Theoretical Foundation of the Minimum-Evolution Method of Phylogenetic Inference[1]

Andre Rzhetsky and Masatoshi Nei

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

The minimum-evolution (ME) method of phylogenetic inference is based on the assumption that the tree with the smallest sum of branch length estimates is most likely to be the true one. In the past this assumption has been used without mathematical proof. Here we present the theoretical basis of this method by showing that the expectation of the sum of branch length estimates for the true tree is smallest among all possible trees, provided that the evolutionary distances used are statistically unbiased and that the branch lengths are estimated by the ordinary least-squares method. We also present simple mathematical formulas for computing branch length estimates and their standard errors for any unrooted bifurcating tree, with the least-squares approach. As a numerical example, we have analyzed mtDNA sequence data obtained by Vigilant et al. and have found the ME tree for 95 human and 1 chimpanzee (outgroup) sequences. The tree was somewhat different from the neighbor-joining tree constructed by Tamura and Nei, but there was no statistically significant difference between them.

Introduction

Rzhetsky and Nei (1992*a*) proposed a minimum-evolution (ME) method of phylogenetic inference in which the branch lengths are estimated, by the ordinary least-squares (OLS) method, from distance matrices. The algorithm of their method is first to construct a neighbor-joining (NJ) tree by using Saitou and Nei's (1987) procedure and to compute the total sum (S) of branch lengths for this tree. Next, all tree topologies that are close to the NJ tree by certain criteria are examined, and the S value for each tree is computed. The S values thus obtained are then compared with each other, and a tree with the smallest S value will be chosen as the final one. This final tree is usually the NJ tree, but the NJ method sometimes fails to identify the ME tree. Of course, if the number of sequences is relatively small, it is possible to examine all topologies, but usually this is unnecessary. A statistical test of the difference in S between different topologies was also developed. Computer simulations have shown that the ME method is more efficient than most other distance methods of phylogenetic inference and that the statistical test proposed is conservative.

This method clearly depends on the assumption that the true tree has the smallest expected value [$E(S)$] of S. We have shown that this is true for the case of four DNA or amino acid sequences when unbiased estimates of evolutionary distances (number of nucleotide or amino acid replacements) are used (Rzhetsky and Nei 1992*a*, 1992*b*).

1. Key words: minimum sum of branch lengths, least-squares estimates of branch lengths, unbiasedness of the estimates of evolutionary distances.

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

Mol. Biol. Evol. 10(5):1073–1095. 1993.

0737-4038/93/1005-0011$02.00

[This is also true when unbiased estimates of genetic distances, such as Nei's (1972) standard genetic distance, are used for estimating population phylogenies.] However, it is still unclear whether this is true irrespective of the number of sequences and topology. The purpose of this paper is to show that this is exactly the case. Before dealing with this problem, however, we shall present a new algorithm for estimating branch lengths, since this is useful for proving our assertion. This algorithm also simplifies the estimation of branch lengths tremendously when the number of sequences used is large.

New Algorithm for Estimating Branch Lengths

Rzhetsky and Nei (1992*a*) used the following equation to estimate the branch lengths by the OLS method:

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

where $\hat{\mathbf{b}}$ is the column vector of the OLS estimates of branch lengths of a tree under consideration, $\mathbf{A}$ is the topological matrix describing the tree structure (see Rzhetsky and Nei 1992*a*), $\mathbf{d}$ is the column vector of the estimates of evolutionary distances between the sequences, and t and -1 stand for the operations of matrix transposing and inversion, respectively. In practice, however, estimation of branch lengths by equation (1) is not always easy, because a large amount of computer memory is required when the number of sequences is large.

This problem can be solved if we estimate branch lengths without using matrix algebra. Consider tree (A) in figure 1 as an example. If we choose one particular interior branch of this tree, this tree can be drawn in the form of tree (B) in the same figure, where A, B, C, and D each represent a cluster of sequences. For example, for interior branch b in figure 1(A), A, B, C, and D represent clusters (3), (1, 2), (4), and (5, 6, 7, 8), respectively. In this case the branch length b in tree (B) can be estimated by the following equation:

$$\begin{aligned}\hat{b} = \frac{1}{2}\{&\gamma[d_{AC}/(n_An_C) + d_{BD}/(n_Bn_D)] \\ &+ (1-\gamma)[d_{BC}/(n_Bn_C) + d_{AD}/(n_An_D)] - d_{AB}/(n_An_B) - d_{CD}/(n_Cn_D)\}\,,\end{aligned} \tag{2}$$

where

$$\gamma = (n_Bn_C + n_An_D)/[(n_A + n_B)(n_C + n_D)]\,, \tag{3}$$

n_A, n_B, n_C, and n_D are the numbers of sequences in the clusters A, B, C, and D, respectively, and d_{AC} is the sum of all intercluster distances, where one sequence belongs to cluster A and the other belongs to cluster C. d_{BD}, d_{BC}, d_{AD}, d_{AB}, and d_{CD} are defined in a similar fashion. By contrast, the OLS estimator of the length b of an exterior branch of tree (C) in figure 1 is given by

$$\hat{b} = [d_{CA}/n_A + d_{CB}/n_B - d_{AB}/(n_An_B)]/2\,, \tag{4}$$

where d_{CA} is the sum of all pairwise distances between sequence C and all sequences belonging to cluster A, d_{CB} is the sum of distances between C and all sequences be-

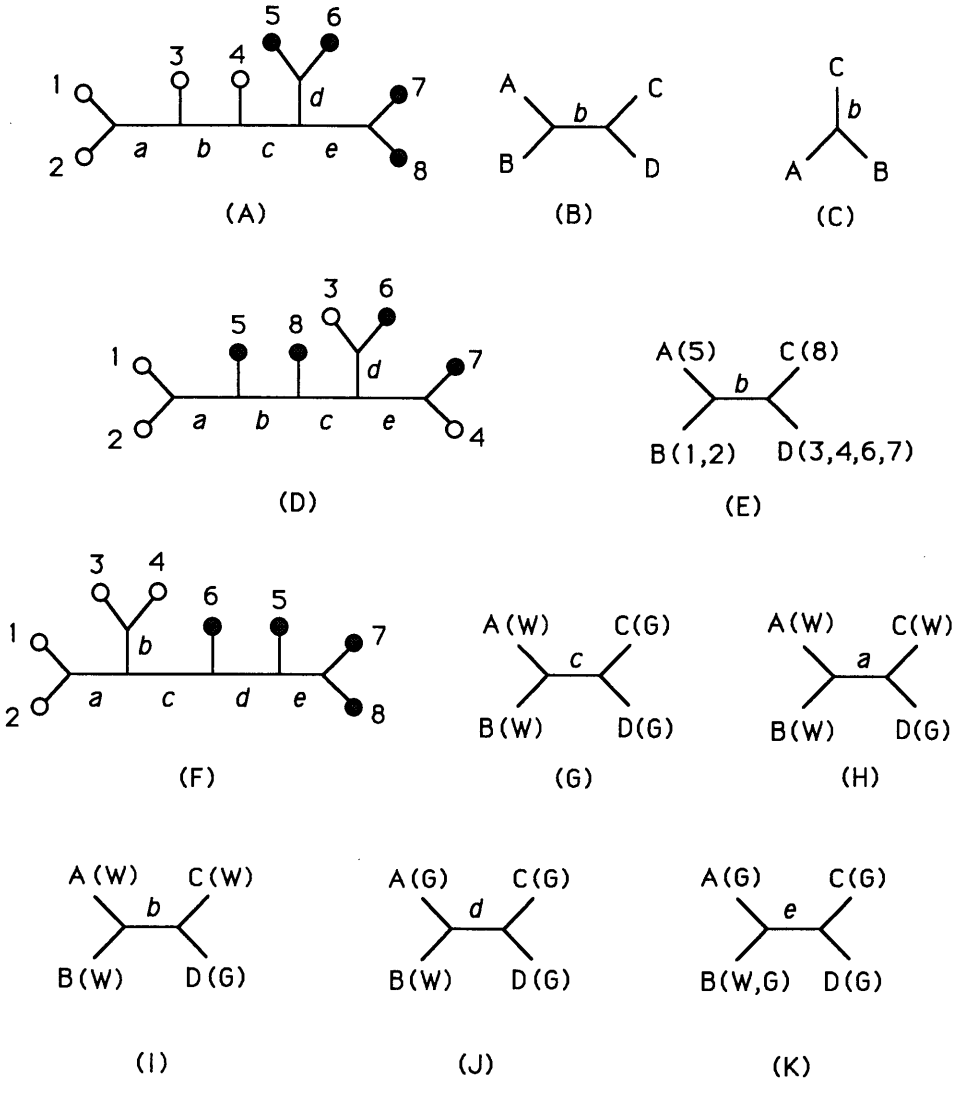


FIG. 1.—(A), Bifurcating true tree for eight sequences. Branch *c* of this tree separates the sequences into two groups, which are denoted by white and gray circles. (B), Four sequence clusters (A, B, C, and D) generated when a particular interior branch (*b*) is considered. (C), One-sequence and multiple-sequence clusters (A, B, and C) generated when a particular exterior branch (*b*) is considered. (D), Wrong tree for the same sequences as those for tree (A). This tree has an incongruent configuration (see text). (E), Four sequence clusters generated when branch *b* in tree (D) is considered. (F), Another wrong tree for the same sequences as those for tree (A). This tree has a congruent configuration. (G)–(K), Four clusters generated when different interior branches of tree (F) were considered. (W) and (G) = "white" and "gray" clusters, respectively; (W,G) = a mixture of white and gray sequences.

longing to cluster B, d_{AB} is the sum of all intercluster distances between sequences in clusters A and B, and n_A and n_B are the numbers of sequences in the clusters A and B, respectively. Derivation of equations (2) and (4) is given in Appendix A.

As was already mentioned, equations (2) and (4) facilitate the computation of OLS branch length estimates enormously, and, even if the number of sequences used is very large (>100), the computer time and memory required are relatively small

(see Appendix B). The variance for each branch length estimate can also be computed in a similar fashion (see Appendix B).

