2	49.9	59.2	44.2	50.8	48.6
	500	58.7	56.0	75.9	59.4	60.7	68.4	52.9	56.1	53.2
2B: (short, short)–(long, long)										
$\kappa = 1$	250	73.9	74.1	73.4	63.0	70.6	64.0	48.6	59.5	51.5
	500	88.1	87.7	87.6	77.4	82.4	77.5	61.6	67.2	57.8
$\kappa = 5$	250	69.6	71.9	74.2	59.5	62.6	62.9	47.9	55.0	49.7
	500	84.6	84.0	87.3	73.1	74.6	79.1	58.1	60.5	57.7
$\kappa = 20$	250	65.5	67.0	67.0	56.5	56.4	53.6	45.2	51.8	47.4
	500	81.0	78.9	80.4	70.0	71.4	68.5	54.6	59.9	59.9
3A: (short, long)–(short, long)										
$\kappa = 1$	250	34.7	78.7	78.1	39.3	63.1	72.4	36.6	45.2	54.2
	500	29.9	90.2	90.3	38.9	70.7	84.9	39.7	47.2	64.0
$\kappa = 5$	250	30.6	73.0	84.8	37.7	62.5	73.6	36.9	46.8	57.9
	500	25.8	84.2	95.5	37.8	73.5	86.2	39.6	49.8	68.2
$\kappa = 20$	250	21.1	57.1	83.0	32.6	53.4	75.4	36.8	45.9	63.4
	500	15.3	59.2	92.8	31.6	61.5	86.6	40.4	53.7	77.2
3B: (short, long)–(short, long)										
$\kappa = 1$	250	82.4	91.6	91.4	67.9	84.6	85.3	47.9	65.5	70.6
	500	95.0	97.3	97.7	81.6	96.0	96.0	56.6	75.8	83.9
$\kappa = 5$	250	75.9	89.1	90.8	61.2	79.2	83.0	43.6	60.6	70.1
	500	87.9	97.7	98.7	73.7	88.1	92.2	50.2	68.4	82.0
$\kappa = 20$	250	68.2	81.5	85.6	52.4	69.2	76.2	37.0	49.,0	60.0
	500	80.1	95.4	95.8	60.2	80.9	88.5	39.3	55.5	74.6

[a] MP stands for maximum parsimony, JC for maximum likelihood assuming the JC69 substitution model, and TRUE for maximum likelihood assuming the true model. The true model used for generating the data is K80 for $\alpha = \infty$, or K80 + dG for $\alpha = 0.8$ and 0.2

lihood methods does not seem to be any better for tree 1B than for tree 1A; the relative performance of MP is not improved by reducing the amount of evolution involved in the sequences.

Trees 2A and 2B: (Short, Short)–(Long, Long)

These two trees produced complicated results concerning the performance of the three methods. All six possible ranks of the three methods in performance were observed for different combinations of κ and α. Many of the differences are so large that it appears safe to draw the conclusion that each of the tree methods performs best for at least one combination of parameters. The performance surfaces of the three methods cross with one another over the κ-α plane. When $\kappa = 1$ and $\alpha = \infty$, the probability of recovering the true tree by MP is much

Table 5. The estimated probabilities (×100) of recovering the true tree by different methods when the K80 model and trees 4A, 4B, 5A, and 5B are assumed[a]

