A Simple Method for Estimating and Testing Minimum-Evolution Trees[1]

Andrey Rzhetsky and Masatoshi Nei

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

A simple method for estimating and testing phylogenetic trees under the principle of minimum evolution (ME) is presented. The basic procedure of this method is first to obtain the neighbor-joining (NJ) tree by Saitou and Nei's method and then to search for a tree with the minimum value of the sum (S) of branch lengths by examining all trees that are closely related to the NJ tree. Once the ME tree is identified, a statistical test is conducted for the difference in S between this tree and other closely related trees. The mathematical method required for conducting this test is developed by using the least-squares approach. Computer simulation has shown that this method identifies the correct tree with a high probability, as long as the number of nucleotides examined is sufficiently large. It has also been shown that the topology of the NJ tree is almost always identical with that of the ME tree. A method for obtaining least-squares estimates (and their standard errors) of branch lengths for a given topology is also presented. This method can be used for testing the reliability of the branching pattern of the ME tree. However, the statistical test of S values is more powerful in rejecting incorrect trees than is the branch-length test or bootstrapping. Furthermore, both a mathematical method for computing the number of trees with a given value of topological difference from the NJ tree and a computer algorithm for identifying all the topologies are developed.

Introduction

The neighbor-joining (NJ) method of phylogenetic inference (Saitou and Nei 1987) seems to be quite efficient in obtaining the correct tree, compared with maximum parsimony and several other methods (Saitou and Nei 1987; Sourdis and Nei 1988; Saitou and Imanishi 1989). In this method pairwise distances with multiple-hit corrections are used, and a tree is constructed by the principle of minimum evolution (ME), i.e., searching for the tree topology that gives the minimum sum of branch lengths. The ME principle used here is different from that of Cavalli-Sforza and Edwards (1967), whose purpose was to construct a Steiner tree. Unlike Saitou and Imanishi's ME method, however, it does not compute the sum (S) of branch lengths for all possible topologies. Instead, examination of different topologies is imbedded in the algorithm, so that only one final tree is produced. Conducting a computer simulation, Saitou and Imanishi (1989) showed that the tree obtained by this method is almost always identical with the ME tree obtained by Saitou and Imanishi's procedure.

Nevertheless, there is some chance that the NJ tree is different from the ME tree.

[1] Key words: neighbor-joining method, minimum-evolution method, estimates of branch lengths, test of topological difference.

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

Mol. Biol. Evol. 9(5):945–967. 1992.

0737-4038/92/0905-0014$02.00

Furthermore, some investigators are interested in comparing the S value for the NJ tree with the S values for other tree topologies. If the NJ tree gives an S value that is significantly smaller than that for any other topology, we can assume that the NJ tree is better than other trees (see Discussion). By contrast, if there is a tree that shows a smaller S value than that for the NJ tree, we should accept it as the most probable tree. Therefore, it is important to know the S values for alternative topologies.

When the number of DNA sequences (n) examined is small ($n \leq 6$), it is relatively easy to examine all possible topologies. For a large n, however, this method requires a prohibitive amount of computer time. For this reason, Nei (1991) suggested that S be computed only for those trees whose topological distance (d_T) from the NJ tree, as measured by Robinson and Foulds's (1981) method, is equal to 2. Since the number of such trees is small compared with the total number of possible trees, this procedure will greatly facilitate the computation and statistical test of S values for most likely trees. Actually, even if we include the trees with $d_T = 4$, the computation is not necessarily burdensome.

The purpose of this paper is to implement this idea and present the statistical methods required, using a new procedure for obtaining S values for different tree topologies. We also present results of a computer simulation to examine the validity of our statistical methods.

Algorithm

The algorithm we present here is as follows: (1) Construct an NJ tree by using Saitou and Nei's (1987) procedure. (2) Obtain all trees whose topological distance from the NJ tree is $d_T = 2$ or 4. (This procedure may be modified as will be mentioned in Discussion.) (3) Estimate S for all these topologies and compute $D = S - S_{NJ}$ and the standard error of D for each tree. (S_{NJ} is the S value for the NJ tree.) (4) If D is significantly greater than 0 for each tree, adopt the NJ tree as the most probable tree. However, if there are alternative trees whose S values are not significantly different from S_{NJ}, they must be regarded as potentially correct trees. Furthermore, if there is any tree whose S is smaller than S_{NJ} ($D < 0$), we must consider it seriously. (5) Estimate branch lengths and their standard errors, to examine the reliability of branch length estimates. In the following we will show how the above algorithm can be implemented, starting with procedure (2).

Finding Topologies with $d_T = 2$ or 4

Penny and Hendy (1985) proposed a topological distance called "partition distance," which is equivalent to Robinson and Foulds's (1981) distance but is simpler to compute. For unrooted bifurcating trees this distance is twice the number of different ways of partitioning sequences between two different trees. (Partitioning of sequences is done only at interior branches.) As an example, consider trees A and B in figure 1. Both trees are for eight sequences and have five interior branches. It is possible to cut the tree at any interior branch and divide the sequences into two groups. Cutting at some interior branches results in the same partition of sequences in trees A and B but not at other branches. For example, a cut at branch a produces two sequence groups—(1,2) and (3,4,5,6,7,8)—in both trees. A cut at branch c, however, produces different partitions in trees A and B. That is, the two groups produced by this cut are (1,2,3,4) and (5,6,7,8) in tree A but are (1,2,3,5) and (4,6,7,8) in tree B. Similarly, a cut at branch d produces different partitions in the two trees. In the present example, only these two cuts result in different partitions. Therefore, the topological distance between

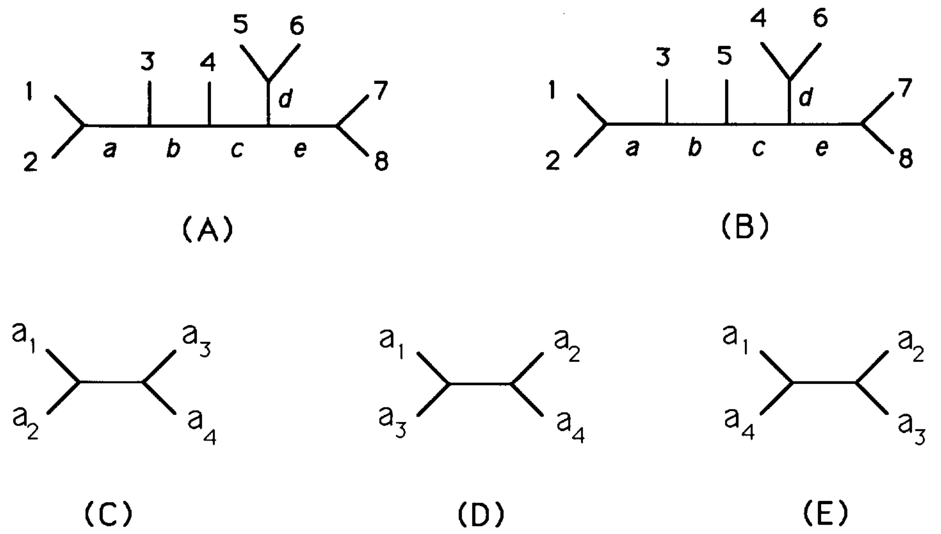

FIG. 1.—Hypothetical trees illustrating various computations discussed in text

the two trees is $d_T = 2 \times 2 = 4$. If the two trees have the same topology, $d_T = 0$, and if all interior branches produce different partitions, $d_T = 10$ in this example. The general formula for d_T for a pair of arbitrary trees for n sequences is given by

$$d_T = 2[\min(q_1, q_2) - p] + |q_1 - q_2| , \qquad (1)$$

where q_1 and q_2 are the total numbers of partitions (interior branches) for trees 1 and 2, respectively, and p is the number of partitions that are identical for the two trees. q_1 and q_2 may not be the same when multifurcating trees are involved. For bifurcating trees, however, q_1 and q_2 are always the same, and d_T takes only even values. In general, unrooted bifurcating trees for n sequences have $n - 3$ interior branches, so that the maximum possible value of d_T for these trees is $2(n-3)$.

Let us now consider a tree for n sequences and derive a formula for the number of trees whose topological distance from a given tree is $d_T = 2$. We first consider the simplest case of $n = 4$ (trees C, D, and E in fig. 1). In this case, there are three different topologies, and the topological distance of tree D or E from C is $d_T = 2$. Therefore, the number of trees whose topological distance from tree C is $d_T = 2$ is 2. In the following we denote this number by $f(d_T = 2)$.