Incidentally, formulas (2) and (4) in combination with the formulas for the corresponding variances (see Appendix B) give an extension of the topological test suggested by Li (1989), to the case of an arbitrary number of sequences. However, as was shown by Rzhetsky and Nei (1992*a*), this test is less efficient in rejecting wrong trees than is the test of the differences in S values between alternative trees.

Proof That $E(S)$ Is Smallest for the True Tree

Special Case

Let us first consider the case of five hypothetical sequences and assume that tree (A) in figure 2 is the correct topology and that tree (B) is a wrong one. An unbiased estimate of the evolutionary distance, $\hat{d}_{ij}$, between sequences i and j is given by

$$
\begin{aligned}
\hat{d}_{12} &= b_1 + b_2 + e_{12}, \\
\hat{d}_{13} &= b_1 + b_3 + b_6 + e_{13}, \\
\hat{d}_{14} &= b_1 + b_4 + b_6 + b_7 + e_{14}, \\
\hat{d}_{15} &= b_1 + b_5 + b_6 + b_7 + e_{15}, \\
\hat{d}_{23} &= b_2 + b_3 + b_6 + e_{23}, \\
\hat{d}_{24} &= b_2 + b_4 + b_6 + b_7 + e_{24}, \\
\hat{d}_{25} &= b_2 + b_5 + b_6 + b_7 + e_{25}, \\
\hat{d}_{34} &= b_3 + b_4 + b_7 + e_{34}, \\
\hat{d}_{35} &= b_3 + b_5 + b_7 + e_{35}, \\
\hat{d}_{45} &= b_4 + b_5 + e_{45},
\end{aligned}
\tag{5}
$$

where b_i's are the true branch lengths, and e_{ij}'s are sampling errors with mean 0 and variance $V(\hat{d}_{ij})$. For any topology, the branch length estimates are linear functions of d_{ij}'s, which are in turn linear functions of b_i's and e_{ij}'s. Therefore, the expected value of the sum of branch length estimates $[E(S)]$ is also a linear function of b_i's. In the case of the true topology we can show that the expectation of S is

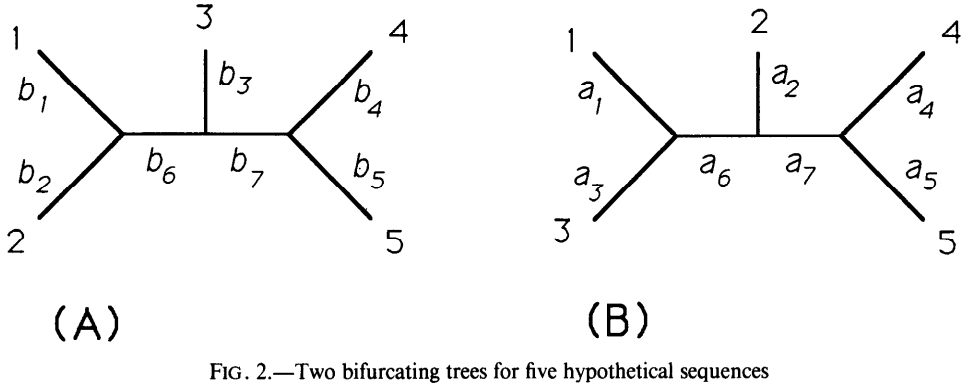


FIG. 2.—Two bifurcating trees for five hypothetical sequences

$$E(S_A) = b_1 + b_2 + b_3 + b_4 + b_5 + b_6 + b_7 , \tag{6}$$

which is intuitively obvious.

By contrast, the branch length estimates ($\hat{a}_i$'s) for tree (B) in figure 2 can be expressed in terms of b_i's and e_{ij}'s by using equations (2), (4), and (5). They become

$$\begin{aligned}
\hat{a}_1 &= b_1 + \frac{2}{3} b_6 + (e_{12} + 3e_{13} + e_{14} + e_{15} - e_{23} - e_{34} - e_{35})/6 , \\
\hat{a}_2 &= b_2 + \frac{1}{2} b_6 + (2e_{12} - e_{14} - e_{15} + 2e_{23} + 2e_{24} + 2e_{25} - e_{34} - e_{35})/8 , \\
\hat{a}_3 &= b_3 + \frac{1}{3} b_6 + (-e_{12} + 3e_{13} - e_{14} - e_{15} + e_{23} + e_{34} + e_{35})/6 , \\
\hat{a}_4 &= b_4 + (e_{14} - e_{15} + e_{24} - e_{25} + e_{34} - e_{35} + 3e_{45})/6 , \\
\hat{a}_5 &= b_5 + (-e_{14} + e_{15} - e_{24} + e_{25} - e_{34} + e_{35} + 3e_{45})/6 , \\
\hat{a}_6 &= -\frac{1}{2} b_6 + (2e_{12} - 4e_{13} + e_{14} + e_{15} + 2e_{23} - 2e_{24} - 2e_{25} + e_{34} + e_{35})/8 , \\
\hat{a}_7 &= \frac{1}{2} b_6 + b_7 + (-2e_{12} + e_{14} + e_{15} - 2e_{23} + 2e_{24} + 2e_{25} + e_{34} + e_{35} - 4e_{45})/8 .
\end{aligned} \tag{7}$$

These equations indicate that the expectation $[E(\hat{a}_i)]$ of $\hat{a}_i$ is a linear function of b_i's, since $E(e_{ij})$'s are all 0. Actually, for any topology $E(\hat{a}_i)$ is a linear function of $E(d_{ij})$'s, which are in turn linear functions of b_i's, as mentioned earlier. This is true irrespective of the number of sequences involved. Therefore, we have

$$E(\hat{a}_i) = \alpha_{i,1} b_1 + \alpha_{i,2} b_2 + \ldots + \alpha_{i,2n-3} b_{2n-3} , \tag{8}$$

where $\alpha_{i,j}$'s are coefficients of b_j's, and $2n - 3$ is the total number of branches of a bifurcating tree for n sequences. This indicates that, if we denote the column vectors of $E(\hat{a}_i)$'s and b_i's by **a** and **b**, respectively, and the matrix of $\alpha_{i,j}$'s by $\boldsymbol{\alpha}$, we can write the following equation:

$$\mathbf{a} = \boldsymbol{\alpha}\mathbf{b} , \tag{9}$$

where $\boldsymbol{\alpha}$ can be computed as

$$\boldsymbol{\alpha} = (\mathbf{A}_W^t \mathbf{A}_W)^{-1} \mathbf{A}_W^t \mathbf{A}_T , \tag{10}$$

where $\mathbf{A}_W$ and $\mathbf{A}_T$ are the topological matrices of the wrong and true topologies under consideration, respectively. (The topological matrix, **A**, for an unrooted bifurcating tree for n sequences is defined as an $[n(n-1)/2] \times [2n-3]$ matrix, where element a_{ij} is equal to 1 if the ith distance includes the jth branch, and $a_{ij} = 0$ otherwise.)

In the present example, $\mathbf{A}_T$ for tree (A) and $\mathbf{A}_W$ for tree (B) in figure 2 are

$$\mathbf{A}_T = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{A}_W = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

Therefore, $\boldsymbol{\alpha}$ is

$$\boldsymbol{\alpha} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \frac{2}{3} & 0 \\ 0 & 1 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 1 \end{bmatrix}. \tag{11}$$

Thus, from equation (9), we have $E(\hat{a}_1) = b_1 + (\frac{2}{3})b_6$, $E(\hat{a}_2) = b_2 + (\frac{1}{2})b_6$, etc., which agree with the values obtained from equation (7).

It is also important to note that, if $b_1 = 1$ and $b_i = 0$ for $i \neq 1$, then $E(\hat{a}_1) = 1 = \alpha_{1,1}$, from equations (7) and (8). Similarly, if $b_6 = 1$ and $b_i = 0$ for $i \neq 6$, then $E(\hat{a}_1) = \frac{2}{3} = \alpha_{1,6}$. In general, $\alpha_{i,j} = E(\hat{a}_i)$ if $b_j = 1$, and $b_k = 0$ for all $k \neq i$. This indicates that $\alpha_{i,j}$'s can be obtained from the equation for $E(\hat{a}_i)$'s. We shall use this property when we consider the general case.

Our purpose is to show that $E(S)$ for the true tree is smallest among all possible topologies, i.e., $E(S_W - S_T) > 0$, where S_T and S_W are the S values for the true topology and a wrong topology, respectively. In the present case, summing up all branch lengths estimates ($\hat{a}_i$'s and $\hat{b}_i$'s) for the trees in figure 2, we have $E(S_B - S_A) = b_6/2 > 0$, if $b_6 > 0$. In general, however, we can write $E(S_W - S_T)$ in the following way:

$$E(S_W - S_T) = \beta_1 b_1 + \beta_2 b_2 + \ldots + \beta_{2n-3} b_{2n-3}, \tag{12}$$

where

$$\beta_j = \alpha_{1,j} + \alpha_{2,j} + \ldots + \alpha_{2n-3,j} - 1. \tag{13}$$

In tree (B) of figure 2, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_7 = 0$, and $\beta_6 = \frac{2}{3} + \frac{1}{2} + \frac{1}{3} + 0 + 0 - \frac{1}{2} + \frac{1}{2} - 1 = \frac{1}{2}$. Therefore, $E(S_W - S_T) = E(S_B - S_A) = b_6/2$, as obtained above.

The present example shows that β_i's [coefficients of b_i's in equation (12)] have two important properties: (1) All β_i's for exterior branches are 0. (2) An interior branch that produces the same partitioning of sequences (see Rzhetsky and Nei 1992*a*) for the true tree and the wrong tree has $\beta_i = 0$. In the present case the interior branch a_7 in tree (B) produces the same partitioning as does b_7 in the true tree. Therefore, $\beta_7 = 0$.