		$\alpha = \infty$			$\alpha = 0.8$			$\alpha = 0.2$		
	N	MP	JC	TRUE	MP	JC	TRUE	MP	JC	TRUE
4A: (short, short)–(short, long)										
$\kappa = 1$	250	97.1	97.7	98.3	92.0	92.4	87.0	76.8	79.6	63.2
	500	99.8	99.5	99.5	98.5	99.0	96.2	90.7	91.8	74.2
$\kappa = 5$	250	95.9	94.7	96.0	89.4	88.6	86.7	74.8	77.6	65.4
	500	99.8	99.0	99.5	98.0	98.0	98.5	89.4	86.7	78.7
$\kappa = 20$	250	95.0	91.8	90.0	86.1	84.1	78.9	72.1	71.1	69.4
	500	98.7	98.0	99.0	96.8	95.6	92.7	86.2	84.7	84.3
4B: (short, short)–(short, long)										
$\kappa = 1$	250	85.2	87.6	87.9	80.8	83.6	82.7	73.7	77.8	77.0
	500	97.5	98.5	98.5	94.8	97.4	96.4	90.5	91.8	89.7
$\kappa = 5$	250	83.6	88.6	89.1	78.9	85.7	84.7	71.0	76.6	74.8
	500	96.9	98.0	98.0	93.8	95.3	93.7	88.3	91.1	89.3
$\kappa = 20$	250	81.9	86.8	86.3	76.6	81.7	79.3	69.1	73.4	72.2
	500	95.9	95.8	96.2	91.6	93.0	91.2	85.3	88.3	84.3
5A: (long, long)–(long, short)										
$\kappa = 1$	250	60.0	64.5	64.2	52.2	58.3	54.2	40.8	46.8	44.1
	500	72.5	73.5	73.4	64.8	68.8	64.2	48.9	57.5	53.8
$\kappa = 5$	250	56.3	59.0	65.0	50.9	54.9	57.5	41.2	48.4	46.7
	500	69.2	71.6	77.4	62.2	67.6	66.7	49.6	55.0	51.8
$\kappa = 20$	250	52.1	52.5	61.7	47.7	51.0	59.0	39.9	44.6	51.2
	500	62.5	63.1	79.4	56.6	60.3	72.4	45.8	50.3	58.9
5B: (long, long)–(long, short)										
$\kappa = 1$	250	68.2	76.5	76.5	58.1	71.0	68.3	46.1	56.1	52.6
	500	84.5	88.4	88.3	74.5	79.2	78.3	59.4	69.2	66.0
$\kappa = 5$	250	63.8	72.9	71.7	54.1	62.4	62.8	44.8	55.0	55.9
	500	80.3	83.6	83.7	68.7	75.2	73.8	55.3	62.8	65.8
$\kappa = 20$	250	60.0	65.6	65.1	51.9	56.4	57.5	42.1	49.7	48.1
	500	75.0	80.2	80.2	64.7	70.0	69.2	51.2	56.2	54.4

[a] See note to Table 4

higher than the probability for JC or TRUE. There are also many cases where JC is much better than both MP and TRUE. The fact that MP or JC can perform better than TRUE deserves special attention. In Fig. 5, the asymptotics as $N \rightarrow \infty$ were examined, with MP and JC (which assumes the true model, JC69) for the case $\kappa = 1$, $\alpha = \infty$ taken as an example. For this purpose, the ratio of the probabilities for the two methods, $P_{\mathrm{ML}}/P_{\mathrm{MP}}$, is not an adequate measure of the efficiency of ML relative to MP. A better measure is $E_{\mathrm{ML,MP}} = (1 - P_{\mathrm{MP}})/(1 - P_{\mathrm{ML}})$, which is plotted in Fig. 5. The limit of $E_{\mathrm{ML,MP}}$ as $N \rightarrow \infty$ appears to be zero although a mathematical proof of this assertion does not seem possible. Another suitable measure is the inverse ratio of the sequence lengths needed by the two methods to recover the true tree with a prespecified probability P, that is, $E^{*}_{\mathrm{ML,MP}} = N_{P,\mathrm{MP}}/N_{P,\mathrm{ML}}$. $N_{P,\mathrm{MP}}$ can be calculated using the approximate formula (Eq. 2) of Zharkikh and Li (1992) for this set of branch lengths, while $N_{P,\mathrm{ML}}$ can be crudely estimated from Fig. 5. For $P = 0.5, 0.6, 0.7, 0.8, 0.9, 0.95$, $N_{P,\mathrm{MP}}$ is calculated to be 167, 310, 538, 906, 1,593, 2,316, respectively, while $N_{P,\mathrm{ML}}$ is estimated as 380, 910, 1600, 3,200, 5,100,

