Syst. Zool., 38(1):1–6, 1989

A THEORETICAL STUDY OF THE UNDERESTIMATION OF BRANCH LENGTHS BY THE MAXIMUM PARSIMONY PRINCIPLE

NARUYA SAITOU

Department of Anthropology, Faculty of Science, The University of Tokyo,
Hongo, Bunkyo-ku, Tokyo 113, Japan

Abstract.—The degree of underestimation of branch lengths by the maximum parsimony principle is studied. The expected number of nucleotide changes per site under the maximum parsimony principle is computed, and it is compared with the expected number of nucleotide substitutions. A tree topology with no hierarchical structure is considered for mathematical simplicity. It is shown that as long as the evolutionary distance is less than 0.2, the maximum parsimony principle gives good estimates of nucleotide substitutions. When the evolutionary distance is greater than 0.2, however, the method gives gross underestimates of nucleotide substitutions. [Branching; parsimony; phylogenetics; topology.]

There are two major problems in constructing a phylogenetic tree from molecular data. One is the determination of the topology of a tree and the other is the estimation of branch lengths. For the first problem, the maximum parsimony method (Camin and Sokal, 1965; Fitch, 1977) has been extensively used for amino acid or nucleotide sequence data. For the estimation of branch lengths, however, this method is expected to underestimate the number of amino acid or nucleotide substitutions. This property comes from the principle of the method itself: minimize the number of changes required. Thus each branch length (estimated by Fitch's [1971] method) is expected to be smaller than the real length on average. In spite of this known shortcoming, the maximum parsimony method seems to be quite appropriate for the estimation of branch length in terms of amino acid or nucleotide substitutions when closely related sequences are compared, since the probability of backward and parallel substitutions is negligible in this situation. But how close should sequences be? The number of sequences compared is also related to this problem, because we expect to extract more and more changes as the number of sequences is increased. So far, there seems to be no theoretical study on these subjects.

In this paper, I show the effect of the amount of divergence and the number of nucleotide sequences on the estimates of branch lengths by the maximum parsimony principle. For simplicity, I consider the model of random nucleotide substitution (Jukes and Cantor, 1969). A constant rate of evolution, or the molecular clock, is also assumed. Further, a tree topology with no hierarchical structure is considered. Under these assumptions, the expected number of required nucleotide changes per site estimated by the maximum parsimony principle is computed.

MATHEMATICAL MODEL

The possible number of tree topologies is astronomical even when sequences from only ten or more taxa are considered (Felsenstein, 1978). Therefore I consider only one of these possible topologies, the tree with no hierarchical structure, as shown in Figure 1. It means that the tree topology is assumed known without applying the maximum parsimony method. Thus the problem considered in this paper is not exactly the maximum parsimony method, but the principle of maximum parsimony is used for the estimation of branch lengths. Thus, we call this narrower usage of the maximum parsimony principle "the maximum parsimony procedure" in the following. It is assumed that all extant sequences started to diverge at the same time

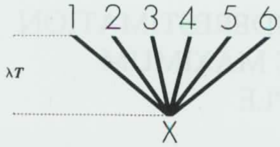

FIG. 1. An example of the topology with no hierarchical structure.

TABLE 1. Five nucleotide configurations for sequences A, B, and C.[a]

Configuration	A	B	C
C_1	X	X	X
C_2	X	X	Y
C_3	X	Y	X
C_4	Y	X	X
C_5	X	Y	Z

[a] X, Y, and Z are three different nucleotides.

from the ancestral sequence X. λT is the expected number of nucleotide substitutions per nucleotide site for each branch, where λ is the rate of nucleotide substitution per site per year, and T is the divergence time. We start from the simplest case, in which there are only two sequences involved, and consider up to six sequences. Then the case of the infinite number of sequences is considered.

To measure the amount of underestimation of nucleotide substitutions by the maximum parsimony procedure, the expected number of required nucleotide changes per branch per site under this procedure is computed. For the tree of Figure 1, this is given by the expectation of the total number of required nucleotide changes divided by the number of sequences compared. This expected number for a tree of n sequences is denoted by E(n).