General Case

In the above example we have seen that β_i's in equation (12) are either 0 or positive, so that $E(S_W - S_T) > 0$. Let us now prove that this is true for any number of sequences. This can be accomplished by showing that at least one term, $\beta_k b_k$, in equation (12) is positive and that all other terms are non-negative. As in the previous example, let $b_1, b_2, \ldots$, and b_{2n-3} be the branch lengths of the true bifurcating tree, and let $E(\hat{a}_1), E(\hat{a}_2), \ldots$, and $E(\hat{a}_{2n-3})$ be the expected values of the OLS branch length estimates for a wrong bifurcating tree.

First, we note that all β_i's associated with exterior branches are always 0 irrespective of the number of sequences and topology. This is because, if the lengths of all interior branches of the true tree are 0, then all the bifurcating trees become the same multifurcating tree, and thus $E(S_W - S_T) = 0$. Therefore, we need to consider only β_i's associated with the interior branches of the true tree.

Consider a particular interior branch, *i*, of the true tree and study the value of β_i for a wrong tree. To simplify our explanation, let us denote all sequences on one side of branch *i* in the true tree by white circles and the sequences on the other side of branch *i* by gray circles. This has been done in tree (A) of figure 1 by considering branch *c*. We also denote each sequence in a wrong tree by the same color as that in the true tree [see fig. 1(D)]. If we do this, there may be two different types of configurations of white and gray sequences in a wrong tree. (1) White and gray sequences are separated at particular branch (k) of a wrong tree in the same way as at that of the true tree. For example, if we consider branch *a* in the true tree (A) of figure 1 and denote sequences 1 and 2 by white circles and the rest of the sequences by gray circles, the same partition of white and gray sequences is obtained in tree (D) as well. In the following we call this type of sequence configuration a "congruent" configuration. (2) White and gray sequences form four or more monochrome clusters. (Three monochrome clusters can always be reduced to two monochrome clusters, as in a congruent configuration.) For example, tree (D) of figure 1 forms three white clusters (1,2; 3; and 4) and four gray clusters (5; 8; 6; and 7). We call this type of configuration an "incongruent" configuration.

In the example discussed in the previous section we have seen that for a congruent configuration $\beta_i = 0$, whereas for an incongruent configuration $\beta_i > 0$. Let us now show that this is true for any topology and any number of sequences. We have seen that, if $\alpha_{j,i}$'s are known, then β_i is obtained from equation (13) and that $\alpha_{j,i}$ can be obtained by putting $b_i = 1$ and $b_m = 0$ for all $m \neq i$ in equation (8). In practice, $\alpha_{j,i}$ can be obtained by using equations (2) and (4), because these equations or equation (1) are equivalent to equation (8) or equation (9). That is, $\alpha_{j,i}$ is obtained by substituting 1 for all pairwise distances that include the ith branch of the true tree (i.e., all distances between white and gray sequences) and by substituting 0 for all other distances

(i.e., distances between sequences of the same color). This can be done easily if we consider four clusters of sequences (A, B, C, and D) with respect to the jth interior branch of the wrong tree [see fig. 1(E)] and count the number of white and gray sequences in each cluster. Let W_A, G_A; W_B, G_B; W_C, G_C; and W_D, G_D be the numbers of white and gray sequences in clusters A, B, C, and D, respectively. The sum of distances between clusters A and C (d_{AC}) should then be replaced by $W_A W_C \cdot 0 + W_A G_C \cdot 1 + G_A W_C \cdot 1 + G_A G_C \cdot 0 = W_A G_C + G_A W_C$. Similarly, d_{BD}, d_{BC}, etc., should be replaced by $W_B G_D + G_B W_D$, $W_B G_C + G_B W_C$, etc., respectively. We can then compute $\alpha_{j,i}$ by

$$\begin{aligned}\alpha_{j,i} = \gamma&\left(\frac{W_A G_C + G_A W_C}{2n_A n_C} + \frac{W_B G_D + G_B W_D}{2n_B n_D}\right) \\ &+ (1-\gamma)\left(\frac{W_B G_C + G_B W_C}{2n_B n_C} + \frac{W_A G_D + G_A W_D}{2n_A n_D}\right) \\ &- \frac{G_A W_B + G_B W_A}{2n_A n_B} - \frac{G_C W_D + G_D W_C}{2n_C n_D},\end{aligned} \tag{14}$$

where γ, n_A, n_B, n_C, and n_D are the same as those in equation (2).

If the jth branch of a wrong tree is an exterior branch (b) connected with clusters A and B [see fig. 1(C)], then $\alpha_{j,i}$ is given by

$$\alpha_{j,i} = \frac{G_A G_B}{n_A n_B}, \text{ if sequence C is white}, \tag{15a}$$

and by

$$\alpha_{j,i} = \frac{W_A W_B}{n_A n_B}, \text{ if sequence C is gray}. \tag{15b}$$

We are now in a position to compute β_i from $\alpha_{j,i}$'s by using equation (13). Let us compute β_i for congruent and incongruent configurations separately.

1. *Congruent Configuration*

In this case white and gray sequences are separated at an interior branch k. Consider tree (A) and tree (F) of figure 1, where the former is the correct tree and the latter is an incorrect tree. However, the latter tree has a congruent configuration at branch $k = c$, with respect to interior branch c of the true tree. In this congruent configuration the four clusters A(3,4), B(1,2), C(6), and D(5,7,8) can be written in the form of tree (G), where W and G represent white and gray sequence clusters, respectively. In this case $G_A = G_B = W_C = W_D = 0$, $W_A = n_A$, $W_B = n_B$, $G_C = n_C$, and $G_D = n_D$ in equation (14). Therefore,

$$\alpha_{k,i} = \gamma\left[\frac{n_A n_C}{2n_A n_C} + \frac{n_B n_D}{2n_B n_D}\right] + (1-\gamma)\left[\frac{n_B n_C}{2n_B n_C} + \frac{n_A n_D}{2n_A n_D}\right] = 1.$$

Note that the above equation holds for any topology and any number of sequences, because it is independent of the values of n_A, n_B, n_C, and n_D.

When $j \neq k$, four different sets of clusters are obtainable [trees (H), (I), (J), and (K) of fig. 1]. In tree (H) $G_A = G_B = G_C = W_D = 0$, $W_A = n_A$, $W_B = n_B$, $W_C = n_C$, and $G_D = n_D$. Therefore, putting these into equation (14), we have $\alpha_{j,i} = 0$. Similarly, we obtain $\alpha_{j,i} = 0$ for all other trees [(I), (J), and (K) of fig. 1]. Note that $\alpha_{j,i} = 0$ ($j \neq k$) also holds for any topology and any number of sequences, as long as the sequence configuration is congruent with respect to branch i of the true tree. Therefore, we have $\beta_i = 0$ from equation (13) for any congruent configuration of sequences.

2. *Incongruent Configuration*

Since $\beta_i = 0$ for any congruent configuration of sequences, our assertion $E(S_W - S_T) > 0$ demands that $\beta_i > 0$ for incongruent configurations. It seems to be difficult to compute β_i for an arbitrary tree with any number of sequences. We therefore prove our assertion by transforming an incongruent configuration step by step into a congruent one and showing that in each step of transformation β_i decreases and that it becomes 0 when the configuration of sequences becomes congruent.

Consider an incongruent configuration given in tree (A) of figure 3. The sequences of this tree form four white clusters (1; 5; 7; and 8) and five gray clusters (2; 3; 4; 6; and 9). This configuration can be transformed into the congruent configuration given in tree (F) through five steps by using two operations given in figure 4. In figure 4, the white and gray circles denote white and gray clusters (one or more sequences), respectively. In operation I, clusters B and C, which are separated by one interior branch, are interchanged so that a group of three monochrome clusters is transformed into a group of two monochrome clusters (C and A + B). (Here the colors of clusters A, B, and C can be white, white, and gray, respectively.) In operation II, clusters B and C, which are separated by two interior branches, are interchanged so that a group of four monochrome clusters is transformed into a group of two monochrome clusters (see fig. 4). At any rate, if we apply these two operations sequentially, any incongruent configuration can be transformed into a congruent one. An example of the application of the above operations is given in figure 3.

It can be shown that in each step of these operations β_i decreases, though proof of this assertion is tedious (see Appendix C). Therefore, one can conclude that an incongruent configuration always has $\beta_i > 0$.

Obviously, the correct tree has a congruent configuration for any interior branch. Therefore, all β_i's are 0. By contrast, any incorrect tree has at least one incongruent configuration of sequences, compared with the correct tree. Therefore, as long as all interior branches of the true tree are positive and the estimates of evolutionary distances are unbiased, the expectation of the sum of the estimates of branch lengths for the true topology is smallest among all possible topologies.

If we use matrix algebra, this finding or theorem may be expressed in the following way:

$$E(S_W - S_T) = \mathbf{u}[(\mathbf{A}_W^t \mathbf{A}_W)^{-1} \mathbf{A}_W^t \mathbf{A}_T - \mathbf{I}]\mathbf{b} > 0\,, \qquad (16)$$

where $\mathbf{A}_T$ is the topological matrix for the true tree, $\mathbf{A}_W$ is the topological matrix for a wrong bifurcating tree, $\mathbf{I}$ is an identity matrix, $\mathbf{u}$ is a row vector consisting of unity, and $\mathbf{b}$ is a column vector of the branch lengths of the true tree.

