J Mol Evol (1991) 32:443–445

Journal of
Molecular Evolution

On the Maximum Likelihood Method in Molecular Phylogenetics

Masami Hasegawa,[1] Hirohisa Kishino,[1] and Naruya Saitou[2]

[1] The Institute of Statistical Mathematics, and Department of Statistical Science, The Graduate University for Advanced Studies, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan
[2] Department of Anthropology, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113, Japan

Summary. The efficiency of obtaining the correct tree by the maximum likelihood method (Felsenstein 1981) for inferring trees from DNA sequence data was compared with trees obtained by distance methods. It was shown that the maximum likelihood method is superior to distance methods in the efficiency particularly when the evolutionary rate differs among lineages.

Key words: Efficiency — Maximum likelihood method — Methods for inferring trees — DNA sequence data

Recently, Saitou (1988) and Saitou and Imanishi (1989) examined the efficiencies of the maximum likelihood (ML) method, which does not assume rate constancy (Felsenstein 1981), and of other methods for inferring trees from DNA sequence data by using computer simulation. Their results contain some inappropriate points, and we reexamine these points in this communication.

Saitou (1988) compared the relative efficiency of the ML method with the efficiencies of the maximum parsimony method and of the distance method. Computer simulations based on Jukes and Cantor's (1969) model (transition rate α is equal to transversion rate β) were done for a tree with four OTUs with varying rate model. In this model, expected numbers of substitutions per site along each of two nonneighboring branches are 1.0 and those along other branches including the internal branch are 0.1 ($v_1 = 1.0$ and $v_2 = 0.1$ in Fig. 1). The ML method chose the correct tree with a frequency of only 43% whereas the distance method with Jukes–Cantor distances produced a correct tree with a frequency of 74% [Table 5 of Saitou (1988)]. However, when we reexamined the efficiencies of the ML method by using Felsenstein's programs as shown below, it turned out that the ML method gave a higher performance in obtaining the correct tree than the distance method (Table 1). This indicates that Saitou's (1988) ML program was not efficient for this model tree.

Assuming the tree of Fig. 1 to be the model tree, a computer simulation was done with $v_1 = 1.0$ and $v_2 = 0.1$ by a procedure similar to Saitou's (1988). Sequences with 500 nucleotides were generated assuming that transition and transversion occur at the same rate, and 100 replications were obtained. The programs for the ML analysis used in this work are two distinct versions of DNAML in Felsenstein's program package PHYLIP; the earlier one (version 2.3) assumes equal rates of occurrence between transition and transversion, whereas the later one (version 3.1) assumes unequal rate. In the latter program, the rate of transition was set to be twice that of transversion by using an option. This contradicts the method of simulation but was used in order to examine the robustness of the likelihood method against the violation of the assumption based on the transition/transversion ratio.

The results are summarized in Table 1. Version 2.3 of DNAML chose the correct tree with a frequency of 92%. Version 3.1, although the α/β ratio assumed in the analysis contradicts the simulation, chose the correct tree with a frequency of 82%, and even better efficiency was available than that for the distance method that assumed the same α/β ratio with that in the simulation. Furthermore, we carried out simulations for $v_1 = 1.0$ and $v_2 = 0.2$ and

Offprint requests to: M. Hasegawa

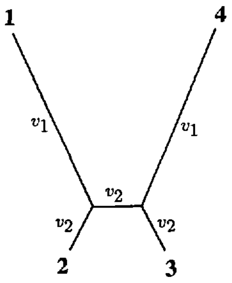

Fig. 1. A model tree used in simulations. v_1 and v_2 refer to the branch lengths; i.e., expected numbers of substitutions along respective branches.

Table 1. Proportions (%) of trees in which the correct topology was reconstructed

	Method			
Branch lengths	Maximum likelihood DNAML version 2.3 (model: $\alpha = \beta$)	Maximum likelihood DNAML version 3.1 (model: $\alpha = 2\beta$)	Distance method (model: $\alpha = \beta$)	Maximum parsimony
$v_1 = 1.0$, $v_2 = 0.1$	92	82	70	0
$v_1 = 1.0$, $v_2 = 0.2$	96	91	88	0
$v_1 = 0.5$, $v_2 = 0.1$	100	100	100	0