For the computation of E(n), we also assume the pattern of nucleotide substitution to be random among the four nucleotides (A, T, G, and C), that is, Jukes and Cantor's (1969) model is used. Under this model, the probability that a nucleotide at an extant sequence is the same as that of the ancestral sequence is given by

$$p = 1/4 + 3\exp(-4\lambda T/3)/4, \qquad (1a)$$

and the probability that a nucleotide at a sequence is different from the ancestral one is given by

$$q = 1/4 - \exp(-4\lambda T/3)/4 \qquad (1b)$$

(see, e.g., Saitou and Nei, 1986). Note that $p + 3q = 1$. These two probabilities are the basis of the following computation.

The basic unit of information for the maximum parsimony procedure is the nucleotide configuration. Nucleotide configuration is the pattern of nucleotide arrangement for sequences compared in terms of nucleotide difference. For example, there are five nucleotide configurations for three sequences (see Table 1). In general, the number (c) of possible nucleotide configurations for n sequences are given by

$$c = (4^{n-1} + 3\cdot 2^{n-1} + 2)/6 \qquad (2)$$

(Saitou and Nei, 1986). For topologies with no hierarchical structure, however, it is not necessary to distinguish sequences, and the computation of E(n) can be greatly simplified, as shown below.

RESULTS

Two sequences.—When two sequences are compared, there are only two nucleotide configurations. Two nucleotides at a site are identical or different from each other. Because the substitution pattern assumed in this study is symmetrical and the process of nucleotide substitution is time-reversal, we can assume one of the sequences to be ancestral and the other to be extant. In this case, the time interval in equations (1a) and (1b) becomes 2T. Noting that any of three different nucleotides can be at the position of the extant site, the probability (π) of observing the event that the extant nucleotide is different from the ancestral one is

$$\pi = 3[1/4 - \exp(-4\lambda\cdot 2T/3)/4]. \qquad (3)$$

Because one nucleotide change is required for sites in which two sequences are different, the expected number [E(2)] of required nucleotide changes for two sequences is

$$E(2) = \pi/2 = 3(1 - \alpha^2)/8, \qquad (4)$$

where $\alpha = \exp(-4\lambda T/3)$.

Three sequences.—There are five nucleotide configurations for three sequences (Table 1). Configuration C_1 requires no nucleotide change, whereas configurations C_2, C_3, and C_4 require one nucleotide change, and configuration C_5 requires two changes. Thus,

$$E(3) = (U_2 + U_3 + U_4 + 2U_5)/3, \quad (5)$$

where U_i is the probability for observing configuration C_i.

There are three possible ways the nucleotide arrangement in configuration C_2 could be observed: (1) two nucleotides are the same as the ancestral one, (2) one nucleotide is the same as the ancestral one, and (3) all three nucleotides are different from the ancestral one (see Fig. 2A). Because nucleotide Y in case (1) can be any of three nucleotides that are different from the ancestral one (X), the probability for this case becomes $3p^2q$. Similarly, probability for case (2) becomes $3pq^2$. On the other hand, there are two nucleotides (Y and Z) that are different from the ancestral one (X) in case (3), and we can have six possibilities for Y and Z. Therefore the probability for case (3) becomes $6q^3$. Thus,

$$U_2 = 3p^2q + 3pq^2 + 6q^3. \quad (6a)$$

Configurations C_3 and C_4 become identical with configuration C_2 if we ignore the label of sequences. Therefore,

$$U_3 = U_2 \text{ and } U_4 = U_2. \quad (6b)$$

As for configuration C_5, there are two possible ways the nucleotide arrangement could be observed: (1) one of three different nucleotides is identical with the ancestral one and (2) all three nucleotides are different from the ancestral one (see Fig. 2B). There are six possibilities for the choice of nucleotides Y and Z for case (1), and the position of nucleotide X can be any of three sequences. Thus, the probability for this case becomes $18pq^2$. Similarly, the probability for case (2) becomes $6q^3$. Thus,

$$U_5 = 18pq^2 + 6q^3. \quad (7)$$

Substituting $U_2 - U_5$ given by equations (6a), (6b), and (7) into equation (5), we obtain

$$E(3) = 3p^2q + 15pq^2 + 10q^3$$
$$= (1 - \alpha)(7 + 7\alpha - 2\alpha^2)/16. \quad (8)$$

