J Mol Evol (1986) 24:189–204

Journal of Molecular Evolution

The Number of Nucleotides Required to Determine the Branching Order of Three Species, with Special Reference to the Human-Chimpanzee-Gorilla Divergence

Naruya Saitou and Masatoshi Nei

Center for Demographic and Population Genetics, University of Texas Health Science Center at Houston, P.O. Box 20334, Houston, Texas 77225, USA

Summary. A mathematical theory for computing the probabilities of various nucleotide configurations among related species is developed, and the probability of obtaining the correct tree (topology) from nucleotide sequence data is evaluated using models of evolutionary trees that are close to the tree of mitochondrial DNAs from human, chimpanzee, gorilla, orangutan, and gibbon. Special attention is given to the number of nucleotides required to resolve the branching order among the three most closely related organisms (human, chimpanzee, and gorilla). If the extent of DNA divergence is close to that obtained by Brown et al. for mitochondrial DNA and if sequence data are available only for the three most closely related organisms, the number of nucleotides (m*) required to obtain the correct tree with a probability of 95% is about 4700. If sequence data for two outgroup species (orangutan and gibbon) are available, m* becomes about 2600–2700 when the transformed distance, distance-Wagner, maximum parsimony, or compatibility method is used. In the unweighted pair-group method, m* is not affected by the availability of data from outgroup species. When these five different tree-making methods, as well as Fitch and Margoliash's method, are applied to the mitochondrial DNA data (1834 bp) obtained by Brown et al. and by Hixson and Brown, they all give the same phylogenetic tree, in which human and chimpanzee are most closely related. However, the trees considered here are "gene trees," and to obtain the correct "species tree," sequence data for several independent loci must be used.

Key words: Molecular phylogeny — Nucleotide substitution — Tree-making methods — Hominoid evolution — Mitochondrial DNA

Offprint requests to: M. Nei

Introduction

During the last two decades, the evolutionary relationships of hominoid species have been studied extensively using molecular data. It is now generally accepted among investigators of molecular evolution that gibbons and orangutans diverged from the human line substantially earlier than chimpanzees and gorillas did. However, no consensus has been reached about the branching order among humans, chimpanzees, and gorillas (Ferris et al. 1981; Brown et al. 1982; Templeton 1983; Sibley and Ahlquist 1984; Bianchi et al. 1985; Hasegawa et al. 1985; Nei et al. 1985; Ueda et al. 1985; Hixson and Brown 1986; Koop et al. 1986). One of the difficulties in resolving this problem is that the three species are so closely related that a large amount of molecular data is required.

Sibley and Ahlquist's (1984) study was based on a hybridization experiment with a large amount of "single-copy" DNA, and their conclusion seems to be quite reasonable. However, since the measurement of $\Delta T_{50}H$ is subject to a rather large experimental error, it is advisable to examine the problem

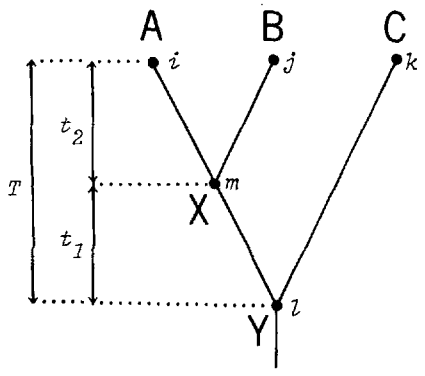

Fig. 1. Model tree for three nucleotide sequences from species A, B, and C. See text for details

using DNA sequences. One of the most extensive sets of sequence data available is Brown et al.'s (1982) for mitochondrial DNA; they sequenced 896 nucleotides. Nei et al.'s (1985) reanalysis of the data, however, indicates that this number of nucleotides is too small to allow discrimination among the three alternative branching orders possible for the three species. How many nucleotides are then required to resolve the branching order among humans, chimpanzees, and gorillas?