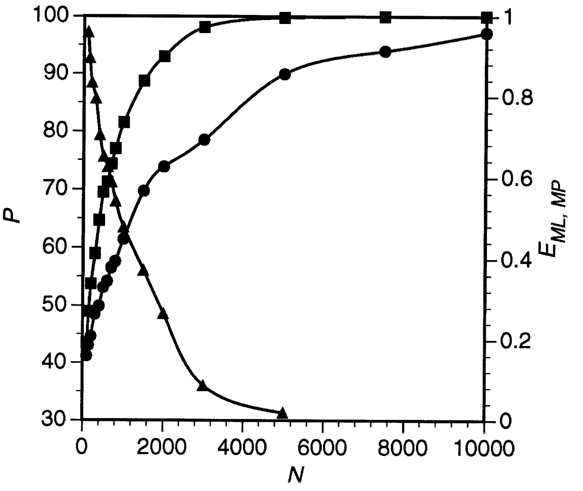

Fig. 5. MP can be better than ML! The estimated probability (P, ×100) that the true tree is recovered by MP (■), and by ML assuming the true model (JC69) (●), as a function of the sequence length (N). The JC69 model was assumed to generate data, with branch lengths taken from tree 2A of Table 3: $\mathbf{t} = (0.1, 0.5, 0.5, 0.6, 1.4)$ (see Fig. 1). The tree has the shape "(short, short)-(long, long)." The number of simulations for ML is 2,000, while that for MP is determined by $\sigma_p < 0.005$. The efficiency (▲) of ML relative to MP is calculated as $E_{\mathrm{ML,MP}} = (1 - P_{\mathrm{MP}})/(1 - P_{\mathrm{ML}})$.

and 7,950, respectively, and their ratio ($E^{*}_{\mathrm{ML,MP}}$) is then 0.439, 0.341, 0.336, 0.283, 0.312, and 0.291, respectively. So both measures suggest deterioration rather than improvement of the performance of ML relative to MP with the increase of *N*.

The fact that it is possible for ML assuming the true model to perform worse than MP or ML assuming a worse model, even when $N \to \infty$, will be discussed later. For the specific cases of trees 2A and 2B, the results are relatively easy to explain. The parsimony is well known to have a tendency to group long branches together and short branches together irrespective of the true topology (e.g., Felsenstein 1978), a property known as ''long branches attract'' (e.g., Penny et al. 1987). Felsenstein (1978) provided a detailed explanation of this property in terms of the probabilities of observing site patterns *xxyy, xyxy,* and *xyyx.* This tendency might be considered a ''bias'' in the parsimony method and is shared by likelihood or distance methods when an overly simplified model is assumed (e.g., Yang 1994b, 1995b). However, when the true tree does have the long branches clustered together, as in trees 2A and 2B, this property actually becomes an advantage and makes it possible for MP or JC to outperform ML assuming the true model.

In the case of tree 2A, the performance of MP becomes poorer with the increase of κ and/or decrease of α and is much poorer than JC or TRUE when $\kappa = 20$ or $\alpha = 0.2$. The two likelihood methods show different patterns with changes of κ and α; the best performance of JC occurs at $\kappa = 1$ and $\alpha = 0.8$ while the best performance of TRUE occurs at $\kappa = 20$ and $\alpha = \infty$.

All methods perform better for tree 2B than for tree 2A, which suggests that tree 2A involves too much evolution. It is noteworthy that the superiority of MP over the likelihood methods found at $\kappa = 1$ and $\alpha = \infty$ for tree 2A disappears when the branches are shortened to those of tree 2B, where MP is either worse or not any better than the likelihood methods. Figure 6 shows the probabilities of recovering the true tree by MP and JC when the true model is JC69, as functions of the amount of evolution measured by *b.* The performance of MP relative to ML improves with the (proportional) increase of branch lengths in the tree, and a small amount of evolution is not a necessary requirement for parsimony.

Trees 3A and 3B: (Short, Long)–(Short, Long)

The parsimony method is inconsistent for all combinations of κ and α when tree 3A is assumed. This is true for the case $\kappa = 20$ and $\alpha = 0.2$, where the probability of recovering the true tree increases when *N* is increased from 250 to 500. The inconsistency of MP means that is it necessarily inefficient (Yang 1995b). JC is uniformly poorer than TRUE, too, especially for large κ and small α. The differences among methods are quite large for almost all parameter combinations. The high probabilities for the TRUE method indicate that the notion of ''long branches attract'' is dependent on the tree reconstruction method.