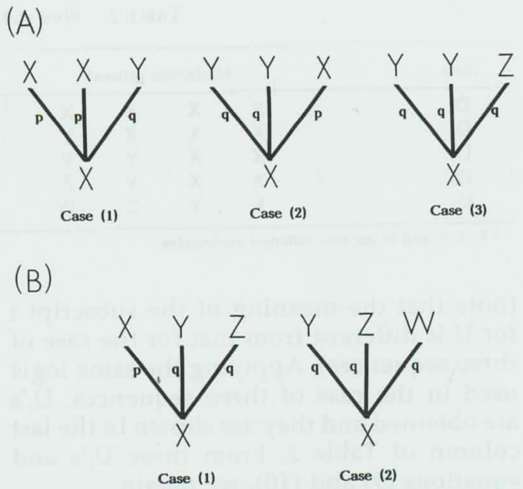

FIG. 2. (A) Three cases for configuration C_2 for three sequences. (B) Two cases for configuration C_5 for three sequences. Three sequences in each tree are in the order of A, B, and C from left to right (see Table 1). p and q are probabilities given in equations (1a) and (1b), respectively.

Four sequences.—There are fifteen nucleotide configurations for four sequences (see equation (2)), but they can be classified into five states (D_1–D_5; see Table 2), if we combine configurations with the same nucleotide pattern. For example, nucleotide pattern (XXYY) of Table 2 includes configurations [XXYY], [XYXY], and [XYYX], in which the order of sequences is specified. This simplification is possible because we are considering topologies with no hierarchical structure, and the label of species can be ignored. Let us define V_i as the probability for state D_i. The required number of nucleotide substitutions for each state under the maximum parsimony procedure is 0 for D_1, 1 for D_2, 2 for D_3 and D_4, and 3 for D_5. Thus,

$$E(4) = (V_2 + 2V_3 + 2V_4 + 3V_5)/4. \quad (9)$$

V_i is given by

$$V_i = n_i U_i, \quad (10)$$

where n_i is the number of configurations for the i-th state and U_i is the probability of having a configuration in the i-th state

TABLE 2. Five states for four sequences.

State	Nucleotide pattern[a]				Number of configurations	U_i
D_1	X	X	X	X	1	$p^4 + 3q^4$
D_2	X	X	X	Y	4	$3p^3q + 3pq^3 + 6q^4$
D_3	X	X	Y	Y	3	$6p^2q^2 + 6q^4$
D_4	X	X	Y	Z	6	$6p^2q^2 + 12pq^3 + 6q^4$
D_5	X	Y	Z	W	1	$24pq^3$

[a] X, Y, Z, and W are four different nucleotides.

(note that the meaning of the subscript i for U is different from that for the case of three sequences). Applying the same logic used in the case of three sequences, U_i's are obtained and they are shown in the last column of Table 2. From these U_i's and equations (9) and (10), we obtain

$$E(4) = 3(p^3q + 9p^2q^2 + 19pq^3 + 11q^4) = 3(1 - \alpha)(1 + \alpha)(5 - \alpha^2)/32. \quad (11)$$

Five sequences.—There are 51 configurations for five sequences, and these can be classified into six states (see Table 3). Then,

$$E(5) = (V_2 + 2V_3 + 2V_4 + 3V_5 + 3V_6)/5. \quad (12)$$

Using equation (10) and U_i's in Table 3, V_i's are obtained and they are substituted into (12).

$$E(5) = 3(p^4q + 12p^3q^2 + 52p^2q^3 + 71pq^4 + 36q^5) = 3(1 - \alpha)(43 + 43\alpha - 11\alpha^2 - 23\alpha^3 + 12\alpha^4)/256. \quad (13)$$

Six sequences.—There are 187 nucleotide configurations for six sequences, and they can be classified into nine states (see Table 4). Then, as in the case of five sequences,

$$E(6) = (V_2 + 2V_3 + 2V_4 + 3V_5 + 3V_6 + 3V_7 + 4V_8 + 4V_9)/6 = 3(p^5q + 15p^4q^2 + 90p^3q^3 + 245p^2q^4 + 261pq^5 + 112q^6) = 3(1 - \alpha)(181 + 181\alpha - 74\alpha^2 - 114\alpha^3 + 101\alpha^4 - 19\alpha^5)/1,024. \quad (14)$$