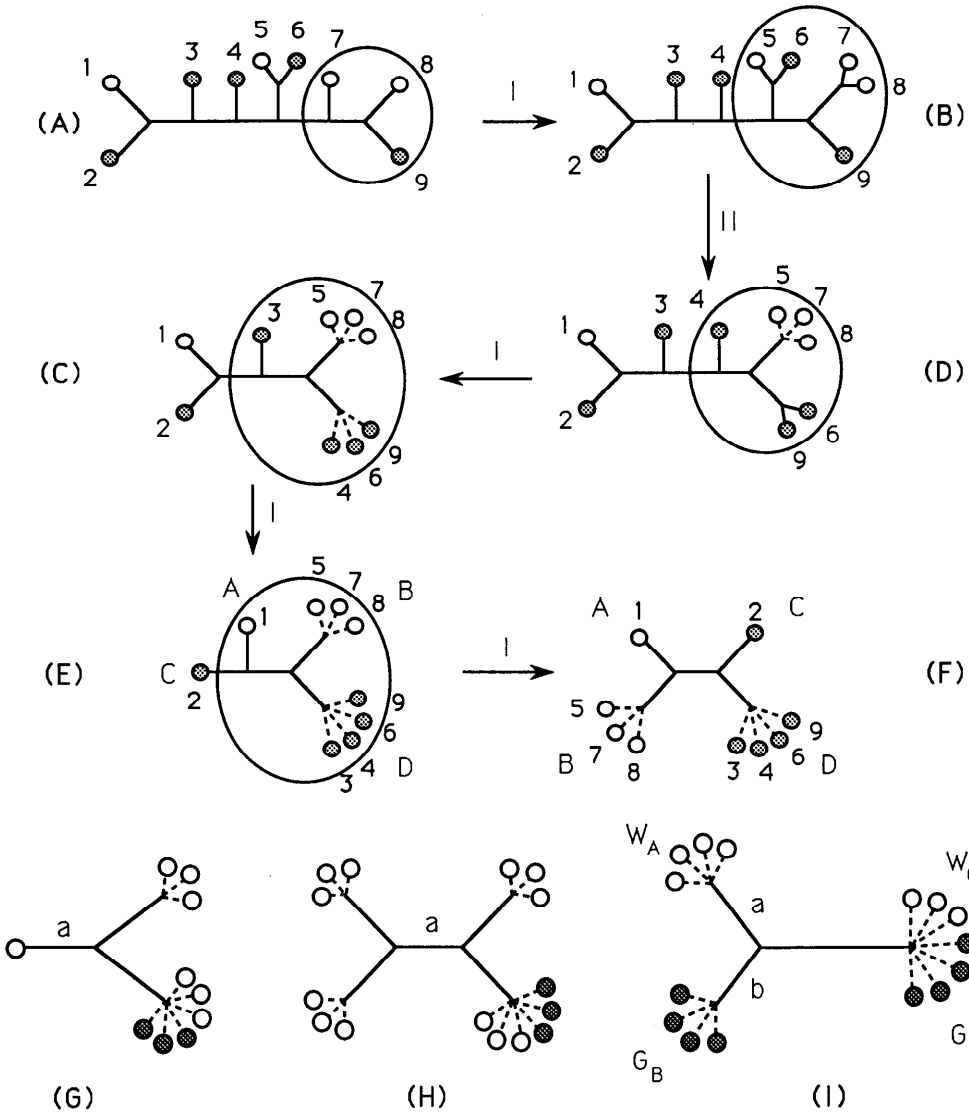


FIG. 3.—(A), Wrong tree with an incongruent configuration. (F), Tree (not necessarily the true tree) with a congruent configuration. (A)–(F), Successive application of operations I and II, which converts an incongruent configuration into a congruent one. (G)–(I), Several patterns of incongruent configurations (see Appendix C). A group of sequences with dashed lines indicates a bifurcating cluster.

Discussion

In proving our theorem $E(S_W - S_T) > 0$, we used the OLS to estimate the branch lengths . One might therefore wonder what will happen if we use the generalized least-squares method to estimate S_W and S_T. Actually, Rzhetsky and Nei (1992*b*) have shown that, if the generalized least-squares method is used, $E(S_W - S_T) > 0$ does not necessarily hold for the case of four sequences. Therefore, this method should not be used for estimating S_W and S_T, though it has several nice statistical properties when applied to data with a multivariate normal distribution. The problem with the present case is that evolutionary distances almost never follow a multivariate normal distribution.

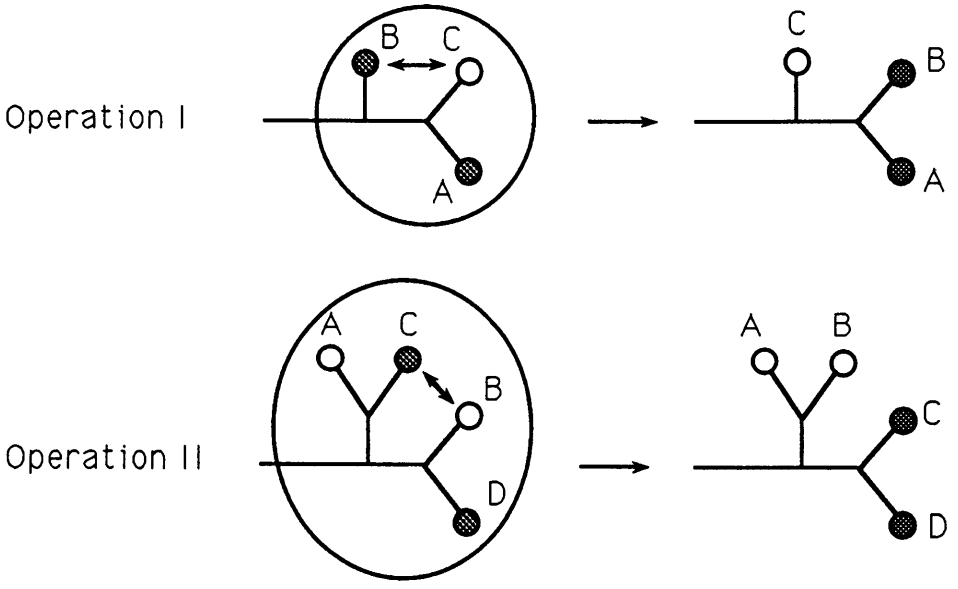


FIG. 4.—Graphic explanation of operations I and II

In this paper we have assumed that the estimates of evolutionary distances used are statistically unbiased. In practice, however, this assumption does not necessarily hold. It is therefore interesting to know the effect of violation of this assumption on $E(S_W - S_T)$. Using simple algebra, one can show that, if there is a systematic error in the estimate of evolutionary distance but the magnitude of the error is small and approximately proportional to the true distance, then the inequality $E(S_W - S_T) > 0$ still holds.

Actually, even when the magnitude of the error is not proportional to the true distance, the inequality $E(S_W - S_T) > 0$ often holds. Let us illustrate this by considering a simple example of four nucleotide sequences [fig. 5, where tree (A) is the correct tree and trees (B) and (C) are wrong trees]. We assume that nucleotide substitution occurs according to the Jukes and Cantor (1969) model but that the number of substitutions per site between the ith and jth sequences (d_{ij}) is estimated by the proportion of different nucleotides between the two sequences (p_{ij}). In this case $E(p_{ij})$ is related to $E(d_{ij})$ by

$$E(p_{ij}) = \left(\frac{3}{4}\right)[1 - \exp(-4E(d_{ij})/3)] \qquad (17)$$

(Jukes and Cantor 1969). Therefore, p_{ij} is not proportional to d_{ij}. However, if we use p_{ij} as the estimate of d_{ij}, $E(S_W - S_T)$ becomes

$$E(S_W - S_T) = [E(p_{13}) + E(p_{24}) - E(p_{12}) - E(p_{34})]/4 . \qquad (18)$$

Note that this equation holds also for the comparison of trees (A) and (C) in figure 5 when subscripts 3 and 4 are interchanged. Equation (18) indicates that, if the rate of evolution is constant, the root of the tree is located at branch 5, $b_1 = b_2$, and $b_3 = b_4$, then $E(S_W - S_T)$ is always positive as long as b_5 is positive and b_1, b_2, b_3, and