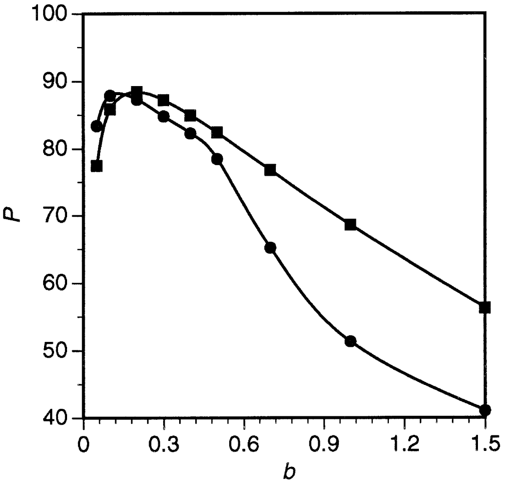

Fig. 6. Does maximum parsimony mean minimum evolution? The estimated probability (*P*, ×100) that the true tree is recovered by MP (■) and by ML (●) that assumes the true model (JC69), as a function of the amount of evolution measured by *b.* The branch lengths are based on tree 2A of Table 3, i.e., $\mathbf{t} = (0.1b, 0.5b, 0.5b, 0.6b, 1.4b)$. The sequence length is $N = 500$, and the results (not shown) for $N = 250$ or 1000 show the same pattern as this graph.

When tree 3B is assumed instead of 3A, MP improves greatly; indeed, for this tree, MP is consistent for all values of κ and α except the combination $\kappa = 20$ and $\alpha = 0.2$. Thus, the small amount of evolution involved in tree 3B remedied the problem of MP to some extent, although the method is still much poorer than both JC and TRUE. The performance of the three methods is invariably in the order MP < JC < TRUE. All methods deteriorate with the increase of κ and decrease of α, with MP and JC particularly so.

Trees 4A and 4B: (Short, Short)–(Short, Long)

All methods generally deteriorate with the increase of κ or decrease of α, although TRUE appears to improve with κ when $\alpha = 0.2$ for tree 4A. The differences among methods are not large in the case of tree 4A. When tree 4B is assumed, the probabilities of recovering the true tree by any method are lower than those for tree 4A. For short sequences ($N = 250$), parsimony is poorer than both JC and TRUE for all combinations of parameters for tree 4B and the differences are considerable in some cases, while for long sequences ($N = 500$), the three methods perform more or less the same.

Trees 5A and 5B: (Long, Long)–(Long, Short)

All three methods are consistent for all combinations of κ and α and for both trees 5A and 5B. Parsimony is poorer than both JC and TRUE for all values of parameters and for both trees. In the case of tree 5A, the differences in the probability range from 1 to 4% at $\kappa = 1$

and $\alpha = \infty$ to over 10% at $\kappa = 20$ and $\alpha = 0.2$, and MP becomes worse when the model becomes more complex. JC appears to be more efficient than TRUE at $\kappa = 1$ and $\alpha = 0.8$ or 0.2, while for other combinations TRUE gives much better performance than both MP and JC. When tree 5B is assumed, the performances of JC and TRUE are similar while MP is much poorer than the likelihood methods.

With four species, if a tree reconstruction method groups together species 1 with 2, species 3 and 4 will be automatically clustered. So one may expect results for trees 4A, 4B, 5A, and 5B to be similar to those for trees 1A and 1B. This appears to be the case, in that parsimony is never too wrong for these three tree shapes—it is consistent for all values of parameters examined—and in that parsimony is almost always poorer than the two likelihood methods. Although the probability of recovering the true tree depends on values of branch lengths as well as the tree shape, the shape ‘‘(short, short)–(short, long)’’ (trees 4A and 4B) seems easier to recover than the shape ‘‘(long, long)–(long, short)’’ (trees 5A and 5B). This is particularly the case with the parsimony method.

Simulation Results When the Substitution Model of Hasegawa et al. (1985) Is Assumed

Due to the similarity of results for trees 4A, 4B, 5A, and 5B to those for trees 1A and 1B, the former trees were not used anymore when the effects of nucleotide-frequency bias were examined by assuming the HKY85 model. With very few exceptions, the probabilities that the three methods recover the true tree for trees 1A, 1B, 2A, 2B, 3A, and 3B when the HKY85 model is assumed are very similar to, but slightly smaller than, the corresponding probabilities of Table 4 obtained for K80. The differences are almost always smaller than 5%. The results under the HKY85 model are not presented due to space limitation but are available (in a table of the form of Table 4) from the author unpon request. The few exceptions mentioned above include several parameter combinations at $\alpha = \infty$ for tree 2A, where JC has higher probabilities of recovering the true tree when HKY85 is assumed instead of K80.

Concerning differences among methods, the same patterns were found for the HKY85 model as those found in Table 4 for K80. For example, all methods generally become poorer with the increase of κ and decrease of α. Parsimony performs better than the two likelihood methods when $\alpha = \infty$, $\kappa = 1$ in the case of tree 2A. For tree 3A, parsimony is inconsistent for all parameter values, while for tree 3B it is inconsistent for the combination $\alpha = 0.2$ and $\kappa = 20$ only. The base-frequency bias therefore seems to reduce the chance of recovering the true tree by any methods, but the effect seems to be minor. Ignoring base-frequency bias does not seem to cause much problem for the tree-reconstruction methods examined here.