Infinite number of sequences.—As the upper limit, we consider the case of the infinite number of sequences. In this case, the procedure for obtaining E(∞) is quite simple. Since we assume that an infinite number of nucleotide sequences is available, the ancestral sequence for each site can be determined unambiguously. Thus, we only need to compare the nucleotide of an extant sequence and the ancestral one. One substitution is required if these are different and no substitutions are required if these are identical. Unless the time T is infinite, $p > q$ from equations (1a) and (1b). This implies that the ancestral nucleotide at a site should be the most frequent one among the present nucleotides. Hence,

TABLE 3. Six states for five sequences.

State	Nucleotide pattern[a]					Number of configurations	U_i
D_1	X	X	X	X	X	1	$p^5 + 3q^5$
D_2	X	X	X	X	Y	5	$3p^4q + 3pq^4 + 6q^5$
D_3	X	X	X	Y	Y	10	$3p^3q^2 + 3p^2q^3 + 6q^5$
D_4	X	X	X	Y	Z	10	$6p^3q^2 + 12pq^4 + 6q^5$
D_5	X	X	Y	Y	Z	15	$12p^2q^3 + 6pq^4 + 6q^5$
D_6	X	X	Y	Z	W	10	$6p^2q^3 + 18pq^4$

[a] X, Y, Z, and W are four different nucleotides.

TABLE 4. Nine states for six sequences.

State	Nucleotide pattern[a]	n_i[b]	U_i
D_1	X X X X X X	1	$p^6 + 3q^6$
D_2	X X X X X Y	6	$3p^5q + 3pq^5 + 6q^6$
D_3	X X X X Y Y	15	$3p^4q^2 + 3p^2q^4 + 6q^6$
D_4	X X X X Y Z	15	$6p^4q^2 + 12pq^5 + 6q^6$
D_5	X X X Y Y Y	10	$6p^3q^3 + 6q^6$
D_6	X X X Y Y Z	60	$6p^3q^3 + 6p^2q^4 + 6pq^5 + 6q^6$
D_7	X X X Y Z W	20	$6p^3q^3 + 18pq^5$
D_8	X X Y Y Z Z	15	$18p^2q^4 + 6q^6$
D_9	X X Y Y Z W	45	$12p^2q^4 + 12pq^5$

[a] X, Y, Z, and W are four different nucleotides.
[b] Number of configurations for the i-th state.

$$E(\infty) = 3q = 3(1 - \alpha)/4. \quad (15)$$

It may be interesting to note that even when the divergence time T becomes infinite, $E(\infty)$ approaches only 0.75, since α $[=\exp(-4\lambda T/3)]$ becomes zero in this case.

Comparison of E(n).—In Table 5, E(2), E(3), E(4), E(5), E(6), and $E(\infty)$ are presented for various values of λT's. These are obtained from equations (4), (8), (11), (13), (14) and (15), respectively, and the same α value, that given by $\exp(-4\lambda T/3)$, is used. For all λT values, E(n) increases as the number of sequences (n) increases. The reason for this is that the determination of the ancestral nucleotide X becomes more and more unambiguous as n is increased, hence more parallel changes are detectable. When λT is small, it is apparent that E(n) quickly approaches $E(\infty)$ as the number of sequences (n) increases. When λT is 0.10, the difference between $E(\infty)$ and λT remains small (the relative difference is 6.4%). However, $E(\infty)$ underestimates λT by more than 10% when the true value is 0.2 or greater. When λT is 1.0, the maximum parsimony procedure gives a gross underestimate of the branch lengths; $E(\infty)$ is only slightly larger than half of the expected length.

DISCUSSION

In this study trees with no hierarchical structure are assumed. In reality, some nucleotide sequences are more closely related than others. For example, sequences 1 and 2 of tree A of Figure 3 are more closely related than the other two sequences. In this case, the behavior of tree A in terms of the estimation of total branch lengths can be approximated as tree B in which sequences 1 and 2 are identical. On the other hand, if we consider tree C where sequences 1 and 2 are assumed to evolve from the same ancestral sequence, the total number of nucleotide substitutions estimated for tree C should be larger than that for tree A. Therefore, if we let 2λT (the expected distance between any two sequences in the tree of Fig. 1) be the largest pairwise distance among all the observed

TABLE 5. Comparison of E(n) for various λT values.