In the case of n sequences $f(d_T = 2)$ can be computed by considering each interior branch separately. As an example, consider tree A in figure 1, which has five interior branches. We first note that any interior branch is connected to four groups of sequences. For example, branch a in tree A of figure 1 is connected to sequences 1, 2, 3 and to the group of sequences 4, 5, 6, 7, and 8. If we denote these four groups of sequences by a_1, a_2, a_3, and a_4, respectively, this tree can be represented by tree C in figure 1. Therefore, branch a of tree A in figure 1 generates two different topologies with distance $d_T = 2$ (trees D and E in fig. 1). In the following we denote trees C, D, and E by $[(a_1,a_2)(a_3,a_4)]$, $[(a_1,a_3)(a_2,a_4)]$, and $[(a_1,a_4)(a_2,a_3)]$, respectively. The same procedure can be used for any interior branch to generate two different topologies with $d_T = 2$. As mentioned above, an unrooted bifurcating tree for n sequences has $n - 3$ interior branches. Thus, we have the following general formula:

$$f(d_{\mathrm{T}}=2) = 2(n-3)\,. \tag{2}$$

The derivation of the formula for the number of trees whose distance from a given tree is $d_{\mathrm{T}} = 4$ is slightly more complicated (Appendix A), but the final result is rather simple. It is given by

$$f(d_{\mathrm{T}}=4) = 2(n^2-4n+3n'-6)\,, \tag{3}$$

where n' is the number of tree nodes that are connected to one interior branch and two exterior branches ($n \geq 4$). For example, in tree A in figure 1, $n = 8$ and $n' = 3$, so $f(d_{\mathrm{T}}=4) = 70$. This number is fairly large but is a small proportion of the total number of possible topologies (10,395).

To execute procedure (2) in our algorithm, it is necessary not only to know $f(d_{\mathrm{T}}=2)$ and $f(d_{\mathrm{T}}=4)$ but also to identify all topologies with $d_{\mathrm{T}} = 2$ or 4. An algorithm for identifying these topologies is presented in Appendix B.

Sum of Branch Lengths for a Given Tree Topology

Once all topologies with $d_{\mathrm{T}} = 2$ or 4 are identified, we must compute S for each of the topologies. Saitou and Imanishi (1989) used Fitch and Margoliash's (1967) algorithm to estimate this value, but this method is not statistically very efficient. So, in the present paper we shall use the least-squares method. [The NJ method gives least-squares estimates of branch lengths when $n \leq 5$ (Saitou and Nei 1987).]

Let us consider a hypothetical tree for five sequences given in tree A of figure 2 and use the least-squares method to estimate the branch lengths of the tree denoted by $b_1, b_2, \ldots,$ and b_7. We represent an unbiased estimate of the evolutionary distance between sequences i and j by d_{ij}. We can than write d_{ij}'s as follows:

$$\begin{aligned}
d_{12} &= b_1 + b_2 + e_{12}\,,\\
d_{13} &= b_1 + b_3 + b_6 + e_{13}\,,\\
d_{14} &= b_1 + b_4 + b_6 + b_7 + e_{14}\,,\\
d_{15} &= b_1 + b_5 + b_6 + b_7 + e_{15}\,,\\
d_{23} &= b_2 + b_3 + b_6 + e_{23}\,,\\
d_{24} &= b_2 + b_4 + b_6 + b_7 + e_{24}\,,\\
d_{25} &= b_2 + b_5 + b_6 + b_7 + e_{25}\,,\\
d_{34} &= b_3 + b_4 + b_7 + e_{34}\,,\\
d_{35} &= b_3 + b_5 + b_7 + e_{35}\,,\\
d_{45} &= b_4 + b_5 + e_{45}\,,
\end{aligned}$$

where e_{ij}'s are sampling errors. We assume that e_{ij} is distributed with mean 0 and variance $V(d_{ij})$. If we use matrix algebra, the above set of equations may be written as

$$\boldsymbol{d} = \boldsymbol{A}\boldsymbol{b} + \boldsymbol{e}\,, \tag{4}$$

where $\boldsymbol{d}$, $\boldsymbol{b}$, and $\boldsymbol{e}$ are the column vectors of d_{ij}'s, b_i's, and e_{ij}'s, respectively, i.e., $\boldsymbol{d}^t = (d_{12}, d_{13}, \ldots, d_{45})$, $\boldsymbol{b}^t = (b_1, b_2, \ldots, b_7)$, and $\boldsymbol{e}^t = (e_{12}, e_{13}, \ldots, e_{45})$. Here t indicates the transpose of vectors or matrices. $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}, \tag{5}$$

The least-squares estimate of $\boldsymbol{b}$ is then given by

$$\hat{\boldsymbol{b}} = (\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{A}^t\boldsymbol{d} = \boldsymbol{L}\boldsymbol{d}\,, \tag{6}$$

where $\boldsymbol{L} = (\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{A}^t$. Obviously, an estimate of the length of the ith branch is

$$\hat{b}_{\mathrm{i}} = \boldsymbol{L}_i\boldsymbol{d}\,, \tag{7}$$

where $\boldsymbol{L}_i$ is the ith row of the matrix L. The variance of $\hat{b}_{\mathrm{i}}$ is then given by

$$V(\hat{b}_{\mathrm{i}}) = \boldsymbol{L}_i\boldsymbol{V}\boldsymbol{L}_i^t\,, \tag{8}$$

where $\boldsymbol{V}$ is the variance and covariance matrix of distances, which will be discussed in detail later. Once the branch lengths are estimated, we can easily compute the S of all branch lengths by

$$S = \sum_{i=1}^{T} \hat{b}_i\,, \tag{9}$$

where $\hat{b}_{\mathrm{i}}$ is the estimate of branch length b_i, and T is the total number of branches.

The above approach can be extended to the case of any number of sequences. When there are n sequences, T in an unrooted bifurcating tree is $2n - 3$, and the total number of pairwise distances (R) is $n(n-1)/2$. In the following, d_{ij} will be renumbered from d_1 to d_R, depending on the convenience.

For obtaining S, however, it is not necessary to estimate all branch lengths. Since S is a linear function of d_{ij}'s, it can be obtained without estimating b_i's. That is,

$$S = \boldsymbol{yd}\,, \tag{10}$$

where $\boldsymbol{y}$ is a row vector of the coefficients of d_i's, i.e., $\boldsymbol{y} = (y_1, y_2, \ldots, y_R)$. The vector y can be obtained by

$$\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{u}\,, \tag{11}$$

where $\boldsymbol{u}$ is a unit column vector of $2n - 3$ elements, i.e., $\boldsymbol{u}^t = (1, 1, \ldots, 1)$. Therefore, the least-squares estimate of S is

$$S = \boldsymbol{yd} = \boldsymbol{A}(\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{ud} = \sum_{i=1}^{R} y_i d_i\,. \tag{12}$$

In the case of tree A in figure 2, $\boldsymbol{y}$ becomes

$$\boldsymbol{y}_{\mathrm{A}} = (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{2})\,. \tag{13}$$

So, S is given by

$$\begin{aligned} S_{\mathrm{A}} = {} & d_{12}/2 + d_{13}/4 + d_{14}/8 + d_{15}/8 + d_{23}/4 \\ & + d_{24}/8 + d_{25}/8 + d_{34}/4 + d_{35}/4 + d_{45}/2\,. \end{aligned} \tag{14}$$

Similarly, for tree B in figure 2, we have

$$\boldsymbol{y}_{\mathrm{B}} = (\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{8}, \tfrac{1}{8}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8}, \tfrac{1}{2})\,, \tag{15}$$

$$\begin{aligned} S_{\mathrm{B}} = {} & d_{12}/4 + d_{13}/2 + d_{14}/8 + d_{15}/8 + d_{23}/4 \\ & + d_{24}/4 + d_{25}/4 + d_{34}/8 + d_{35}/8 + d_{45}/2\,. \end{aligned} \tag{16}$$

Test of the Null Hypothesis $D \equiv S_{\mathrm{B}} - S_{\mathrm{A}} = 0$

As mentioned earlier, we are primarily interested in the difference in S between two topologies. This difference can be written as

$$D = S_{\mathrm{B}} - S_{\mathrm{A}} = (\boldsymbol{y}_{\mathrm{B}} - \boldsymbol{y}_{\mathrm{A}})\boldsymbol{d} = \sum_{i=1}^{R} (y_{\mathrm{B}i} - y_{\mathrm{A}i})d_i\,. \tag{17}$$

Therefore, if $\boldsymbol{y}_{\mathrm{A}}$ and $\boldsymbol{y}_{\mathrm{B}}$ are computed for a pair of topologies, D can easily be obtained.

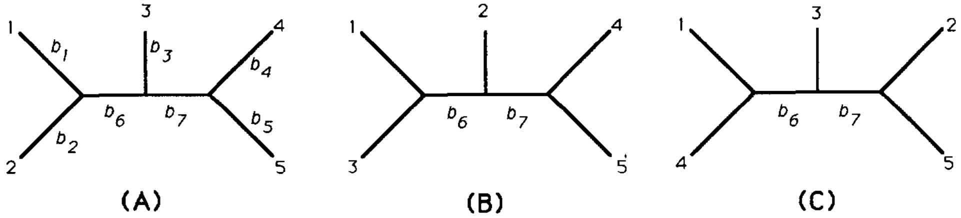

FIG. 2.—Hypothetical trees illustrating computation of branch lengths and S values

For this purpose, it is not necessary to know individual S values. In the case of trees A and B in figure 2, we know y_{Ai}'s and y_{Bi}'s, so D is given by

$$D = -d_{12}/4 + d_{13}/4 + d_{24}/8 + d_{25}/8 - d_{34}/8 - d_{35}/8\ . \qquad (18)$$

Our null hypothesis is $D = 0$, as mentioned earlier. If S_A is smaller than S_B and D is significantly greater than 0, we conclude that tree A is better than tree B. However, what is the biological meaning of this null hypothesis when two different topologies are compared? Actually, this hypothesis is equivalent to the null hypothesis that the lengths of the interior branches that produce different partitions of sequences for the two topologies are 0. For example, the test of $D = S_B - S_A = 0$ in the trees of figure 2 is actually testing the hypothesis that $b_6 = 0$. Indeed, we can show that

$$D = S_B - S_A = b_6/2 - (2e_{12} - 2e_{13} - e_{24} - e_{25} + e_{34} + e_{35})/8\ . \qquad (19)$$

Therefore, when $b_6 = 0$, the expectation of D is 0. In practice, we do not know which of trees A and B is the correct one. So, D can be positive or negative. Similarly, $D = S_C - S_A$ can be written as

$$D = S_C - S_A = 3(b_6 + b_7)/4 - 3(e_{12} - e_{14} - e_{25} + e_{45})/8\ . \qquad (20)$$

Therefore, we are testing the null hypothesis that both b_6 and b_7 in tree A in fig. 2 are equal to 0. This principle applies to any pair of bifurcating trees, irrespective of the number of sequences.