Estimation of a Statistical Parameter and Estimation of a Tree Topology

In this section, the similarities and differences of phylogenetic tree estimation when compared with the estimation of a statistical parameter will be examined. A review of concepts in statistical estimation highlights the complexities of phylogenetic tree reconstruction. Recognizing these complexities will be helpful for improving current tree reconstruction methods or for devising new ones.

The Problem of Statistical Estimation: Concepts

In statistical parameter estimation, ‘‘[w]e know, or assume as a working hypothesis, that the parent population is distributed in a form which is completely determined but for the value of some parameter θ’’ (Kendall and Stuart 1979:1). In other words, the probability function of observing data X, $f(X;\theta)$, is fully specified although the value of the parameter θ may be unknown. A function of the data, $t(X)$, is called a statistic, and may be taken as an estimator of the parameter θ. The estimator t is said to be unbiased if $E(t) = \theta$, where the expectation is taken over X, and is said to be consistent if, for any given small positive numbers ϵ and η, we can find an N_0 such that prob$(t - \theta > \epsilon) < \eta$ for all $N > N_0$ (Kendall and Stuart 1979:3–4). The variance of a consistent and unbiased estimator cannot be smaller than the Cramér-Rao lower bound; that is, for any unbiased estimator t of θ, we have (Kendall and Stuart 1979:8–10)

$$\mathrm{var}(t) \geqslant -1/E\left[\frac{\partial^2 \log\{f(X;\theta)\}}{\partial\theta^2}\right] \tag{4}$$

An estimator that attains this lower bound for all θ is called a minimum variance bound estimator. When this bound is not attainable, there may still be an estimator that has, uniform in θ, smaller variance than any other estimator. Such an estimator may not exist, but if it does, it is unique and is called a minimum variance estimator (Kendall and Stuart 1979:17).

Maximum likelihood is a methodology for estimating parameter θ after the data X are observed. The probability of observing the data, $f(X;\theta)$, is taken as a function of the parameter θ, and the value of θ that maximizes $f(X;\theta)$ is the maximum likelihood estimator (MLE) of θ. Note that changing the value of θ will change the value of f but not its functional form. Under very general ‘‘regularity’’ conditions, the MLE has desirable asymptotic properties: that is, when $N \rightarrow \infty$, an MLE is consistent, asymptotically unbiased, and normally distributed, and attains the minimum variance bound of (4) (Kendall and Stuart 1979:38–81).

The Problem of Phylogenetic Tree Estimation: Complexities

Following the pioneering work of Cavalli-Sforza and Edwards (1967; Edwards 1970), Felsenstein (1981; see also Felsenstein 1973; Thompson 1975; Bishop and Friday 1985) suggested a method for reconstructing the phylogenetic tree using DNA sequence data which has been known as the maximum likelihood estimation (or, more often, maximum likelihood inference!) of the phylogeny. Given a tree topology, the probability of observing the data, $f(X;\ \theta)$, is used as the ''likelihood function'' for estimating parameters θ, which include branch lengths in the tree and parameters in the substitution model. The optimum value of the likelihood for this tree is also obtained. The same process is repeated for other tree topologies, and the (optimum) likelihood values for different tree topologies are compared to select the best topology (see Table 2 for a example). This method has similarity with the conventional maximum likelihood in that $f(X;\ \theta)$ will be the probability of observing the data if the true tree, true branch lengths, and true values of other parameters are used in the function. It, however, does not fit in the framework of statistical parameter estimation as the aforementioned working hypothesis is invalid in the context of topology estimation. Different topologies involve different sets of branch-length parameters; and the functional form of $f(X;\ \theta)$ changes with tree topology, so that the distribution of the data is not fully specified without knowledge of the true topology (Yang 1994b; Yang et al. 1995; see also Nei 1987:325). The problem of phylogeny reconstruction concerns more the question of what the (branch length) parameters are than the question of what numerical values the parameters take. The implementation of the method is remarkably more complicated than the traditional maximum likelihood, in that the maximization of the likelihood function(s) by numerical iteration has, at least in theory, to be performed as many times as the number of tree topologies. Literally it is a *maximum maximum likelihood method.*

The simple substitution model of Felsenstein (1981) has been improved in many ways, for example, to account for more complicated patterns of nucleotide substitution (e.g., Hasegawa et al. 1985; Yang 1994a) and to accommodate the variation and dependence of substitution rates across nucleotide sites (Yang 1993, 1994c, 1995a). The same methodology was applied to homologous protein sequences (Bishop and Friday 1985, 1987; Kishino et al. 1990) and protein-coding DNA sequences with the codon structure taken into account (Goldman and Yang 1994; Muse and Gaut 1994). Nevertheless, the methodology has remained unchanged.