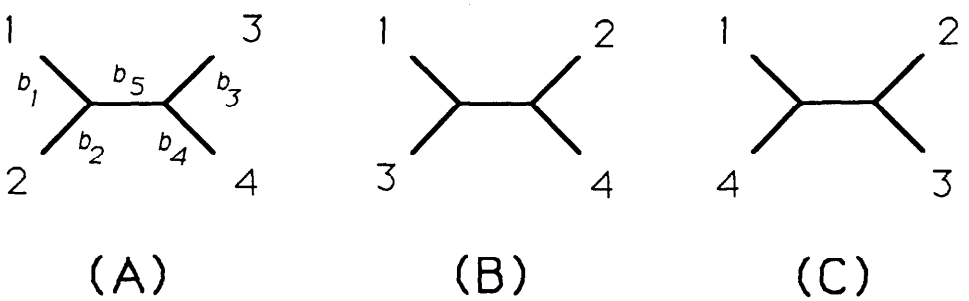


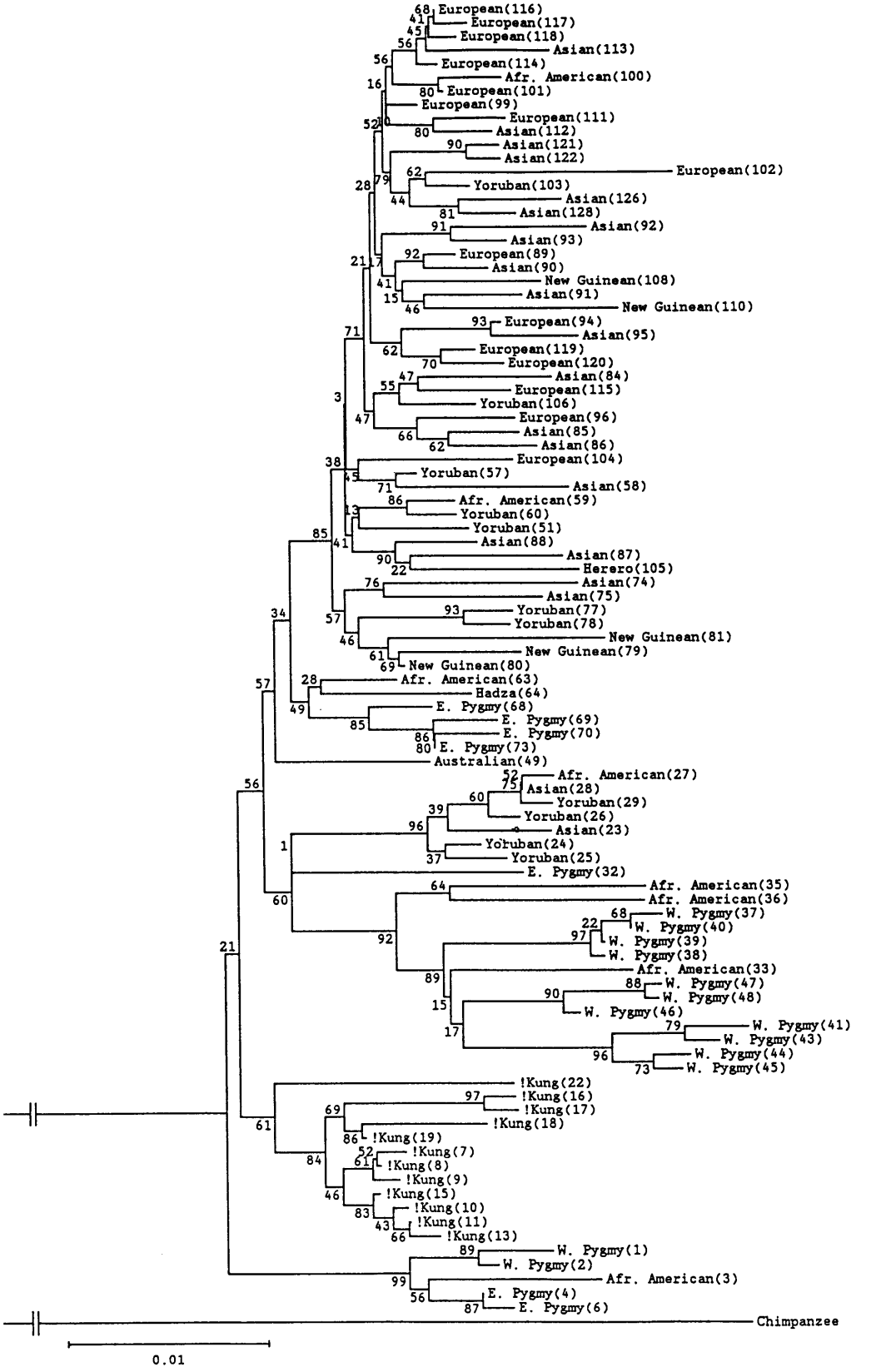
FIG. 5.—(A), Hypothetical true tree for four sequences. (B) and (C), Wrong trees for the same sequences.

b_4 are finite. This suggests that, even if a biased estimator (p_{ij}) is used, the ME method gives the correct topology, unless p_{ij}'s are large and the number of nucleotides examined is small. Even if the evolutionary rate is not the same for all branches, the inequality $E(S_W - S_T) > 0$ still holds for a wide variety of cases. For example, if $b_1 = 0.4$, $b_2 = 0.2$, and $b_4 = b_5 = 0.1$, then $E(S_W - S_T)$ is positive if $b_3 < 0.72$, and only when $b_3 > 0.72$ does $E(S_W - S_T)$ become negative.

Nevertheless, it is important to remember that $E(S_W - S_T)$ may become negative when biased estimates of d_{ij}'s are used. This is particularly so when the number of sequences used is large and the rate of nucleotide substitution varies with sequence. By contrast, if we use unbiased estimates of evolutionary distances, $E(S_W - S_T)$ is always positive. It is therefore advisable to use unbiased estimates of evolutionary distances whenever these are obtainable. [However, when (*a*) the number of nucleotide differences per site is small (say, $p_{ij} \leq 0.1$) for all sequence comparisons and (*b*) the evolutionary rate is nearly constant, p_{ij}'s often give slightly better results than do d_{ij}'s, apparently because the former have smaller variances than do d_{ij}'s (Saitou and Nei 1987).] Tajima (1993) recently showed that even Jukes and Cantor's (1969) estimator, i.e., $d_{ij} = -(3/4)\log_e[1 - 4p_{ij}/3]$, may give biased estimates when d_{ij} is large and the number of nucleotides examined is small. For this case he has presented another formula that gives unbiased estimates. For a general guideline about the distance measures to be used for phylogenetic inference, see the work of Nei (1991).

As mentioned earlier, equations (2) and (4) are very useful for estimating branch lengths when the number of sequences is large. Taking advantage of these equations, we constructed the ME tree for the data of human mtDNA obtained by Vigilant et al. (1991). Our strategy for finding the ME tree was as follows: We first generated the NJ tree and then searched for trees with smaller $\hat{S}$ values than that of the NJ tree ($\hat{S}_{NJ}$), examining all topologies whose topological distance (d_T) from the NJ tree was 2 or 4. If this search found trees with a smaller $\hat{S}$, we repeated the same search around the trees, and this process was continued until a tree with the smallest $\hat{S}$ was obtained.

FIG. 6.—ME tree for the human mtDNA sequences published by Vigilant et al. (1991). Ninety-five human and one chimpanzee (outgroup) sequences were used, as in the case of Tamura and Nei (1993). The proportion of different nucleotides was used as a distance measure. All sites containing ambiguous nucleotides or gaps were excluded from the analysis. [The same data set as that used by Tamura and Nei (1993) was used.] The number for each interior branch shows the significance level of the difference of the corresponding branch length from 0 (see Rzhetsky and Nei 1992*a*).


European(116)
European(117)
European(118)
Asian(113)
European(114)
Afr. American(100)
European(101)
European(99)
European(111)
Asian(112)
Asian(121)
Asian(122)
European(102)
Yoruban(103)
Asian(126)
Asian(128)
Asian(92)
Asian(93)
European(89)
Asian(90)
New Guinean(108)
Asian(91)
New Guinean(110)
European(94)
Asian(95)
European(119)
European(120)
Asian(84)
European(115)
Yoruban(106)
European(96)
Asian(85)
Asian(86)
European(104)
Yoruban(57)
Asian(58)
Afr. American(59)
Yoruban(60)
Yoruban(51)
Asian(88)
Asian(87)
Herero(105)
Asian(74)
Asian(75)
Yoruban(77)
Yoruban(78)
New Guinean(81)
New Guinean(79)
New Guinean(80)
Afr. American(63)
Hadza(64)
E. Pygmy(68)
E. Pygmy(69)
E. Pygmy(70)
E. Pygmy(73)
Australian(49)
Afr. American(27)
Asian(28)
Yoruban(29)
Yoruban(26)
Asian(23)
Yoruban(24)
Yoruban(25)
E. Pygmy(32)
Afr. American(35)
Afr. American(36)
W. Pygmy(37)
W. Pygmy(40)
W. Pygmy(39)
W. Pygmy(38)
Afr. American(33)
W. Pygmy(47)
W. Pygmy(48)
W. Pygmy(46)
W. Pygmy(41)
W. Pygmy(43)
W. Pygmy(44)
W. Pygmy(45)
!Kung(22)
!Kung(16)
!Kung(17)
!Kung(18)
!Kung(19)
!Kung(7)
!Kung(8)
!Kung(9)
!Kung(15)
!Kung(10)
!Kung(11)
!Kung(13)
W. Pygmy(1)
W. Pygmy(2)
Afr. American(3)
E. Pygmy(4)
E. Pygmy(6)
Chimpanzee
0.01

In this case we used the p_{ij} distance, because p_{ij}'s were all <0.05 and were nearly equal to d_{ij}'s for human sequences. Ninety-five human and one chimpanzee (outgroup) sequences were used in this study, as in the analysis by Tamura and Nei (1993), who constructed the NJ tree for the same set of data. (These authors also used p_{ij}'s.) The ME tree obtained is presented in figure 6. This tree has $\hat{S} = 0.652$, which is considerably smaller than the value (0.663) for the NJ tree given by Tamura and Nei (1993), though the difference is not statistically significant. In the ME tree all branch length estimates except 2 were positive, whereas in the NJ tree 10 branch lengths were negative.

Acknowledgments

We thank Koichiro Tamura, Tatsya Ota, Sudhir Kumar, and Blair Hedges for many helpful comments on the earlier versions of this paper. This work was supported by grants from NIH and NSF to M.N.

APPENDIX A

Derivation of Equations (2) and (4)

To derive equations (2) and (4), let us consider a particular interior branch of a bifurcating tree and group all sequences into four clusters A, B, C, and D as in tree (B) of figure 1. We then consider all possible linear estimators of the length (b) of the branch by introducing unknown coefficients. Such a family of estimators may be expressed as

$$\hat{b} = \sum_{i=1}^{K} \alpha_i d^i_{AC} + \sum_{i=1}^{L} \beta_i d^i_{BD} + \sum_{i=1}^{M} \gamma_i d^i_{AD} + \sum_{i=1}^{N} \delta_i d^i_{BC} - \sum_{i=1}^{P} \varepsilon_i d^i_{AB} - \sum_{i=1}^{Q} \zeta_i d^i_{CD}, \quad \text{(A1)}$$

where α_i's, β_i's, γ_i's, δ_i's, ε_i's, and ζ_i's are unknown coefficients and d^i_{AC} is the ith distance between two sequences, one belonging to cluster A and the other to cluster C. d^i_{BD}, d^i_{AD}, d^i_{BC}, d^i_{AB}, and d^i_{CD} are similarly defined. K, L, M, N, P, and Q stand for products $n_A n_C$, $n_B n_D$, $n_A n_D$, $n_B n_C$, $n_A n_B$, and $n_C n_D$, respectively, where n_A, n_B, n_C, and n_D are as defined in text. The first four terms in equation (A1) are weighted sums of all evolutionary distances that include the branch length b. The last two terms are for subtracting all branch lengths except b from the sums of evolutionary distances.