Variance of $D \equiv S_A - S_B$

To test the statistical significance of D, however, we must compute the variance of D. Nei (1991) suggested that the variance of D be computed by the bootstrap method (Efron 1982, chap. 5). However, the variance of D, $V(D)$, can be obtained analytically. That is,

$$\begin{aligned} V(D) &= (\mathbf{y}_B - \mathbf{y}_A)^t \mathbf{V} (\mathbf{y}_B - \mathbf{y}_A) \\ &= \sum_{i=1}^{R} (y_{Bi} - y_{Ai})^2 V(d_i) + 2 \sum_{j>i}^{R} (y_{Bi} - y_{Ai})(y_{Bj} - y_{Aj}) \mathrm{Cov}(d_i, d_j)\ . \end{aligned} \qquad (21)$$

Here, $\mathbf{V}$ stands for the variance-covariance matrix of d_{ij}'s. In the case of trees A and B in figure 2, $V(D)$ can be written as

$$\begin{aligned} V(D) = {} & V(d_{12})/16 + V(d_{13})/16 + V(d_{24})/64 + V(d_{25})/64 \\ & + V(d_{34})/64 + V(d_{35})/64 - \mathrm{Cov}(d_{12},d_{13})/8 - \mathrm{Cov}(d_{12},d_{24})/16 \\ & - \mathrm{Cov}(d_{12},d_{25})/16 + \mathrm{Cov}(d_{12},d_{34})/16 + \mathrm{Cov}(d_{12},d_{35})/16 \\ & + \mathrm{Cov}(d_{13},d_{24})/16 + \mathrm{Cov}(d_{13},d_{25})/16 - \mathrm{Cov}(d_{13},d_{34})/16 \qquad (22) \\ & - \mathrm{Cov}(d_{13},d_{35})/16 + \mathrm{Cov}(d_{24},d_{25})/32 - \mathrm{Cov}(d_{24},d_{34})/32 \\ & - \mathrm{Cov}(d_{24},d_{35})/32 - \mathrm{Cov}(d_{25},d_{34})/32 \\ & - \mathrm{Cov}(d_{25},d_{35})/32 + \mathrm{Cov}(d_{34},d_{35})/32\ . \end{aligned}$$

There is no difficulty in evaluating $V(d_{ij})$. Most methods for estimating d_{ij} from sequence data give an approximate formula for $V(d_{ij})$ (e.g., see Kimura and Ohta 1972; Kimura 1980; Tajima and Nei 1984). However, the computation of $\mathrm{Cov}(d_{ij}, d_{lk})$ is more complicated.

A number of authors (e.g., Nei et al. 1985; Bulmer 1989; Nei and Jin 1989) have proposed use of the formula

$$\mathrm{Cov}(d_{ij},d_{kl}) = V(d'_{ij,kl}) , \qquad (23)$$

where $d'_{ij,kl}$ is the length of branches that are shared by the path connecting sequences i and j and the path connecting sequences k and l. Hence, the estimation of $\mathrm{Cov}(d_{ij}, d_{kl})$ depends on tree topology. Since two different tree topologies are involved in our case, this approach is inapplicable. We therefore use the method suggested by Bulmer (1991).

Let d_{ij} be the distance between species i and j, and let d_{kl} be the distance between species k and l. The correlation coefficient between d_{ij} and d_{kl} may be estimated by

$$r_{ij,kl} = \frac{(p_{ij,kl} - p_{ij}p_{kl})}{[(1-p_{ij})p_{ij}(1-p_{kl})p_{kl}]^{1/2}} , \qquad (24)$$

where $p_{ij,kl}$ is the proportion of sites at which sequence i differs from sequence j and sequence k differs from sequence l, and where p_{ij} is the proportion of sites at which sequence i differs from sequence j. Therefore, $\mathrm{Cov}(d_{ij},d_{kl})$ can be estimated by

$$\mathrm{Cov}(d_{ij},d_{kl}) = r_{ij,kl}s(d_{ij})s(d_{kl}) , \qquad (25)$$

where $s(d_{ij})$ and $s(d_{kl})$ are the standard deviations of d_{ij} and d_{kl}, respectively. In the case of Jukes and Cantor's (1969) one-parameter model the evolutionary distance between a pair of sequences is estimated by

$$d_{ij} = -b \ln(1-p_{ij}/b) , \qquad (26)$$

where $b = 3/4$, and the variance of d_{ij} is given by

$$V(d_{ij}) = \frac{p_{ij}(1-p_{ij})}{m(1-p_{ij}/b)^2} \qquad (27)$$

(Kimura and Ohta 1972). Here, m stands for the number of nucleotides examined. The covariance of d_{ij} and d_{kl} is therefore given by

$$\mathrm{Cov}(d_{ij},d_{kl}) = \frac{(p_{ij,kl} - p_{ij}p_{kl})}{m(1-p_{ij}/b)(1-p_{kl}/b)} \qquad (28)$$

(Bulmer 1991). The formula can also be used for amino acid sequence data by putting $b = 19/20$. In the case of Kimura's (1980) two-parameter model, we have

$$d_{ij} = -\tfrac{1}{2}\ln[(1-2\mathrm{P}_{ij}-\mathrm{Q}_{ij})\sqrt{1-2\mathrm{Q}_{ij}}]\,, \tag{29}$$

$$\mathrm{Var}(d_{ij}) = (1/m)[(a_{ij}^2\mathrm{P}_{ij}+b_{ij}^2\mathrm{Q}_{ij})-(a_{ij}\mathrm{P}_{ij}+b_{ij}\mathrm{Q}_{ij})^2]\,, \tag{30}$$

$$\begin{aligned}\mathrm{Cov}(d_{ij},d_{kl}) = {} & \frac{(p_{ij,kl}-p_{ij}p_{kl})}{m[(1-p_{ij})p_{ij}(1-p_{kl})p_{kl}]^{1/2}} \\ & \times [(a_{ij}^2\mathrm{P}_{ij}+b_{ij}^2\mathrm{Q}_{ij})-(a_{ij}\mathrm{P}_{ij}+b_{ij}\mathrm{Q}_{ij})^2]^{1/2} \\ & \times [(a_{kl}^2\mathrm{P}_{kl}+b_{kl}^2\mathrm{Q}_{kl})-(a_{kl}\mathrm{P}_{kl}+b_{kl}\mathrm{Q}_{kl})^2]^{1/2}\,,\end{aligned} \tag{31}$$

where P_{ij} is the proportion of transitional differences between species i and j, Q_{ij} is the proportion of transversional differences between species i and j, $a_{ij} = (1-2\mathrm{P}_{ij}-\mathrm{Q}_{ij})^{-1}$, and $b_{ij} = (\tfrac{1}{2})(1-2\mathrm{P}_{ij}-\mathrm{Q}_{ij})^{-1}+(\tfrac{1}{2})(1-2\mathrm{Q}_{ij})^{-1}$. (See note added in proof.)

Computer Simulation

The method of obtaining the ME tree presented in this paper depends on a number of assumptions. In particular, it assumes that the ME tree is included in the group of trees whose topological distance d_{T} from the NJ tree is 2 or 4 and that the ME tree is likely to be the correct tree. (Of course, when the number of sequences is small, it is possible to examine all possible trees.) Furthermore, the resolving power of our test of $D = S-S_{\mathrm{NJ}}$ is dependent not only on the lengths of the interior branches involved but also on the error terms e_{ij}'s [eqq. (19) and (20)]. We therefore examined the validity of the assumption and the resolving power of our phylogenetic test by using computer simulation.