The failure to recognize the complexity of the problem has caused much controversy in theoretical studies of tree reconstruction methods. Felsenstein (1973, 1978) referred to the regularity conditions of Wald (1949) for a proof of the consistency of the maximum likelihood method for estimating the tree topology. These conditions would include the continuity and differentiability of the likelihood function with respect to the topology parameter. Such concepts are not defined. Concerning the nature of the tree topology, there have been suggestions that it is a statistical parameter (e.g., Thompson 1975; Goldman 1990), and there are also suggestions that it is not (e.g., Felsenstein 1988). As the role of the topology is to specify the branch-length parameters and the form of the likelihood function, the topology certainly differs from a parameter which takes only discrete numerical values. Note that the topology also differs from a statistical hypothesis, which concerns the values of parameters rather than what the parameters are. The controversy concerning the use of the nonparametric bootstrap in the test of phylogenies suggested by Felsenstein (1985) appears in a large part to be due to our lack of understanding of the nature of the topology (e.g., Zharkihk and Li 1992; Felsenstein and Kishino 1993; Li and Zharkikh 1994).

It is, nevertheless, no surprise that the likelihood method of phylogenetic tree estimation does not share all asymptotic properties of the maximum likelihood estimator of a statistical parameter. Yang (1994b) suggested a proof that the likelihood estimation of the tree topology is consistent despite the complexity of the parameter space. Results of this simulation suggest that the method is not always asymptotically most efficient. Intuitively, almost all tree-reconstruction methods select the tree according to some criterion, which measures the compatibility of the tree with the data. The likelihood method recovers the true tree if and only if the likelihood value for the true tree is greater than those for the wrong trees. When a wrong model is used, the likelihood values of all tree topologies are decreased (normally by a great margin; see, e.g., Yang et al. 1994, 1995), but it is possible for the aforementioned condition (i.e., the likelihood of the true tree being the greatest) to be satisfied more often if a wrong model is assumed than if the true model is assumed. The fact that using a wrong model can recover the true tree with a higher probability has been observed in computer simulations where distance-matrix methods were used (e.g., Saitou and Nei 1987; Sourdis and Krimbas 1987; Tateno et al. 1994). Schoeniger and von Haeseler (1993) and Tajima and Takezaki (1994) experimented with several distance estimates based on such ''wrong'' models and found that they could give better performance than the distances based on the true model.

In sum, phylogenetic tree reconstruction remains a difficult estimation problem. Use of the true model fails quite often to give the best performance in the simulations. The concept of unbiasedness of an estimated topology is yet to be defined, but one can ask the important question of whether there exists a method that has higher probability of recovering the true tree than any other methods for the whole parameter space; such a method would be equivalent to the minimum-variance estimator of a statistical parameter. Perhaps it is not that using the true model sometimes fails to give the best performance but rather that we have not found the right way of using the true model.

Yang et al. (1995) noted that different tree topologies have similarities to different statistical distributions. One possibility that might be worth exploring is construction of a ''super model'' that has all tree topologies as its special cases, in the same way that a family of distributions can encompass many different distributions. For example, the Katz family of discrete distributions (random variable $x = 0, 1, \ldots,$ with probability $p_0, p_1, \ldots$) can be specified as

$$p_{x+1}/p_x = (a + bx)/(1 + x),\ x = 0, 1, \ldots \qquad (5)$$

where $a > 0$, $b < 1$. Katz (1946, 1965; see also Johnson et al. 1992:77–81) showed that $b < 0$, $b = 0$, and $0 < b < 1$ in the Katz family give rise to the binomial, Poisson, and negative binomial distributions, respectively. The distinction among the distributions is then equivalent to a test of hypotheses concerning the parameter b in the Katz family of distributions. Such an approach, if it could work, would drastically reduce the computation involved in phylogenetic estimation.