Let us now determine the coefficients α_{ij}'s, β_{ij}'s, etc., by considering the conditions that should be satisfied by any linear estimator and those required for OLS estimators. We first consider the case where all evolutionary distances are estimated without errors and all branch lengths except b are 0. In this case equation (A1) reduces to

$$b = \sum_{i=1}^{K} \alpha_i b + \sum_{i=1}^{L} \beta_i b + \sum_{i=1}^{M} \gamma_i b + \sum_{i=1}^{N} \delta_i b .$$

Therefore, we obtain

$$\sum_{i=1}^{K} \alpha_i + \sum_{i=1}^{L} \beta_i + \sum_{i=1}^{M} \gamma_i + \sum_{i=1}^{N} \delta_i = 1 . \quad \text{(A2)}$$

We next consider the case where all evolutionary distances are estimated without errors and where the length (b) of the interior branch (k) under consideration is 0. Denote the sum of branch lengths from sequence i in cluster A to branch k by A_i, and denote the sum of branch lengths from sequence j in cluster C to branch k by C_j.

Then, we have $d_{ij} = A_i + C_j$. Similarly, we can define B_i's and D_i's. Equation (A1) can then be written as

$$0 = \sum_{i=1}^{n_A}\sum_{j=1}^{n_C} \alpha_{ij}(A_i + C_j) + \sum_{i=1}^{n_B}\sum_{j=1}^{n_D} \beta_{ij}(B_i + D_j) + \sum_{i=1}^{n_A}\sum_{j=1}^{n_D} \gamma_{ij}(A_j + D_j) + \sum_{i=1}^{n_B}\sum_{j=1}^{n_C} \delta_{ij}(B_i + C_j) - \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \varepsilon_{ij}(A_i + B_j) - \sum_{i=1}^{n_C}\sum_{j=1}^{n_D} \zeta_{ij}(C_i + D_j) , \quad \text{(A3)}$$

where α_{ij}'s, β_{ij}'s, γ_{ij}'s, δ_{ij}'s, ε_{ij}'s, and ζ_{ij}'s are two-subscript versions of the coefficients in equation (A1). In the present case, α_{ij}'s, β_{ij}'s, γ_{ij}'s, δ_{ij}'s, ε_{ij}'s, and ζ_{ij}'s are independent of A_i, B_i, C_i, and D_i, and the above equation must hold even when only one of A_i's, B_i's, C_i's, and D_i's is positive. Therefore,

$$\sum_{i=1}^{n_A}\sum_{j=1}^{n_C} \alpha_{ij}A_i + \sum_{i=1}^{n_A}\sum_{j=1}^{n_D} \gamma_{ij}A_i - \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \varepsilon_{ij}A_i = 0 ; \quad \text{(A4)}$$

$$\sum_{i=1}^{n_B}\sum_{j=1}^{n_D} \beta_{ij}B_i + \sum_{i=1}^{n_B}\sum_{j=1}^{n_C} \delta_{ij}B_i - \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \varepsilon_{ij}B_j = 0 ; \quad \text{(A5)}$$

$$\sum_{i=1}^{n_A}\sum_{j=1}^{n_C} \alpha_{ij}C_j + \sum_{i=1}^{n_B}\sum_{j=1}^{n_C} \delta_{ij}C_j - \sum_{i=1}^{n_C}\sum_{j=1}^{n_D} \zeta_{ij}C_i = 0 ; \quad \text{(A6)}$$

$$\sum_{i=1}^{n_B}\sum_{j=1}^{n_D} \beta_{ij}D_j + \sum_{i=1}^{n_A}\sum_{j=1}^{n_D} \gamma_{ij}D_j - \sum_{i=1}^{n_C}\sum_{j=1}^{n_D} \zeta_{ij}D_j = 0 . \quad \text{(A7)}$$

Equation (A4) can be rewritten as

$$\sum_{i=1}^{n_A} A_i \sum_{j=1}^{n_C} \alpha_{ij} + \sum_{i=1}^{n_A} A_i \sum_{j=1}^{n_D} \gamma_{ij} - \sum_{i=1}^{n_A} A_i \sum_{j=1}^{n_B} \varepsilon_{ij} = 0 . \quad \text{(A8)}$$

Thus,

$$\sum_{j=1}^{n_C} \alpha_{ij} + \sum_{j=1}^{n_D} \gamma_{ij} - \sum_{j=1}^{n_B} \varepsilon_{ij} = 0 . \quad \text{(A9)}$$

Summation of both sides of equation (A9) over i from 1 to n_A yields

$$\sum_{i=1}^{n_A}\sum_{j=1}^{n_C} \alpha_{ij} + \sum_{i=1}^{n_A}\sum_{j=1}^{n_D} \gamma_{ij} = \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \varepsilon_{ij} , \quad \text{(A10)}$$

which becomes

$$\sum_{i=1}^{K} \alpha_i + \sum_{i=1}^{M} \gamma_i = \sum_{i=1}^{P} \varepsilon_i , \quad \text{(A11)}$$

if we use the single-subscript notation. Applying the same procedure to equations (A5), (A6), and (A7), we have

$$\sum_{i=1}^{L} \beta_i + \sum_{i=1}^{N} \delta_i = \sum_{i=1}^{P} \varepsilon_i \; ; \tag{A12}$$

$$\sum_{i=1}^{K} \alpha_i + \sum_{i=1}^{N} \delta_i = \sum_{i=1}^{Q} \zeta_i \; ; \tag{A13}$$

$$\sum_{i=1}^{L} \beta_i + \sum_{i=1}^{M} \gamma_i = \sum_{i=1}^{Q} \zeta_i \; . \tag{A14}$$

Therefore, we obtain

$$\sum_{i=1}^{P} \varepsilon_i = \sum_{i=1}^{Q} \zeta_i = \frac{1}{2} \; ; \tag{A15}$$

$$\sum_{i=1}^{K} \alpha_i = \sum_{i=1}^{L} \beta_i \; ; \tag{A16}$$

$$\sum_{i=1}^{M} \gamma_i = \sum_{i=1}^{N} \delta_i \; ; \tag{A17}$$

$$\sum_{i=1}^{K} \alpha_i + \sum_{i=1}^{M} \gamma_i = \sum_{i=1}^{L} \beta_i + \sum_{i=1}^{N} \delta_i = \sum_{i=1}^{K} \alpha_i + \sum_{i=1}^{N} \delta_i = \sum_{i=1}^{L} \beta_i + \sum_{i=1}^{M} \gamma_i = \frac{1}{2} \; . \tag{A18}$$

[Note that relations (A15)–(A18) are applicable to *any* linear estimator of branch length b, including the *generalized least-squares* estimator]. Using these relationships, we can exclude five unknown coefficients in equation (A1)—namely, β_L, γ_M, δ_N, ε_P, and ζ_Q. Therefore,

$$\begin{aligned} \hat{b} = & \sum_{i=1}^{K} \alpha_i d^i_{AC} + \sum_{i=1}^{L-1} \beta_i d^i_{BD} + \left\{ \sum_{i=1}^{K} \alpha_i - \sum_{i=1}^{L-1} \beta_i \right\} d^L_{BD} + \sum_{i=1}^{M-1} \gamma_i d^i_{AD} \\ & + \left\{ \frac{1}{2} - \sum_{i=1}^{K} \alpha_i - \sum_{i=1}^{M-1} \gamma_i \right\} d^M_{AD} + \sum_{i=1}^{N-1} \delta_i d^i_{BC} + \left\{ \frac{1}{2} - \sum_{i=1}^{K} \alpha_i - \sum_{i=1}^{N-1} \delta_i \right\} d^N_{BC} \\ & - \sum_{i=1}^{P-1} \varepsilon_i d^i_{AB} - \left\{ \frac{1}{2} - \sum_{i=1}^{P-1} \varepsilon_i \right\} d^P_{AB} - \sum_{i=1}^{Q-1} \zeta_i d^i_{CD} - \left\{ \frac{1}{2} - \sum_{i=1}^{Q-1} \zeta_i \right\} d^Q_{CD} \; . \end{aligned} \tag{A19}$$

Now we must find the coefficients α_i's, β_i's, γ_i's, δ_i's, ε_i's, and ζ_i's for the OLS estimator. To do this, we need some additional information. According to the theory of least squares (see Rao 1973, p. 220), the OLS method gives estimates with the minimum variance in the class of all linear estimators if all variances of the estimates of evolutionary distances are equal and all corresponding covariances are zero. Under this condition, the covariance matrix for evolutionary distances, **V**, can be expressed as

$$\mathbf{V} = v\mathbf{I} \; , \tag{A20}$$

where **I** is the unit matrix, and v is some constant. Let us assume that the matrix **V** in equation (A20) is indeed the covariance matrix for evolutionary distances. The variance of estimate $\hat{b}$ in expression (A19) then becomes

$$V(\hat{b}) = v\left\{\sum_{i=1}^{K}\alpha_i^2 + \sum_{i=1}^{L-1}\beta_i^2 + \left(\sum_{i=1}^{K}\alpha_i - \sum_{i=1}^{L-1}\beta_i\right)^2 + \sum_{i=1}^{M-1}\gamma_i^2 + \left(\frac{1}{2} - \sum_{i=1}^{K}\alpha_i - \sum_{i=1}^{M-1}\gamma_i\right)^2 + \sum_{i=1}^{N-1}\delta_i^2 + \left(\frac{1}{2} - \sum_{i=1}^{K}\alpha_i - \sum_{i=1}^{N-1}\delta_i\right)^2 - \sum_{i=1}^{P-1}\varepsilon_i^2 - \left(\frac{1}{2} - \sum_{i=1}^{P-1}\varepsilon_i\right)^2 - \sum_{i=1}^{Q-1}\zeta_i^2 - \left(\frac{1}{2} - \sum_{i=1}^{Q-1}\zeta_i\right)^2\right\}. \quad \text{(A21)}$$

Solving the system of equations

$$\partial V(\hat{b})/\partial\alpha_i = 0, \quad \partial V(\hat{b})/\partial\beta_i = 0, \quad \partial V(\hat{b})/\partial\gamma_i = 0,$$
$$\partial V(\hat{b})/\partial\delta_i = 0, \quad \partial V(\hat{b})/\partial\varepsilon_i = 0, \quad \partial V(\hat{b})/\partial\zeta_i = 0,$$

for α_i's, β_i's, γ_i's, δ_i's, ε_i's, and ζ_i's, we obtain

$$\alpha_i = \frac{1}{2}\frac{1}{n_A n_C}\frac{(n_B n_C + n_A n_D)}{(n_A + n_B)(n_C + n_D)}, \qquad \beta_i = \frac{1}{2}\frac{1}{n_B n_D}\frac{(n_B n_C + n_A n_D)}{(n_A + n_B)(n_C + n_D)},$$
$$\gamma_i = \frac{1}{2}\frac{1}{n_A n_D}\frac{(n_A n_C + n_B n_D)}{(n_A + n_B)(n_C + n_D)}, \qquad \delta_i = \frac{1}{2}\frac{1}{n_B n_C}\frac{(n_A n_C + n_B n_D)}{(n_A + n_B)(n_C + n_D)}, \quad \text{(A22)}$$
$$\varepsilon_i = \frac{1}{2}\frac{1}{n_A n_B}, \qquad \text{and } \zeta_i = \frac{1}{2}\frac{1}{n_C n_D}.$$

It can be shown that these values of α_i's, β_i's, γ_i's, δ_i's, ε_i's, and ζ_i's in equations (A22) indeed minimize the variance in expression (A21). Therefore, they give the OLS estimator of $\hat{b}$. Substituting equations (A22) into equation (A1), we can obtain equations (2) and (3) in the main text.