Probability of Obtaining the True Tree among the Neighboring Trees of the NJ Tree

Our first simulation was conducted to see how close the NJ tree is, compared with the true tree. We set up a model tree and simulated nucleotide substitution according to this model tree. The method of the simulation of nucleotide substitution was the same as that of Sourdis and Nei (1988), but we considered only the Jukes-Cantor model of substitution. The model trees considered were the same as those of Saitou and Imanishi (1989) and consisted of six DNA sequences (fig. 3) with various topologies and branch length. (We also considered several other trees with 10 sequences, but, since the results obtained were essentially the same as those for fig. 3, we shall not present them.) Nucleotide substitution was assumed to be stochastic, so that the sequences generated for a given model tree varied from replication to replication. Once a set of six extant sequences were generated for a given tree, a tree was reconstructed by the NJ method, and this tree's topological distance from the true tree was computed. This was repeated 10,000 times for each tree. In trees A and B in figure 3 a constant rate of evolution was assumed for all sequences, whereas in trees C and D in figure 3 the rate of nucleotide substitution varied with sequence. In these trees the branch lengths are proportional to the expected number of substitutions per site. In the case of constant rate of substitution (i.e., trees A and B), the expected number of substitutions per site, from the ancestral sequence to the extant sequence, is denoted by U, and the length of each branch is expressed as multiples of a, which is related to U as specified in each figure. We examined two different values of U, namely, 0.05

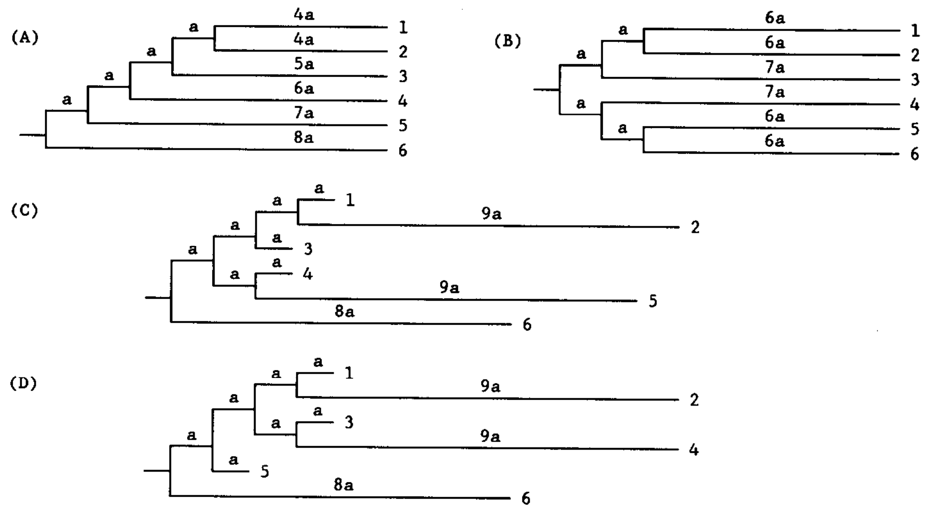

FIG. 3.—Four model trees used for computer simulation. These trees are identical with those used by Saitou and Imanishi (1989).

and 0.5. In the case of varying rate of evolution (i.e., trees C and D), $a = 0.01$ and $a = 0.05$ were used. These values are identical with those used by Saitou and Imanishi (1989). In each of the four trees, we considered five different numbers of nucleotides examined, *m*, i.e., 300, 600, 900, 1,200, and 2,400.

Table 1 shows the relative frequencies of the NJ trees differing from the true tree by $d_T = 0$, 2, 4, and 6, for model trees A and B. Here $d_T = 0$ denotes the case where the NJ method recovered the correct tree. The frequencies in table 1 are similar to

Table 1
Frequencies of NJ Trees Differing from Correct Tree by d_T = 0, 2, 4, and 6

	FREQUENCY OF (%)							
	Tree A: d_T =				Tree B: d_T =			
	0	2	4	6	0	2	4	6
U = 0.05:								
300 bp	58.1	31.7	8.7	1.5	62.9	31.7	5.1	0.3
600 bp	82.1	16.3	1.5	0.1	84.3	14.8	0.9	0
900 bp	92.6	7.0	0.4	0	92.9	6.9	0.2	0
1,200 bp	96.2	3.8	0.1	0	96.6	3.3	0.1	0
2,400 bp	99.7	0.3	0	0	99.8	0.2	0	0
U = 0.50:								
300 bp	53.6	34.7	10.3	1.3	52.8	39.2	7.9	0.1
600 bp	79.8	18.6	1.6	0.1	76.7	21.9	1.4	0
900 bp	93.0	7.0	0	0	91.0	8.8	0.2	0
1,200 bp	96.3	3.7	0	0	96.4	3.6	0	0
2,400 bp	99.6	0.4	0	0	99.7	0.3	0	0

NOTE.—d_T = 0 represents the correct topology. These results are based on 10,000 replications.

those obtained by Saitou and Imanishi (1989), though these authors examined only the cases of $m = 300$ and $m = 600$. When $m = 300$, the NJ method may choose a wrong tree, with a substantial probability. However, if we consider the trees with $d_T \leq 4$, the correct tree is included among them with a probability of 98.5%. This probability increases to nearly 100% if $m \geq 600$. Actually, in the case of $m \geq 600$, the probability of finding the correct tree among trees with $d_T \leq 2$ from the NJ tree is >0.98. The results are essentially the same both for trees A and B and for $U = 0.05$ and 0.50. The same thing can be said about trees C and D in figure 3 (table 2). Furthermore, we examined four more trees with 10 DNA sequences, as mentioned earlier. In these cases, too, the probability that the correct tree is included among the trees with $d_T \leq 2$ was very high, as long as m is large ($\geq$900). Therefore, our procedure seems to be sufficient for obtaining the correct tree when topologies with $d_T \leq 2$ are considered, as long as $m > 900$. (Of course, if a is very small, this would not be the case, and some other procedures are necessary, as will be mentioned in Discussion.)

Rank of S for the True Tree among the Neighboring Trees with $d_T = 2$

Theoretically, the true tree is expected to give the smallest S value, compared with other trees, as long as d_{ij}'s are unbiased estimates and m is large (authors' unpublished data). In practice, however, sampling errors may disturb the ranks of S values. We therefore examined the rank of the S for the true tree among the trees with $d_T \leq 2$. Table 3 shows that in both tree A and tree B the probability that the S for the true tree is smallest is ~50%–60% when 300 nucleotides are examined but that the probability rapidly increases as m increases. Yet, it is <98% even for $m = 1{,}200$. It is interesting to see that these values are virtually identical with the probabilities that the NJ tree is the correct tree in table 1. They are also similar to the probabilities that the ME tree is the correct tree (Saitou and Imanishi 1989). These results support Saitou and Imanishi's (1989) conclusion that the NJ tree is almost always identical

Table 2
Frequencies of NJ Trees Differing from Correct Tree by d_T = 0, 2, 4, and 6

	Frequency of (%)							
	Tree C: d_T =				Tree D: d_T =			
	0	2	4	6	0	2	4	6
a = 0.01:								
300 bp	74.6	21.7	3.3	0.4	75.6	19.6	4.6	0.2
600 bp	93.7	6.2	0.1	0	94.5	5.0	0.5	0
900 bp	98.1	1.9	0	0	98.7	1.3	0	0
1,200 bp	99.4	0.6	0	0	99.6	0.4	0	0
2,400 bp	100.0	0	0	0	100.0	0	0	0
a = 0.05:								
300 bp	72.3	24.0	3.3	0.4	75.8	17.5	6.5	0.1
600 bp	90.6	8.6	0.8	0	94.5	4.5	1.0	0
900 bp	97.3	2.7	0	0	99.0	0.9	0.1	0
1,200 bp	98.5	1.5	0	0	99.7	0.3	0	0
2,400 bp	100.0	0	0	0	100.0	0	0	0

NOTE.—d_T = 0 represents the correct topology. These results are based on 10,000 replications.

Table 3
Frequencies of Rankings of S Value for True Tree among Trees with $d_T \leq 2$

	Frequency of (%)									
	Tree A: Rank of S =					Tree B: Rank of S =				
	1	2	3	4	5	1	2	3	4	5
U = 0.05:										
300 bp	57	23	16	3	0	63	21	14	2	0
600 bp	83	12	5	0	0	84	12	4	0	0
900 bp	94	5	1	0	0	94	6	1	0	0
1,200 bp	97	3	0	0	0	95	4	1	0	0
2,400 bp	100	0	0	0	0	100	0	0	0	0
U = 0.5:										
300 bp	51	31	16	2	0	54	31	13	3	0
600 bp	77	18	5	0	0	77	17	5	1	0
900 bp	90	9	1	0	0	92	6	1	0	0
1,200 bp	96	4	0	0	0	97	3	0	0	0
2,400 bp	99	1	0	0	0	99	1	0	0	0

NOTE.—1, 2, 3, . . . represent the smallest, second smallest, third smallest, etc. These results are based on 1,000 replications.

with the ME tree. Table 4 gives the same quantities for trees C and D. In this case the probability that the true tree shows the smallest S is even higher than that for trees A and B. From these results, we may conclude that, if all trees whose distance from the NJ tree is $d_T = 2$ are examined, the true tree is included with a high probability, unless m and a are very small.

Resolving Power of the Statistical Test Proposed

To examine the resolving power of our phylogenetic test, we studied the frequencies of occurrence of the following four cases: (1) the S for the true tree (S_T) is minimum and is significantly (5% level) smaller than any other S. (2) S_T is smallest but is not statistically significant from the second smallest S. (3) A wrong tree shows the smallest S value, but it is not statistically significant from S_T. (4) A wrong tree shows the smallest S, which is significantly smaller than S_T.