Discussion

It is noteworthy that the simple JC method performed quite well in the simulations, even though the model (JC69) is wrong. In almost all cases, it is much more efficient than parsimony, and, when the tree is difficult to reconstruct, it is less likely to be misleading than parsimony (results for tree 3A, Table 4). The JC69 model has computational advantages over more complex models and they may be made use of. It should also be noted that other factors may be more important in affecting the probability of recovering the true tree than the choice of methods. The shape of the tree and the branch lengths affect the success of tree reconstruction significantly. The base frequency bias does not seem to have posed a great problem, but the transition/transversion rate bias and, in particular, the variation of substitution rates across sites generally reduce the probability of recovering the true tree. By careful sampling of species as well as gene sequences, biologists may be able to control these factors to their advantage.

The author of this paper has been unable to see any connection between the parsimony method of phylogenetic tree reconstruction and the parsimony or simplicity principle of science and philosophy, or any scientific merit of discussions that claim such a connection. In this study, parsimony is considered a well-defined method of data analysis and its performance for this purpose is examined. Nevertheless, both the intuitive examination and the computer simulation employed in this study for identifying the assumptions underlying parsimony involve difficulties. First of all, an analytical method may not be sensitive to its assumptions, and indeed, the reconstruction of the tree topology by model-based methods is known to be quite tolerant to violations of the assumptions (e.g., Fukami-Kobayashi and Tateno 1991; Debry 1992; Gaut and Lewis 1995). It is also commonplace that an assumption made to derive certain results was later shown to be unnecessary. The greatest difficulty is our lack of a method that is known to give the best performance for all values of parameters, with which other methods such as parsimony can be compared.

At any rate, the hierarchy of models assumed in this study can be arranged, say, in the order JC69, K80, K80 + dG, HKY85 + dG, so that the model becomes more and more complex and realistic. Then the intuitive examination suggests that the model underlying parsimony is closer to JC69 than to HKY85 + dG. This expectation has certainly been confirmed by the simulation results. The pattern is clearest when there exist considerable differences among the methods, for example, in the case of tree 2A where MP and JC can outperform ML assuming the true model, and in the case of tree 3A where both MP and JC do not perform well. Further evidence is that the performance of MP (and JC) in comparison with ML assuming the true model generally deteriorates when the model becomes more complex (Tables 4 and 5). Recent attempts to use ''step matrices'' in the parsimony analysis (Maddison and Maddison 1992; Swofford 1993), despite their arbitrary nature, have made it possible for the method to use, to some extent, information and knowledge concerning the process of sequence evolution such as the transition/transversion rate bias, and have relaxed some of the stringent assumptions about the substitution process as identified in this paper. An important factor that does not seem to have been considered is the difference of branch lengths. Failure to take into account different branch lengths of the tree appears to be the major reason for the failure of parsimony in cases such as trees 3A and 3B. It seems possible to modify the method so that differences in branch lengths are explicitly considered in the calculation of the parsimony score. The resulting method might be a parsimony with a time structure and, at the same time, a likelihood without iteration, and might have advantages of both methods.

Acknowledgments. I thank Brandon Gaut, Nick Goldman, Masatoshi Nei, and Jeff Thorne for many helpful comments on the manuscript. This study was supported by a grant from the National Natural Science Foundation of China to Z.Y. and by NIH and NSF grants to M. Nei.

References

Bishop MJ, Friday AE (1985) Evolutionary trees from nucleic acid and protein sequences. Proc R Soc Lond [Biol] 226:271–302

Bishop MJ, Friday AE (1987) Tetropad relationships: the molecular evidence. In: Patterson C (ed) Molecules and morphology in evolution: conflict or compromise? Cambridge University Press, Cambridge, pp 123–129

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates, tempo and mode of evolution. J Mol Evol 18:225–239

Camin J, Sokal R (1965) A method for deducing branching sequences in phylogeny. Evolution 19:311–326

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Evolution 21:550–570

Debry RW (1992) The consistency of several phylogeny-inference methods under varying evolutionary rates. Mol Biol Evol 9:537–551

Eck RV, Dayhoff MO (1966) Inference from protein sequence comparisons. In: Dayhoff MO (ed) Atlas of protein sequence and structure. National Biomedical Research Foundation, Silver Spring, MD, pp 161–202

Edwards AWF (1970) Estimation of the branch points of a branching diffusion process with discussion. J R Stat Soc B 32:155–174