Equation (4) in the main text can be obtained from equation (2) by letting $n_C = n_D = 1$, $d_{AC} = d_{AD}$, $d_{BC} = d_{BD}$, and $d_{CD} = 0$ [see fig. 1(B) and (C)], i.e., by considering a special case, where clusters C and D become a single pending vertex, C.

In the same way one can derive equations for estimating branch lengths of an arbitrary *multifurcating* tree, but this is not our concern in this paper.

APPENDIX B

Computer Algorithms for Computing $\hat{b}$ and $V(\hat{b})$

In developing a computer program it is convenient to compute the estimate of a branch length as

$$\hat{b} = \sum_{i<j}\omega_{ij}d_{ij}, \quad \text{(B1)}$$

where b stands for either the length of an interior branch [as b in fig. 1(B)] or that of an exterior branch [as b in fig. 1(C)]. Coefficients ω_{ij}'s are computed in the following way: For an interior branch of any bifurcating tree, we have

$$\omega_{ij} = \begin{cases} -1/(2n_A n_B), & \text{if } i \in A \text{ and } j \in B, \\ (n_B n_C + n_A n_D)/[(n_A + n_B)(n_C + n_D)(2n_A n_C)], & \text{if } i \in A \text{ and } j \in C, \\ (n_A n_C + n_B n_D)/[(n_A + n_B)(n_C + n_D)(2n_A n_D)], & \text{if } i \in A \text{ and } j \in D, \\ (n_A n_C + n_B n_D)/[(n_A + n_B)(n_C + n_D)(2n_B n_C)], & \text{if } i \in B \text{ and } j \in C, \\ (n_B n_C + n_A n_D)/[(n_A + n_B)(n_C + n_D)(2n_B n_D)], & \text{if } i \in B \text{ and } j \in D, \\ -1/(2n_C n_D), & \text{if } i \in C \text{ and } j \in D, \\ 0, \text{ if both } i \text{ and } j \text{ belong to the same cluster}, \end{cases} \quad \text{(B2)}$$

where "$i \in A$" stands for "the sequence i belonging to cluster A." By analogy, in the case of an exterior branch [see fig. 1(C)],

$$\omega_{ij} = \begin{cases} 1/(2n_A), & \text{if } i = C \text{ and } j \in A, \\ 1/(2n_B), & \text{if } i = C \text{ and } j \in B, \\ -1/(2n_A n_B), & \text{if } i \in A \text{ and } j \in B, \\ 0, \text{ if both } i \text{ and } j \text{ belong to the same cluster (A or B)}. \end{cases} \quad \text{(B3)}$$

Obviously, the variance of a branch length estimate can be found as

$$V(\hat{b}) = \sum_{i<j} \omega_{ij}^2 V(d_{ij}) + 2 \sum_{ij<kl} \omega_{ij}\omega_{kl}\mathrm{Cov}(d_{ij}, d_{kl}), \quad \text{(B4)}$$

where $V(d_{ij})$ is the variance of the estimate of the distance between sequences i and j, and $\mathrm{Cov}(d_{ij}, d_{kl})$ is the covariance between estimates of distances d_{ij} and d_{kl}.

The estimate of S is obtained by

$$\hat{S} = \sum_{i<j} y_{ij} d_{ij}, \quad \text{(B5)}$$

where $y_{ij} = \sum_k \omega_{ij}^{(k)}$ and $\omega_{ij}^{(k)}$ stands for the ω_{ij} for the kth branch of the tree and $\omega_{ij}^{(k)}$'s are summed over all branches of the tree under consideration. The estimate of difference in S between two alternative trees (D) is then given by

$$\hat{D} = \hat{S}_1 - \hat{S}_2 = \sum_{i<j} (y_{ij}^{(1)} - y_{ij}^{(2)}) d_{ij}, \quad \text{(B6)}$$

and the variance of $\hat{D}$ is estimated by

$$V(\hat{D}) = \sum_{i<j} (y_{ij}^{(1)} - y_{ij}^{(2)})^2 V(d_{ij}) + 2 \sum_{ij<kl} (y_{ij}^{(1)} - y_{ij}^{(2)})(y_{kl}^{(1)} - y_{kl}^{(2)})\mathrm{Cov}(d_{ij}, d_{kl}). \quad \text{(B7)}$$

The computation of $\mathrm{Cov}(d_{ij}, d_{kl})$ has been described by Bulmer (1991) and, in more detail, by Rzhetsky and Nei (1992*a*).

Let us now briefly discuss the computer time and memory required for using equations (1) and (2). If the number of sequences used is n, then a straightforward application of equation (1) requires $O(n^4)$ time and $O(n^3)$ memory. By contrast, equation (2) formally requires $O(n^5)$ time and $O(n)$ computer memory. However, in the case of equation (2) we can store a tree topology as a list of partitions of

sequences for each branch. That is, for each branch of a tree the sequences are classified into either four (for the case of an interior branch) or three (for the case of an exterior branch) clusters, as shown in figure 1(B) and (C), respectively. (This method of storing tree topologies is also very convenient for generating neighboring topologies for a given tree.) Thus, computation of $\hat{b}_i$'s with equations (B1)–(B3) can be accomplished by using $O(n^2)$ time. Since there are $(2n - 3)$ branches, the total computer time required for estimating all branch lengths by equation (2) is $O(n^3)$. Therefore, both computer time and memory required are smaller for equation (2) than for equation (1).

APPENDIX C

Operations I and II Always Decrease β_i

Let us first consider a special case where an incongruent configuration is composed of four monochrome clusters [see fig. 3(E)]. Let us examine the change of β_i when configuration (E) is transformed into congruent configuration (F). Let n_A, n_B, n_C, and n_D be the number of the sequences in clusters A, B, C, and D, respectively. Then, using equation (14), one can obtain the following equation:

$$\beta_i^{(E)} - \beta_i^{(F)} = \beta_i^{(E)} = \frac{(n_A n_B + n_C n_D)}{(n_A + n_C)(n_B + n_D)} > 0 , \qquad (C1)$$

where $\beta_i^{(E)}$ and $\beta_i^{(F)}$ are the β_i's for trees (E) and (F) in figure 3, respectively. This proves our assertion that operation I decreases β_i in this special case.

Let us now consider the general case where operation I or II transforms an incongruent configuration into another incongruent configuration (see fig. C1). In figure C1, operation I transforms configuration (B) into configuration (A). To determine $\beta_i^{(B)} - \beta_i^{(A)}$, we must evaluate $\alpha_{j,i}$'s for configurations (A) and (B). In the following we drop subscript i of $\alpha_{j,i}$'s, since i refers to the same interior branch of the true tree.

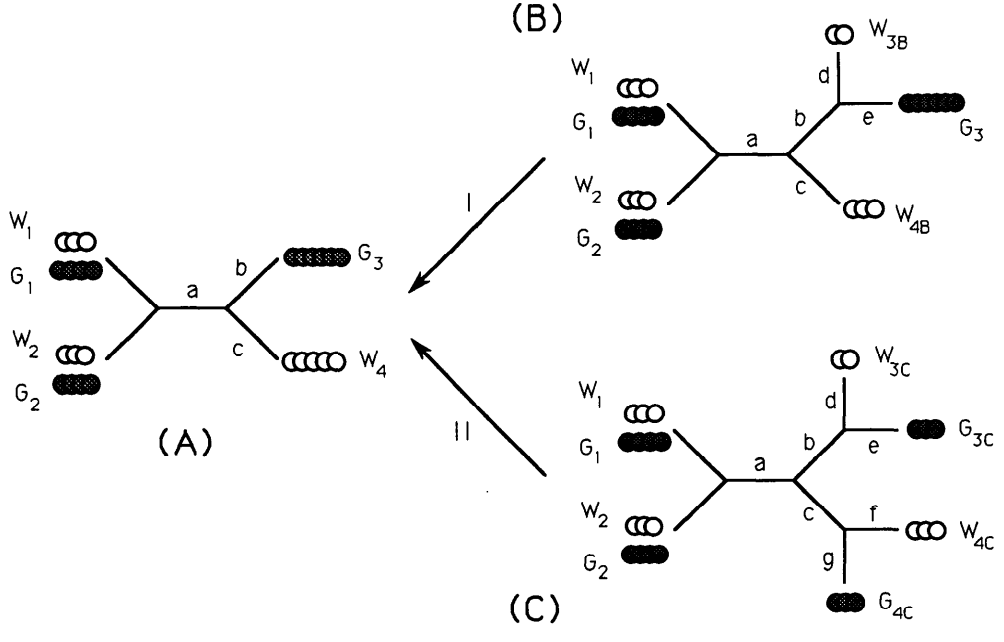


FIG. C1.—Transformation of an incongruent configuration into another (simpler) incongruent configuration, by operations I and II.