Table 5 shows the frequencies of these four different cases for trees A and B. When $m = 300$, the frequency of case (1) is quite low, indicating the inefficiency of our statistical test. Furthermore, for $m = 300$, case (4) occurs with a frequency of 0.5%–1%. Even when m is as large as 2,400, the frequency of case (1) is not very high. Table 6 shows that the results for trees C and D are better than those for trees A and B. With these model trees the frequency of case (1) is $\sim \geq 90\%$ when 2,400 nucleotides are examined. Our test is more powerful in the case where substitution rate varies with evolutionary lineage than in the case where the molecular clock applies. In tables 1 and 2 we saw that the ME tree or the NJ tree is correct with a probability of >99% in the case of $m = 2{,}400$. Yet, the S for the correct tree is not always significantly smaller than the second smallest S. In this sense, our test is quite conservative.

Numerical Example

At the present time, the crocodilian group of reptiles is believed to be the type of organism that is most closely related to birds. However, molecular data often suggest

Table 4
Frequencies of Rankings of *S* Value for True Tree among Trees with $d_T \leq 2$

	Frequency of (%)									
	Tree C: Rank of *S* =					Tree D: Rank of *S* =				
	1	2	3	4	5	1	2	3	4	5
a = 0.01:										
300 bp	73	19	7	1	0	76	17	7	0	0
600 bp	92	7	1	0	0	97	3	0	0	0
900 bp	98	2	0	0	0	99	1	0	0	0
1,200 bp	99	1	0	0	0	100	0	0	0	0
2,400 bp	100	0	0	0	0	100	0	0	0	0
a = 0.05:										
300 bp	70	24	6	0	0	82	14	4	0	0
600 bp	90	9	1	0	0	97	3	0	0	0
900 bp	97	3	0	0	0	100	0	0	0	0
1,200 bp	99	1	0	0	0	100	0	0	0	0
2,400 bp	100	0	0	0	0	100	0	0	0	0

NOTE.—1, 2, 3, . . . represent the smallest, second smallest, third smallest, etc. These results are based on 1,000 replications.

that birds are more closely related to mammals than to reptiles. Hedges et al. (1990) recently studied this problem by using the nucleotide sequences of the 18S ribosomal RNA gene and some other data. We therefore examined this problem by using our new statistical method.

Table 5
Results of Significance Tests of $D = S\text{-}S_T$

	Frequency of (%)							
	Tree A: Case				Tree B: Case			
	1	2	3	4	1	2	3	4
U = 0.05:								
300 bp	0	56.9	42.7	0.4	0	64.8	34.7	0.5
600 bp	0.4	81.8	17.6	0.2	1.6	83.2	15.0	0.2
900 bp	3.7	89.2	6.9	0.2	8.7	83.0	8.2	0.1
1,200 bp	12.3	83.7	4.0	0	20.8	75.2	4.0	0
2,400 bp	71.3	28.7	0	0	71.2	28.6	0.2	0
U = 0.5:								
300 bp	0	52.2	47.3	0.5	0	50.3	48.8	0.9
600 bp	0.6	76.9	22.4	0.1	0.9	77.7	21.2	0.2
900 bp	5.6	85.3	9.1	0	7.2	84.4	8.3	0.1
1,200 bp	17.1	79.9	3.0	0	20.7	74.9	4.4	0
2,400 bp	58.4	41.1	0.5	0	57.9	41.2	0.9	0

NOTE.—1 = Significant minimum of the *S* for the true tree (S_T); 2 = nonsignificant minimum of S_T; 3 = nonsignificant minimum of *S* for a wrong tree; and 4 = significant minimum of *S* for a wrong tree. These results are based on 2,000 replications.

Table 6
Results of Significance Tests of $D = S\text{-}S_T$

	FREQUENCY OF (%)							
	Tree C: Case				Tree D: Case			
	1	2	3	4	1	2	3	4
a = 0.01:								
300 bp	0.1	73.4	26.2	0.3	0.1	78.1	21.5	0.3
600 bp	4.9	88.5	6.6	0	7.8	87.6	4.6	0
900 bp	23.6	73.8	2.5	0.1	31.0	67.7	1.3	0
1,200 bp	44.4	54.3	1.3	0	57.7	42.2	0.1	0
2,400 bp	95.2	4.7	0.1	0	98.6	1.4	0	0
a = 0.05:								
300 bp	0	70.7	29.2	0.1	0.2	80.7	19.1	0
600 bp	2.0	87.0	11.0	0	9.6	86.6	3.8	0
900 bp	16.1	79.6	4.3	0	42.2	57.0	0.8	0
1,200 bp	40.4	58.2	1.4	0	71.4	28.5	0.1	0
2,400 bp	89.4	10.5	0.1	0	99.3	0.7	0	0

NOTE.—1 = Significant minimum of the S for the true tree (S_T); 2 = nonsignificant minimum of S_T; 3 = nonsignificant minimum of S for a wrong tree; and 4 = significant minimum of S for a wrong tree. These results are based on 2,000 replications.

In this numerical example we will use the nucleotide sequences of the 18S ribosomal RNA gene for six species: *Homo sapiens* (mammal), *Turdus migratorius* (bird), *Heterodon platyrhinos* (snake), *Xenopus laevis* (frog), *Pseudemus scripta* (turtle), and *Alligator mississippiensis* (crocodilian). Hedges et al. have presented the entire sequences available for these species. In this analysis, however, we eliminated all deletions/insertions (gaps) and ambiguous nucleotide sites and used 1,297 sites which were shared by all the six species. We first estimated the number of nucleotide substitutions per site (d) for all pairs of sequences by using the Jukes-Cantor formula. [In the present case, most distance measures give essentially the same results, since the d is small (Nei 1987, pp. 72–73).] The results obtained are presented in table 7. Using the NJ method, we then reconstructed a phylogenetic tree, which is presented as tree A in figure 4. The branch lengths of this tree and their standard errors were estimated by the least-squares method described earlier [eqq. (6) and (8)]. The sum of all branch lengths therefore becomes 8.27×10^{-2}. For this case, the matrix $\boldsymbol{L}$ in equation (6) is given in figure 5.

Table 7
Pairwise Distances ($d \times 100$) for Six Species of Tetrapods

	Bird	Snake	Frog	Turtle	Crocodilian
Mammal	3.31	3.31	4.61	3.06	2.82
Bird		3.39	5.35	3.31	2.98
Snake			3.63	1.63	1.16
Frog				3.06	2.74
Turtle					0.46
Crocodilian					

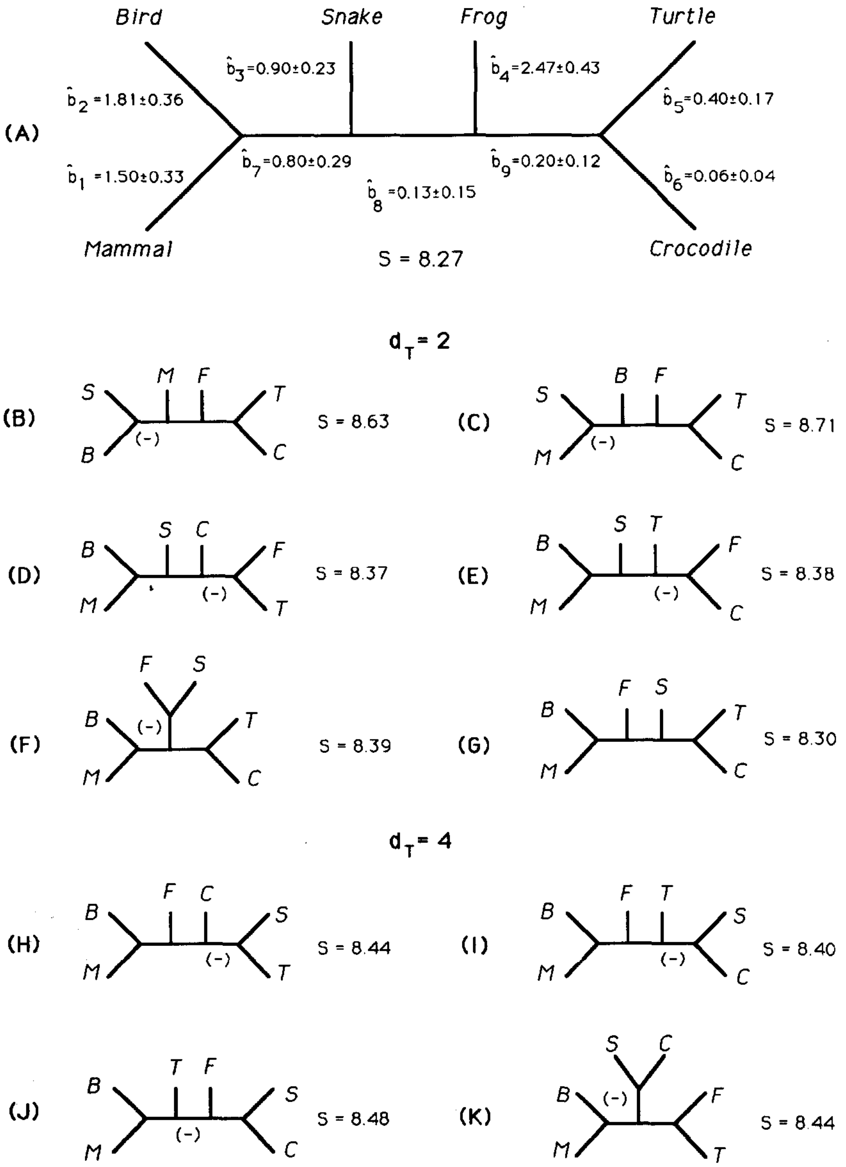

FIG. 4.—NJ tree (A) and its closely related trees (B–K) for the 18S rRNA genes from six tetrapod species [mammals (M), birds (B), snakes (S), frogs (F), turtles (T), and crocodiles (C)]. The estimates of branch lengths (i.e., $\hat{b} \times 100$) and their standard errors were obtained by the least-squares method. The estimate of the length of a branch with a (−) sign was negative. The S value in this figure represents the true value multiplied by 100.