	λT					
n	0.01	0.05	0.10	0.20	0.50	1.00
2	0.00987	0.0468	0.0878	0.1550	0.2762	0.3489
3	0.00990	0.0476	0.0905	0.1637	0.3061	0.4007
4	0.00993	0.0483	0.0929	0.1710	0.3270	0.4301
5	0.00993	0.0483	0.0934	0.1736	0.3416	0.4592
6	0.00993	0.0484	0.0936	0.1747	0.3497	0.4788
∞	0.00993	0.0484	0.0936	0.1756	0.3649	0.5523
λT − E(∞)	0.00007	0.0016	0.0064	0.0245	0.1351	0.4477
dif[a]	0.7	3.3	6.4	12.2	27.0	44.8

[a] Ratio of $[\lambda T - E(\infty)]/\lambda T$ in percentage.

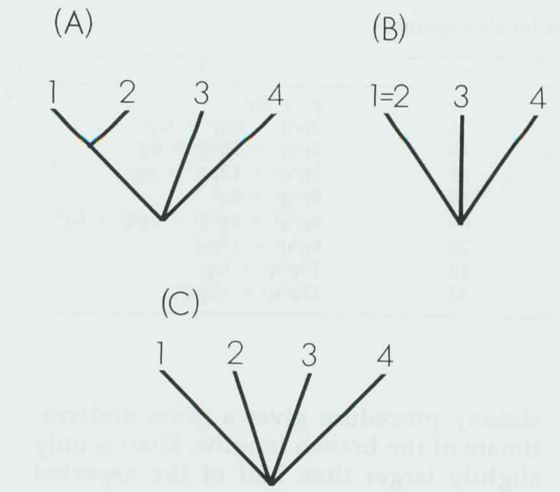

FIG. 3. A relationship of three topologies for four sequences.

distances, the results derived in this study can be considered as the upper limit of the degree of underestimation of the total number of nucleotide substitutions for a given number of sequences.

There are three types of hidden nucleotide changes which may or may not be detected by the maximum parsimony method: (1) parallel changes, (2) backward changes, and (3) successive changes. If there are only two characters such as + and −, types (1) and (2) are sufficient. In the case of nucleotide sequences, however, we need to consider case (3) because there are four nucleotides. For example, changes like A → G → T belong to this case. In this connection, it should be noted that parallel changes alone can be detected with the present model tree of no hierarchical structure. To detect changes of the other two types, we need a tree of hierarchical structure. In that case the detectability of these hidden changes depends on the number of branching events between two sequences compared in a given topology. Recognizing this property, Fitch and Brushi (1987) recently presented a new method for correcting branch lengths estimated by the maximum parsimony method. The analytical method presented in this study can be used for studying the reliability of their method.

ACKNOWLEDGMENTS

I thank Drs. J. W. Archie, D. P. Faith, M. D. Hendy, and an anonymous reviewer for their helpful comments. This study was supported from a grant-in-aid (No. 62790190) from the Ministry of Education, Science, and Culture, Japan, to N.S.

REFERENCES

CAMIN, J. H., AND R. R. SOKAL. 1965. A method for deducing branching sequences in phylogeny. Evolution, 19:311–326.

FELSENSTEIN, J. 1978. The number of evolutionary trees. Syst. Zool., 27:27–33.

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

FITCH, W. M. 1977. On the problem of discovering the most parsimonious tree. Amer. Natur., 111:223–257.

FITCH, W. M., AND M. BRUSHI. 1987. The evolution of prokaryotic ferredoxins—With a general method correcting for unobserved substitutions in less branched lineages. Mol. Biol. Evol., 4:381–394.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism, Volume 3 (H. N. Munro, ed.). Academic Press, New York.

SAITOU, N., AND M. NEI. 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J. Mol. Evol., 24:189–204.

Received 6 November 1987; accepted 13 June 1988.