Simulations were replicated 100 times based on Jukes and Cantor's model (the transition rate α is equal to the transversion rate β)

for $v_1 = 0.5$ and $v_2 = 0.1$. In the former case, the model of $\alpha = \beta$ chose the correct tree with a frequency of 96% whereas the model of $\alpha = 2\beta$ chose it with a frequency of 91%, and in the latter case both versions of DNAML chose the correct tree with a frequency of 100% (the distance method also gave 100% efficiency), suggesting extensive robustness of the ML method against violation of the assumption based on the transition/transversion ratio. Fukami-Kobayashi and Tateno (1990) studied the robustness of the ML method by more extensive simulations. It should be noted that, although the maximum parsimony method is positively misleading for the cases examined here (Felsenstein 1978; Hasegawa and Yano 1984), the ML method and also the distance method can provide good efficiency in these cases.

Table 2. Proportions (%) of trees in which the correct topology was reconstructed

	Maximum likelihood method		
	DNAML version 2.3 (model: $\alpha = \beta$)	DNAML version 3.1[a] (model: $\alpha = 2\beta$)	Neighbor-joining method[a] (model: $\alpha = \beta$)
A			
8a = 0.05			
300 bp	52	38	40
600 bp	88	80	82
8a = 0.50			
300 bp	48	48	46
600 bp	78	70	82
B			
8a = 0.05			
300 bp	80	62	70
600 bp	90	88	86
8a = 0.50			
300 bp	66	56	60
600 bp	80	76	70
C			
a = 0.01			
300 bp	84	78	72
600 bp	96	98	92
a = 0.05			
300 bp	92	92	68
600 bp	100	100	96
D			
a = 0.01			
300 bp	86	80	74
600 bp	96	96	92
a = 0.05			
300 bp	96	96	78
600 bp	100	100	100

Simulation data are from Saitou and Imanishi (1989). Constant rate of nucleotide substitution is assumed for A and B, whereas there is a large variation in the rate for C and D. a is a unit branch length. Simulations were replicated 50 times based on Jukes and Cantor's model

[a] From Saitou and Imanishi (1989)

Saitou and Imanishi (1989) studied the relative efficiencies of the Fitch–Margoliash, minimum-evolution, and neighbor-joining (NJ) methods in addition to the ML and maximum parsimony methods for the four model trees with six OTUs. They used Felsenstein's DNAML program (version 3.1) for the ML method. For the constant rate models, they concluded that the ML method is slightly less efficient than the NJ and minimum-evolution methods [Table 1 of Saitou and Imanishi (1989)]. However, as they acknowledged and as was also mentioned above, in the DNAML of version 3.1 we cannot assume equal rates between transition and transversion, and they assumed that the rate of transition is twice that of transversion while the simulation was performed based on Jukes and Cantor's model. Because the distance analysis was performed based on the assumption of an equal transition/transversion rate, their way of evaluation was biased against the ML method. Therefore, we reexamined their simulation data by using the earlier version (2.3) of DNAML and assuming an equal transition/transversion rate (Table 2).

In the case of rate constancy among branches, although the ML method with an inappropriate model of $\alpha = 2\beta$ is slightly less efficient than the NJ method with $\alpha = \beta$, the former with $\alpha = \beta$ is more efficient than the latter except in the case of A with 8a = 0.50, 600 bp. This suggests that the ML method is somewhat sensitive to the transition/transversion ratio under this model. In the case of variable rate among branches, the ML method is more efficient even with the model of $\alpha = 2\beta$ than the NJ method with $\alpha = \beta$ (Saitou and Imanishi 1989). The ML method that does not assume rate constancy is better when the evolutionary rate varies widely among lineages (cases C and D), whereas it is as efficient as the NJ method when the evolutionary rate is constant (cases A and B).

Although the NJ method was shown to be slightly inferior to the ML method in obtaining the correct tree topology, it is more efficient than other existing tree-making methods (Saitou and Nei 1987; Sourdis and Nei 1988; Saitou and Imanishi 1989). In any case, those simulation studies were based on simple assumptions, and further study may be necessary to evaluate the efficiencies of different tree-making methods. Because of the simplicity of the NJ method algorithm, it seems to be a useful method that is complementary to the ML method.

Acknowledgments. We are grateful to Miss M. Fujiwara for help with the computer. This work was supported by grants from the Ministry of Education, Science, and Culture of Japan, and by the Joint Studies Program of The Graduate University for Advanced Studies.

References

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

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

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

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

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism, vol III. Academic Press, New York, pp 21–132

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

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 M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

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

Received September 17, 1990/Revised December 10, 1990