With six species, there are 105 different topologies. Six of them have a topological distance of $d_T = 2$ from the NJ tree. These six topologies are given as trees B–G in figure 4. To compute S and $D = S{-}S_{NJ}$ for all of these topologies, we must know the vector ($\mathbf{y}$) of y coefficients [eq. (10)]. These vectors for topologies A–G can be obtained by using equation (11). For example, the vectors for trees A and G are as follows:

$$\mathbf{L} = \begin{bmatrix}
\frac{1}{2} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{4} & -\frac{1}{12} & -\frac{1}{12} & -\frac{1}{12} & \frac{1}{4} & -\frac{1}{12} & -\frac{1}{12} & -\frac{1}{12} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 \\
0 & 0 & \frac{1}{6} & -\frac{1}{12} & -\frac{1}{12} & 0 & \frac{1}{6} & -\frac{1}{12} & -\frac{1}{12} & \frac{1}{6} & -\frac{1}{12} & -\frac{1}{12} & \frac{1}{4} & \frac{1}{4} & 0 \\
0 & 0 & 0 & \frac{1}{8} & -\frac{1}{8} & 0 & 0 & \frac{1}{8} & -\frac{1}{8} & 0 & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{2} \\
0 & 0 & 0 & -\frac{1}{8} & \frac{1}{8} & 0 & 0 & -\frac{1}{8} & \frac{1}{8} & 0 & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & \frac{1}{2} \\
-\frac{1}{2} & \frac{1}{4} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{4} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} & 0 & 0 & 0 \\
0 & -\frac{1}{4} & \frac{5}{36} & \frac{1}{18} & \frac{1}{18} & -\frac{1}{4} & \frac{5}{36} & \frac{1}{18} & \frac{1}{18} & \frac{2}{9} & \frac{5}{36} & \frac{5}{36} & -\frac{1}{4} & -\frac{1}{4} & 0 \\
0 & 0 & -\frac{1}{6} & \frac{1}{12} & \frac{1}{12} & 0 & -\frac{1}{6} & \frac{1}{12} & \frac{1}{12} & -\frac{1}{6} & \frac{1}{12} & \frac{1}{12} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{2}
\end{bmatrix}$$

FIG. 5.—Matrix **L** in eq. (6), for tree (A) in fig. 4

$$\boldsymbol{y}_{\mathrm{A}} = (1/2\ 1/4\ 5/36\ 1/18\ 1/18\ 1/4\ 1/18\ 1/18\ 2/9\ 5/36\ 5/36\ 1/4\ 1/4\ 1/2)\,;$$

$$\boldsymbol{y}_{\mathrm{G}} = (1/2\ 1/8\ 1/8\ 1/8\ 1/8\ 1/8\ 1/8\ 1/8\ 1/8\ 1/2\ 1/8\ 1/8\ 1/8\ 1/8\ 1/2)\,.$$

The vectors $\boldsymbol{y}_{\mathrm{B}}$, $\boldsymbol{y}_{\mathrm{C}}$, $\boldsymbol{y}_{\mathrm{D}}$, $\boldsymbol{y}_{\mathrm{E}}$, and $\boldsymbol{y}_{\mathrm{F}}$ can be readily obtained from $\boldsymbol{y}_{\mathrm{A}}$ by renumbering of y_i's. The S values obtained from these vectors [see eq. (10)] are given for each of the topologies in figure 4. These values indicate that the NJ tree (tree A) has the smallest S but that several other trees have an S that is close to the smallest S, i.e., S_{NJ}.

To test the statistical significance of the difference (D) in S between tree A and another tree, we must compute the variance of D given by equation (21). The variances and covariances in matrix V in this equation can be obtained by using equations (27) and (28). One can therefore compute the standard error of $D_i \equiv S_i - S_{\mathrm{A}}$ by

$$s(D_i) = [(\boldsymbol{y}_i - \boldsymbol{y}_{\mathrm{A}})^t V (\boldsymbol{y}_i - \boldsymbol{y}_{\mathrm{A}})]^{1/2}\,, \tag{32}$$

where the subscript i refers to the ith tree. This computation gives the following results $\{[D_i \pm s(D_i)] \times 100\}$: 0.36 ± 0.16 for tree B, 0.44 ± 0.15 for tree C, 0.10 ± 0.07 for tree D, 0.11 ± 0.06 for tree E, 0.12 ± 0.08 for tree F, and 0.03 ± 0.08 for tree G. Therefore, a Z test [$Z = D/s(D)$] shows that the NJ tree is significantly better than trees B and C (at the 5% level), whereas the S values for other trees (D–G) are not significantly different from S_{NJ}. Therefore, we have to consider any of the latter four trees and the NJ tree as a potentially correct tree.

In addition to these trees presented in figure 4 (i.e., trees A–G), there are 98 more different topologies in the present case. We computed the S values for all of these topologies and found that the S value for four more topologies are not significantly different from S_{NJ}. All of them had $d_T = 4$, and they are given in figure 4 (trees H–K).

Trees A and D–K in figure 4 are all consistent with the view that birds are evolutionarily closer to mammals than to crocodilians. Furthermore, our computation showed that the S values for the topologies in which birds and crocodilians are sister groups are all significantly greater than S_{NJ} at the 1% level. Therefore, our results support Hedges et al.'s (1990) conclusion that the bird 18S rRNA gene is closer to the mammalian gene than to the crocodilian gene. Also, nine of the 11 trees in figure 4 have one negative branch, and thus they are unlikely to be correct. The remaining two trees (A and G) have no negative branches and have the smallest S values. It is therefore likely that one of them is the correct tree.

In tree A in figure 4 the standard errors of the estimates of branch lengths are presented, so that one can test the statistical significance of the estimate of each interior branch. The Z test with these standard errors shows that $\hat{b}_7$ is significantly different from 0 but that $\hat{b}_8$ and $\hat{b}_9$ are not. This result is consistent with the above conclusion that all trees without the cluster of bird and mammal show an S value significantly greater than S_{NJ}. However, the test with S values is more powerful than the branch-length test, because the former test examines the significance of the sum of $\hat{b}_8$ and $\hat{b}_9$ [see eq. (20)] whereas the latter examines the significance of $\hat{b}_8$ and $\hat{b}_9$ separately. Thus, the latter test cannot reject any of the 10 trees with $d_T = 4$ that are associated with the partitions of sequences at branches b_8 and b_9, whereas the former can reject six of the trees, as indicated in figure 4.

Discussion

Although there seems to be some confusion about the concept of ME in the literature, the ME method of phylogenetic inference discussed here is different from maximum-parsimony methods. In the latter methods discrete characters are used, and a topology that requires the smallest number of character-state changes is adopted as the most likely tree. Maximum-parsimony methods do not use all phylogenetic information contained in the data, contrary to Penny's (1982) statement (see Sourdis and Nei 1988), and are known to suffer from multiple nucleotide substitutions at the same sites and may produce an incorrect tree even if a large number of nucleotides are used (e.g., see Felsenstein 1978; Jin and Nei 1990). By contrast, the ME method deals with distance data, and, as long as unbiased estimates of distances are used, this method generates the correct tree as the number of nucleotides used increases (authors' unpublished data).

In this paper we have presented a simple method to find the ME tree and to test the statistical significance of topological differences in terms of S values. This method is much simpler than the bootstrap method suggested by Nei (1991). In the latter method, the variance of D must be computed by resampling nucleotides with replacement, for each pair of topologies tested, to take into account the correlation between S_A and S_B. This requires a large amount of computer time when many topologies have to be examined. Yet, according to our computer simulation, the result of the test is virtually identical for the two methods (data not shown).

In a foregoing section we suggested that all topologies related to the NJ tree with distance $d_T = 2$ or 4 should be examined to find the ME tree. We have already developed a computer algorithm for this method. However, there are several other methods to examine the topologies that are close to the NJ tree. One such method is first to examine the statistical significance of each interior branch of the NJ tree and then to compute S for all topologies associated with branches whose lengths are not significantly different from 0. In this procedure we can avoid the computation of S for trees that have $d_T = 2$ or 4 but that will probably have a significantly larger S value as in the case of trees B and C in figure 4. We are currently developing a computer algorithm for this method. When there are many small interior branches and the number of nucleotides examined is not large, this method may identify many nonsignificant interior branches, and consequently one may have to examine a large number of topologies, including those with $d_T \geq 6$. In this case, however, it would not be very meaningful to compute the D values for all of these topologies, because there are too many possible trees to permit any meaningful conclusion. In such a case the best way to resolve the topological problem would be to increase the number of nucleotides examined. At any rate, the test of interior branch lengths provides important information for phylogenetic inference.