To compute β_i, we first note that α_a associated with branch a in configuration (G) or (H) in figure 3 is always 0, as we have already mentioned in the main text. Second, α_a and α_b associated with branches a and b in configuration (I) in figure 3 are given by

$$\alpha_a = W_C/(W_C + G_C), \qquad \alpha_b = G_C/(W_C + G_C), \tag{C2}$$

where W_C and G_C are the numbers of white and gray sequences in cluster C. Therefore, $\alpha_a + \alpha_b = 1$. Equation (C2) can be derived by using equation (14).

Using the above properties and noting the fact that the left-hand side of branch a is the same for both configuration (A) and configuration (B) in figure C1, we obtain

$$\begin{aligned}\beta_i^{(B)} - \beta_i^{(A)} &= \alpha_{aB} + \alpha_{bB} + \alpha_{cB} + \alpha_{dB} + \alpha_{eB} - \alpha_{aA} - \alpha_{bA} - \alpha_{cA} \\ &= \alpha_{aB} + \alpha_{bB} + \alpha_{cB} - \alpha_{aA},\end{aligned} \tag{C3}$$

where the second subscripts A and B for α refer to configurations (A) and (B), respectively. Here note that $\alpha_{bA} + \alpha_{cA} = 1$ and $\alpha_{dB} + \alpha_{eB} = 1$.

Application of equations (14) and (15) in the main text then gives

$$\alpha_{aA} = \frac{(\gamma_A - 1)W_1G_2 - \gamma_A G_1 W_2}{(W_1 + G_1)(W_2 + G_2)}, \tag{C4}$$

$$\text{with } \gamma_A = \frac{(W_1 + G_1)W_4 + (W_2 + G_2)G_3}{(W_1 + G_1 + W_2 + G_2)(G_3 + W_4)};$$

$$\alpha_{aB} = \frac{\gamma_{aB}(W_1G_2 - G_1W_2)G_3 + (G_1W_{3B} - W_1G_3)G_2}{(W_1 + G_1)(W_2 + G_2)(W_{3B} + G_3)}, \tag{C5}$$

$$\text{with } \gamma_{aB} = \frac{(W_1 + G_1)W_{4B} + (W_2 + G_2)(W_{3B} + G_3)}{(W_1 + G_1 + W_2 + G_2)(G_3 + W_4)};$$

$$\alpha_{bB} = \frac{(\gamma_{bB} - 1)(G_1 + G_2)}{(W_1 + G_1 + W_2 + G_2)}, \tag{C6}$$

$$\text{with } \gamma_{bB} = \frac{(W_1 + G_1 + W_2 + G_2)G_3 + W_{3B}W_{4B}}{(W_1 + G_1 + W_2 + G_2 + W_{4B})(W_{3B} + G_3)};$$

$$\alpha_{cB} = \frac{(G_1 + G_2)G_3}{(W_1 + G_1 + W_2 + G_2)(W_{3B} + G_3)}. \tag{C7}$$

In the present case we have $n = W_1 + G_1 + W_2 + G_2 + W_{3B} + G_3 + W_{4B}$, and

$$W_1 \geq 0,\ G_1 \geq 1,\ W_2 \geq 1,\ G_2 \geq 0,\ W_{3B} \geq 1,\ G_3 \geq 1,\ W_{4B} \geq 1,$$

or

$$W_1 \geqslant 1,\ G_1 \geqslant 0,\ W_2 \geqslant 0,\ G_2 \geqslant 1,\ W_{3B} \geqslant 1,\ G_3 \geqslant 1,\ W_{4B} \geqslant 1\ .$$

Therefore, the minimum value of equation (C3) is obtained in the following two cases:

$$1)\ G_1 = G_3 = W_{4B} = 1, \qquad G_2 = 0, \qquad W_{3B} \simeq (W_1 + W_2)/2 \simeq (n-3)/3$$

or

$$2)\ G_1 = 0, \qquad G_2 = G_3 = W_{4B} = 1, \qquad W_{3B} \simeq (W_1 + W_2)/2 \simeq (n-3)/3\ .$$

In both cases equation (C3) reduces to

$$\mathrm{Min}[\beta_i^{(B)} - \beta_i^{(A)}] = \frac{27(n-2)}{n^2(4n-6)}\ . \tag{C8}$$

Obviously, for $n \geqslant 3$, expression (C8) is always positive. Therefore, operation I always decreases the value of β_i.

Using the same argument, we can study the effect of transformation of configuration (C) into (A) in figure C1 (operation II) on β_i. For configuration (C), we have

$$\begin{aligned}\beta_i^{(C)} - \beta_i^{(A)} &= \alpha_{aC} + \alpha_{bC} + \alpha_{cC} + \alpha_{dC} + \alpha_{eC} + \alpha_{fC} + \alpha_{gC} - \alpha_{aA} - \alpha_{bA} - \alpha_{cA} \\ &= \alpha_{aC} + \alpha_{bC} + \alpha_{cC} + 1 - \alpha_{aA}\ .\end{aligned} \tag{C9}$$

Here $\alpha_{dC} + \alpha_{eC} = 1$ and $\alpha_{fC} + \alpha_{gC} = 1$. By analogy with equations (C4)–(C7) we obtain

$$\alpha_{aC} = \frac{-\gamma_{aC}(W_1G_2 - G_1W_2)(W_{3C}G_{4C} - G_{3C}W_{4C}) + (W_1G_{3C} - G_1W_{3C})(W_2G_{4C} - G_2W_{4C})}{(W_1 + G_1)(W_2 + G_2)(W_{3C} + G_{3C})(W_{4C} + G_{4C})}, \tag{C10}$$

$$\text{with } \gamma_{aC} = \frac{(W_1 + G_1)(W_{4C} + G_{4C}) + (W_2 + G_2)(W_{3C} + G_{3C})}{(W_1 + G_1 + W_2 + G_2)(G_3 + W_4)}\ ;$$

$$\alpha_{bC} = \frac{(\gamma_{bC} - 1)W_{4C}(G_1 + G_2) - \gamma_{bC}G_{4C}(W_1 + W_2)}{(W_1 + G_1 + W_2 + G_2)(W_{4C} + G_{4C})}, \tag{C11}$$

$$\text{with } \gamma_{bC} = \frac{(W_1 + G_1 + W_2 + G_2)G_{3C} + (W_{4C} + G_{4C})W_{3C}}{(W_1 + G_1 + W_2 + G_2 + W_{4C} + G_{4C})(W_{3C} + G_{3C})}\ ;$$

$$\alpha_{cC} = \frac{(\gamma_{cC} - 1)W_{3C}(G_1 + G_2) - \gamma_{cC}G_{3C}(W_1 + W_2)}{(W_1 + G_1 + W_2 + G_2)(W_{3C} + G_{3C})}, \tag{C12}$$

$$\text{with } \gamma_{cC} = \frac{(W_1 + G_1 + W_2 + G_2)G_{4C} + (W_{3C} + G_{3C})W_{4C}}{(W_1 + G_1 + W_2 + G_2 + W_{3C} + G_{3C})(W_{4C} + G_{4C})}\ .$$

In the present case we have $n = W_1 + G_1 + W_2 + G_2 + W_{3C} + G_{3C} + W_{4C} + G_{4C}$ and

$$W_1 \geq 0,\ G_1 \geq 1,\ W_2 \geq 1,\ G_2 \geq 0,\ W_{3C} \geq 1,\ G_{3C} \geq 1,\ W_{4C} \geq 1, \text{ and } G_{4C} \geq 1\ ,$$

or

$$W_1 \geq 1,\ G_1 \geq 0,\ W_2 \geq 0,\ G_2 \geq 1,\ W_{3C} \geq 1,\ G_{3C} \geq 1,\ W_{4C} \geq 1, \text{ and } G_{4C} \geq 1\ .$$

Therefore, the minimum value of equation (C9) is obtained in the following four cases:

$$(1)\ G_1 = W_2 = G_{3C} = W_{4C} = 1, \quad W_1 = G_2 = 0, \quad W_{3C} \simeq G_{4C} \simeq (n-4)/2\ ,$$

$$(2)\ G_1 = W_2 = W_{3C} = G_{4C} = 1, \quad W_1 = G_2 = 0, \quad G_{3C} \simeq W_{4C} \simeq (n-4)/2\ ,$$

$$(3)\ G_1 = W_2 = 0, \quad W_1 = G_2 = G_{3C} = W_{4C} = 1, \quad W_{3C} \simeq G_{4C} \simeq (n-4)/2\ ,$$

$$(4)\ G_1 = W_2 = 0, \quad W_1 = G_2 = W_{3C} = G_{4C} = 1, \quad G_{3C} \simeq W_{4C} \simeq (n-4)/2\ .$$

In each of these cases the minimum value of $\beta_i^{(C)} - \beta_i^{(A)}$ is given by

$$\text{Min}[\beta_i^{(C)} - \beta_i^{(A)}] = \frac{2(5n^2 - 30n + 48)}{n^3 - 2n^2 - 4n - 8}\ . \qquad \text{(C13)}$$

Expression (C13) is always positive when $n \geq 4$, so operation II also always decreases the value of β_i.

LITERATURE CITED

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

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.

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

NEI, M. 1972. Genetic distance between populations. Am. Nat. **106**:283–292.

———. 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.

RAO, C. R. 1973. Linear statistical inference and its applications, 2d ed. John Wiley & Sons, New York.

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

———. 1992*b*. Statistical properties of the ordinary least-squares, generalized least-squares and minimum-evolution methods of phylogenetic inference. J. Mol. Evol. **35**:367–475.

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

TAJIMA, F. 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **10**:677–688.

TAMURA, K., and M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10:** 512–526.

VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES, and A. C. WILSON. 1991. African populations and the evolution of human mitochondrial DNA. Science **253:**1503–1507.

MICHAEL BULMER, reviewing editor

Received November 19, 1992; revision received February 25, 1993

Accepted March 8, 1993