Edwards AWF, Cavalli-Sforza LL (1963) The reconstruction of evolution. Heredity 18:553

Farris J (1973) On the use of the parsimony criterion for inferring evolutionary trees. Syst Zool 22:250–256

Farris J (1977) Phylogenetic analysis under Dollo's law. Syst Zool 26:77–88

Felsenstein J (1973) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst Zool 22:240–249

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

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet 22:521–565

Felsenstein J, Kishino H (1993) Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Syst Biol 42:193–200

Felsenstein J, Sober E (1986) Parsimony and likelihood: an exchange. Syst Zool 35:617–626

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst Zool 20:406–416

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

Gaut BS, Lewis PO (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. Mol Biol Evol 12:152–162

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

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736

Hartigan JA (1973) Minimum evolution fits to a given tree. Biometrics 29:53–65

Hasegawa M, Fujiwara M (1993) Relative efficiencies of the maximum likelihood, maximum parsimony, and neihbor joining methods for estimating protein phylogeny. Mol Phyl Evol 2:1–5

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

Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160–174

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

Hill ID (1973) Algorithm AS 66: the normal integral. Appl Stat 22: 424–427

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

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

Johnson NJ, Kotz S, Kemp AW (1992) Univariate discrete distributions, 2nd ed. Wiley, New York

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–123

Katz L (1946) On the class of functions defined by difference equation $(x + 1)f(x + 1) = (a + bx)f(x)$ (abstract). Ann Math Stat 17:501

Katz L (1965) Unified treatment of a broad class of discrete probability distributions. In: Patil GP (ed) Classical and contagious discrete distributions. Pergamon Press, Oxford, pp 175–182

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

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

Kishino H, Miyata T, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin or chloroplasts. J Mol Evol 31:151–160

Kuhner MK, Felsenstein J (1994) A simulation of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459–468

Li WH, Zharkikh A (1994) What is the bootstrap technique? Syst Biol 43:424–430

Maddison WP, Maddison DR (1992) MacClade: analysis of phylogeny and character evolution, version 3. Sinauer, Sunderland, MA

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to chloroplast genome. Mol Biol Evol 11:715–724

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Pauling L, Zuckerkandl E (1963) Chemical paleogenetics: molecular ''restoration studies'' of extinct forms of life. Acta Chem Scand 17:S9–S16

Penny D, Hendy MD, Henderson IM (1987) The reliability of evolutionary trees. Cold Spring Harb Symp Quant Biol 52:857–862

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

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

Schoniger M, von Haeseler A (1993) A simple method to improve the reliability of tree reconstructions. Mol Biol Evol 10:471–483

Sober E (1988) Reconstructing the past: parsimony, evolution, and inference. MIT Press, Cambridge, MA

Sourdis J, Krimbas C (1987) Accuracy of phylogenetic trees estimated from DNA sequence data. Mol Biol Evol 4:159–166

Swofford DL (1993) Phylogenetic analysis using parsimony (PAUP), version 3.1. University of Illinois, Champaign

Swofford DL, Olsen GJ (1990) Phylogeny reconstruction. In: Hillis

DM, Moritz G (eds) Molecular systematics. Sinauer, Sunderland, MA, pp 411–501

Tajima F, Takezaki N (1994) Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. Mol Biol Evol 11: 277–286

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

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

Thompson EA (1975) Human evolutionary trees. Cambridge University Press, Cambridge

Wald A (1949) Note on the consistency of the maximum likelihood estimate. Ann Math Statist 20:595–601

Wiley E (1975) Karl P. Popper, systematics, and classification: a reply to Walter Bock and other evolutionary taxonomists. Syst Zool 24: 233–242

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

Yang Z (1994a) Estimating the pattern of nucleotide substitution. J Mol Evol 39:105–111

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

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

Yang Z (1995a) A space-time process model for the evolution of DNA sequences. Genetics 139:993–1005

Yang Z (1995b) Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. J Mol Evol 40:689–697

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

Yang Z, Goldman N, Friday AE (1995) Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst Biol 44:384–399

Zharkikh A, Li WH (1992) Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: I. Four taxa with a molecular clock. Mol Biol Evol 9:1119–1147

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

Zuckerkandl E (1964) Further principles of chemical paleogenetics as applied to the evolution of hemoglobin. In: Peeters H (ed) Protides of the biological fluids. Elsevier, Amsterdam, pp 102–109