The statistical method presented in this paper depends on the least-squares estimation of branch lengths. One might therefore wonder whether the ME method is superior to the least-squares phylogenetic-inference method proposed by Cavalli-Sforza and Edwards (1967) and Fitch and Margoliash (1967). Our theoretical study has indicated that, in obtaining the correct tree, the ME method is statistically more efficient than the least-squares method. Therefore, the ME method is superior to the least-squares method. This result will be published elsewhere.

One of the commonly used methods for testing the reliability of the branching pattern of a tree is the bootstrap test (Efron 1982, chap. 5; Felsenstein 1985; Whittam 1990). This method is equivalent to our test of interior branch lengths (see also Li 1989), and our computer simulation has shown that it gives essentially the same statistical conclusion. In practice, however, it is much easier to use the branch-length test, since this test can be carried out by using analytical formulas. Nevertheless, it should be noted that, in rejecting incorrect trees, the bootstrap and branch-length tests are less powerful than the S-value test, as mentioned earlier.

Computer Program

A computer program for computing all the quantities presented in this paper is available on request.

Acknowledgments

We thank Michael Bulmer, Tatsya Ota, Blair Hedges, and Koichiro Tamura for their comments on an earlier version of this paper. This study was supported by research grants from the National Institute of Health and the National Science Foundation to M.N.

APPENDIX A

Number of Topologies with $d_T \geq 4$

To obtain the number of trees with $d_T = 4$, we must consider two interior branches simultaneously. Because there are $n - 3$ interior branches for an unrooted bifurcating tree with n sequences, it is necessary to consider $[(n-3)(n-4)]/2$ pairs of interior branches. Some of these pairs of branches are connected with each other, but others are separated by some other branches. In the following, these two different cases will be considered separately.

1. Case of Two Contiguous Interior Branches

When two interior branches are connected with each other, the sequences involved can be divided into five groups, a_1, a_2, a_3, a_4, and a_5 (see fig. A1). Any of these groups may include one or more sequences. To generate a topology with $d_T = 4$, a cut of both branches a and b in the new topology must produce different partitions of sequences, compared with those of the original topology (i.e., tree A). A cut of branch a of tree A produces a partition of sequences which may be represented by $[(a_1,a_2)(a_3,a_4,a_5)]$, and a cut of branch b produces a partition represented by $[(a_1,a_2,a_3)(a_4,a_5)]$. Therefore, tree A may be described by $[(a_1,a_2)(a_3)(a_4,a_5)]$. All topologies with $d_T = 4$ can then be generated by rearranging the five sequence groups as follows:

$$\begin{array}{ll} [(a_1,a_3)(a_4)(a_2,a_5)]\,, & [(a_1,a_3)(a_5)(a_2,a_4)]\,, \\ [(a_1,a_4)(a_2)(a_3,a_5)]\,, & [(a_1,a_4)(a_3)(a_2,a_5)]\,, \\ [(a_1,a_4)(a_5)(a_2,a_3)]\,, & [(a_1,a_5)(a_3)(a_2,a_4)]\,, \\ [(a_1,a_5)(a_4)(a_2,a_3)]\,, & [(a_1,a_5)(a_2)(a_3,a_4)]\,, \\ [(a_2,a_4)(a_1)(a_3,a_5)]\,, & [(a_2,a_5)(a_1)(a_3,a_4)]\,. \end{array} \tag{A1}$$

These 10 topologies are shown as trees B–K in figure A1.

2. Case of Two Noncontiguous Interior Branches

In this case, each of two noncontiguous branches, say a and c in tree A in figure 1, produces four different groups of sequences and generates two topologies with $d_T = 2$. Therefore, if we consider branches a and c simultaneously, they generate four topologies with $d_T = 4$. In this case, we write the original tree as

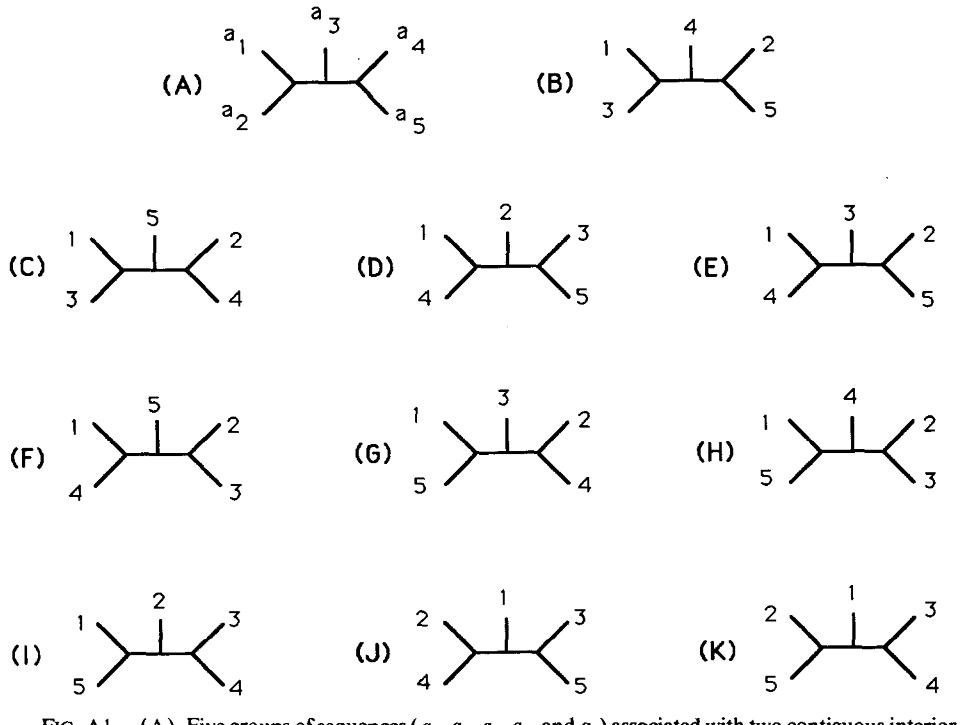

FIG. A1.—(A), Five groups of sequences (a_1, a_2, a_3, a_4, and a_5) associated with two contiguous interior branches. The rearrangement of the five groups of sequences generates 10 different topologies [(B)–(K)] with $d_T = 4$.

$[(a_1,a_2)(a_3,a_4)(b_1,b_2)(b_3,b_4)]$, where $a_1,a_2,a_3,a_4,b_1,b_2,b_3$, and b_4 represent sequence groups $\{1\}$, $\{2\}$, $\{3\}$, $\{4,5,6,7,8\}$, $\{1,2,3\}$, $\{4\}$, $\{5,6\}$, and $\{7,8\}$, respectively, in tree A in figure 1. The four topologies may then be written as follows:

$$\begin{aligned}&[(a_1,a_3)(a_2,a_4)(b_1,b_3)(b_2,b_4)]\,,\\&[(a_1,a_3)(a_2,a_4)(b_1,b_4)(b_2,b_3)]\,,\\&[(a_1,a_4)(a_2,a_3)(b_1,b_3)(b_2,b_4)]\,,\\&[(a_1,a_4)(a_2,a_3)(b_1,b_4)(b_2,b_3)]\,.\end{aligned} \tag{A2}$$

3. Number of Topologies with d_T = 4

In a tree with n sequences there are $n + n' - 6$ contiguous pairs of interior branches, and each of these pairs generates 10 topologies with $d_T = 4$. On the other hand, the number of noncontiguous pairs of interior branches is given by $[(n-3)(n-4)]/2 - (n+n'-6)$, and each of these pairs generates four topologies with $d_T = 4$. Therefore, the total number of topologies with $d_T = 4$ is given by

$$\begin{aligned}f(d_T{=}4) &= 10(n+n'-6)+4\{[(n-3)(n-4)]/2-(n+n'-6)\}\\&= 2(n^2-4n+3n'-6)\,,\end{aligned} \tag{A3}$$

which is identical with equation (3).

4. Number of Topologies with $d_T \geq 6$

To compute the number of topologies with $d_T = 2i$, where i is an integer ≥ 3, we must consider all possible combinations of i interior branches, whether they are contiguous or not. In the case of $i = 3$, for example, there are four different types of combinations of interior branches: (i) three noncontiguous branches, (ii) two contiguous branches plus one noncontiguous branch, (iii) three linear contiguous branches (e.g., branches a–c in tree A in fig. 1), and (iv) three interior branches through a tree node (e.g., branches c–e in tree A in fig. 1). In the case of type i branches, each noncontiguous branch generates two new topologies with $d_T = 2$. Therefore, any triplet of noncontiguous branches generates eight new topologies with $d_T = 6$. Thus, if the number of triplets of such branches for a tree is n_i, they generate $8n_i$ topologies with $d_T = 6$. With type ii branches, a pair of contiguous branches generates 10 topologies with $d_T = 4$, whereas a noncontiguous branch produces two topologies with $d_T = 2$. Therefore, if the number of type ii branches is n_{ii}, they generate $20n_{ii}$ topologies with $d_T = 6$.

For type iii branches, the number of topologies with $d_T = 6$ is computed by noting that the total number of possible topologies for a tree with i interior branches (with $i+3$ sequences) is

$$B = \prod_{j=3}^{i+3} (2j-5) \tag{A4}$$

(Cavalli-Sforza and Edwards 1967). In the case of $i = 3$, B becomes 105. Therefore, the number of topologies with $d_T = 6$ that are generated by one type iii branch is obtained by

$$g = B-f(d_T{=}0)-f(d_T{=}2)-f(d_T{=}4)\,. \tag{A5}$$

When $i = 3$, $f(d_T{=}0) = 1$, $f(d_T{=}2) = 6$, and $f(d_T{=}4) = 24$. Therefore, $g = 74$. Thus, if there are n_{iii} sets of type iii branches, they produce $74n_{iii}$ new topologies.

In the case of type iv branches, each set of these branches can be decomposed into three different sets of two contiguous branches each. For example, in the case of branches c–e in tree A in figure 1, we have $\{d,e\}$, $\{c,e\}$, and $\{c,d\}$. In each of these cases two contiguous branches generate 10 topologies with $d_T = 4$. Therefore, for a set of type iv branches, the number of topologies with $d_T = 6$ can be obtained by $g = B - f(d_T{=}0) - f(d_T{=}2) - f(d_T{=}4) = 105-1-6-30 = 68$. Thus, if there are n_{iv} sets of type iv branches for the tree, they generate $68n_{iv}$ topologies with $d_T = 6$. Hence, for all types of branches, the total number of topologies with $d_T = 6$ will be

$$f(d_T{=}6) = 8n_i + 20n_{ii} + 74n_{iii} + 68n_{iv}\,. \qquad \text{(A6)}$$

Of course, in order to know $f(d_T{=}6)$, we still have to determine n_i, n_{ii}, n_{iii}, and n_{iv}. When n is small, these values can be determined by counting all different types of branch combinations. When n is large, this counting is facilitated by using the topology designation in the computer algorithm mentioned below. However, for a "caterpillar" tree [unrooted bifurcating tree with $n'{=}2$ and $n{\ge}4$ (Penny and Hendy 1985)], we have obtained the following formulas:

$$f(d_T{=}4) = 2(n^2-4n)\,; \qquad \text{(A7)}$$

$$f(d_T{=}6) = (2/3)(2n^3-6n^2-5n-75)\,; \qquad \text{(A8)}$$

$$f(d_T{=}8) = (2/3)(n^4+2n^2)-48n-624\,. \qquad \text{(A9)}$$

The above method can be extended to any value of d_T, though the computation becomes more complicated.

APPENDIX B

Algorithm for Identifying Topologies with d_T = 2, 4, etc.

For writing a computer program that generates topologies with a given value of d_T, the following method is convenient: In computing $f(d_T{=}2)$, we previously showed that a cut of an interior branch generates four groups of sequences, i.e., a_1, a_2, a_3, and a_4, and that different topologies with $d_T = 2$ can be obtained by rearranging these groups. This rearrangement of groups can be done in the following way: As an example, let us consider tree A in figure 1. We first note that this tree can be defined by a series of partitions of sequences. That is, a cut of branch a of this tree produces two groups of sequences: $\{1,2\}$ and $\{3,4,5,6,7,8\}$. We denote this partition by [11000000]. Here 1 stands for a sequence that exists in the left-hand side of the branch that is cut, whereas 0 represents a sequence that exists in the right-hand side of the branch. Similarly, cuts at branches b, c, etc. generate partitions [11100000], [11110000], etc. Therefore, when all five interior branches are taken into account, tree A can be defined by the following partitions:

$$\begin{array}{ll} a, & 1\,1\,0\,0\,0\,0\,0\,0\,; \\ b, & 1\,1\,1\,0\,0\,0\,0\,0\,; \\ c, & 1\,1\,1\,1\,0\,0\,0\,0\,; \\ d, & 1\,1\,1\,1\,0\,0\,1\,1\,; \\ e, & 1\,1\,1\,1\,1\,1\,0\,0\,. \end{array} \qquad \text{(B1)}$$

Previously we mentioned that partition a can be written as $[(a_1,a_2)(a_3,a_4)]$, where $a_1 = \{1\}$, $a_2 = \{2\}$, $a_3 = \{3\}$, and $a_4 = \{4,5,6,7,8\}$, and that this partition generates two new topologies or partitions: $[(a_1,a_3)(a_2,a_4)]$ and $[(a_1,a_4)(a_2,a_3)]$. $[(a_1,a_3)(a_2,a_4)]$ represents the case where a_1 and a_3 exist in the left-hand of the branch under consideration and where a_2 and a_4 exist in the right-hand site. Therefore, if we use the binary notation, it can be written as [10100000]. Similarly, $[(a_1,a_4)(a_2,a_3)]$ can be written as [10011111]. Each of these new partitions plus partitions *b–e* in set (B1) now define a new topology. The rearrangement of sequence groups a_1, a_2, a_3, and a_4 for each of the partitions *b–e* in set (B1) also generates two new partitions. Therefore, we can define all new topologies with $d_T = 2$, by a binary system similar to that of set (B1). Obviously, this procedure can be used for any type of tree.

To generate a new topology with $d_T = 4$, we consider two sets of four groups of sequences, i.e., $[(a_1,a_2)(a_3,a_4)]$ and $[(b_1,b_2)(b_3,b_4)]$. If a_1, a_2, a_3, and a_4 are all different from b_1, b_2, b_3, and b_4, then the interior branches corresponding to the two partitions are noncontiguous. In this case, each pair of interior branches generates four new sequence partitions with $d_T = 4$, as discussed earlier. These new partitions also can be described by a binary notation. Therefore, if we combine one of these partitions with the partitions with respect to the remaining interior branches, we can define a new topology with a binary system similar to that in set (B1). By contrast, if any of a_1, a_2, a_3, and a_4 is identical with any of b_1, b_2, b_3, and b_4, then the corresponding interior branches are contiguous. In this case, each pair of interior branches generates 10 partitions with $d_T = 4$, but all topologies can be identified by using the procedure mentioned above.

Note added in proof.—Equation (31) was derived by a method analogous to that of equation (25), and numerical computation has shown that it is quite accurate for most cases. However, a more accurate equation for $\mathrm{Cov}(d_{ij},d_{kl})$ can be derived by the delta technique (see Kendall and Stuart 1958, pp. 231–232). It becomes

$$\mathrm{Cov}(d_{ij},d_{kl}) = [a_{ij}a_{kl}(\mathrm{P}_{ij,kl}-\mathrm{P}_{ij}\mathrm{P}_{kl})+a_{ij}b_{kl}(\mathrm{R}_{ij,kl}-\mathrm{P}_{ij}\mathrm{Q}_{kl}) \\ +b_{ij}a_{kl}(S_{ij,kl}-\mathrm{Q}_{ij}\mathrm{P}_{kl})+b_{ij}b_{kl}(\mathrm{Q}_{ij,kl}-\mathrm{Q}_{ij}\mathrm{Q}_{kl})]/m\,,$$

where P_{ij} and Q_{ij} are the same as those in (31). $\mathrm{P}_{ij,kl}$ is the proportion of sites where transitional differences are observed between sequences *i* and *j* and between sequences *k* and *l*, $\mathrm{Q}_{ij,kl}$ is the proportion of sites where transversional differences are observed between sequences *i* and *j* and between sequences *k* and *l*, $\mathrm{R}_{ij,kl}$ is the proportion of sites where sequences *i* and *j* have transitional differences but sequences *k* and *l* have transversional differences, and $S_{ij,kl}$ is the proportion of sites where sequences *i* and *j* have transversional differences but sequences *k* and *l* have transitional differences. a_{ij}, b_{ij}, a_{kl}, and b_{kl} are the same as those in equation (31).

LITERATURE CITED

BULMER, M. 1989. Estimating the variability of substitution rates. Genetics **123**:615–619.

———. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. **8**:868–883.

CAVALLI-SFORZA, L. L., and A. W. F. EDWARDS. 1967. Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet. **19**:233–257.

EFRON, B. 1982. The jackknife, the bootstrap and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia.

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

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

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

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

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

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

KENDALL, G., and A. STUART. 1958. The advanced theory of statistics. Vol. **1**. Hafner, New York.

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.

KIMURA, M., and T. OHTA. 1972. On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol. **2**:87–90.

LI, W.-H. 1989. A statistical test of phylogenies estimated from sequence data. Mol. Evol. Biol. **6**:424–435.

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

———. 1991. Relative efficiencies of different tree-making methods for molecular data. Pp. 90–128 *in* M. M. MIYAMOTO and J. L. CRACRAFT, eds. Recent advances in phylogenetic studies of DNA sequences. Oxford University Press, Oxford.

NEI, M., and L. JIN. 1989. Variances of the average numbers of nucleotide substitutions within and between populations. Mol. Biol. Evol. **6**:290–300.

NEI, M., J. C. STEPHENS, and N. SAITOU. 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. **2**:66–85.

PENNY, D. 1982. Towards a basis for classification: incompleteness of distance measures, incompatability analysis, and phenetic classification. J. Theor. Biol. **96**:129–142.

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

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

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

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

TAJIMA, F., and M. NEI. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1**:269–285.

WHITTAM, T. 1990. NJBOOT: a computer program. Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, Pa.

MICHAEL BULMER, reviewing editor

Received December 15, 1991; revision received April 20, 1992

Accepted April 20, 1992