The purpose of this paper is to examine this problem. More specifically, we will study the probability of obtaining the correct topology (branching order) for a given number of nucleotides, considering DNA sequences whose divergences are similar to those of mitochondrial DNA. Various methods of constructing phylogenetic trees will be considered, since the probability of obtaining the correct topology is not the same for all tree-making methods. The tree-making methods considered here are the unweighted pair-group method (UPGMA) (Sneath and Sokal 1973), the distance-Wagner method (Farris 1972), Fitch and Margoliash's (1967) method, the transformed distance method (Farris 1977; Klotz and Blanken 1981; Li 1981), the maximum parsimony method (Eck and Dayhoff 1966; Fitch 1977), and the compatibility method (Le Quesne 1969). The first three methods estimate both the topology and branch lengths of a tree, whereas the other three obtain only the topology. In the first four methods, genetic distances are computed for all species pairs, and an evolutionary tree is constructed using information on genetic distances. In the remaining two methods, a tree is constructed by comparing nucleotide sequences site by site.

Before going into a detailed discussion, we should mention that two types of errors occur in the construction of phylogenetic trees: topological errors and branch length errors (Tateno et al. 1982). In this paper, we will consider only topological errors. We will also consider the case where only one DNA sequence (gene) is studied from each species and assume that the evolutionary pathway of the genes studied (gene tree) is identical with that of the species considered (species tree), though this assumption does not always hold in the presence of polymorphism within species (see Discussion). When the number of nucleotides examined is small, however, a phylogenetic tree reconstructed from a set of genes may not always give the true phylogeny of the genes even if we disregard the discrepancy between the gene tree and the species tree. Our main concern in this paper will be the accuracy of a reconstructed tree under the assumption that the true gene tree is identical with the species tree.

We will be concerned primarily with the resolution of the branching order among humans, chimpanzees, and gorillas. However, the resolution of the branching order among three species is a fundamental problem in phylogenetic studies, and the results obtained here will be applicable to any similar situation.

Theoretical Basis

Although we do not know the real evolutionary relationship of humans, chimpanzees, and gorillas, it can be represented by the diagram in Fig. 1. Here A, B, and C each stand for one of the three species, whereas T and t_2 are the times since divergence between A and C and between A and B, respectively, with $t_1 = T - t_2$. In reality, the true relationship is not known, and is inferred from the nucleotide sequences of the genes sampled from the three species. If the rate of nucleotide substitution is constant over time, we should be able to determine the relationship by examining a large number of nucleotides. If the number of nucleotides examined is small, the tree inferred is not necessarily the correct one because of sampling error. Our task is to determine the relationship between the number of nucleotides examined and the probability of obtaining the correct tree (or topology). To compute this probability, we must know the probabilities of different nucleotide configurations among the species studied.

Probabilities of Nucleotide Configurations

The basic units of information required for tree making are the nucleotide differences at each nucleotide site among the species compared; the relative frequencies of different nucleotide configurations among all sites determine the branching pattern of the species. Let i, j, and k be the nucleotides for species A, B, and C, respectively, at a given nucleotide site (see Fig. 1). Here i, j, and k are any of the four nucleotides A, T, C, and G. In the case of three

species, there are five different nucleotide configurations; they are listed in Table 1. In general, the number (c) of possible nucleotide configurations for n species is given by

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

To compute the probability of obtaining the correct tree for a given number of nucleotides examined, we must know the relative frequencies of the nucleotide configurations. We use two models of nucleotide substitution to determine the relative frequencies. The first model is that of Jukes and Cantor (1969), in which all nucleotides (A, T, C, and G) change to one another with equal probability. In this model (the one-parameter model), the probability that the nucleotide for a given site at time t is identical with that at time 0 is

$$P_{ii}(t) = \tfrac{1}{4} + \tfrac{3}{4} \exp(-\tfrac{4}{3}\lambda t) \quad (2a)$$

and the probability that nucleotide i at time 0 changes to nucleotide j at time t is

$$P_{ij}(t) = \tfrac{1}{4} - \tfrac{1}{4} \exp(-\tfrac{4}{3}\lambda t) \quad (2b)$$

where λ is the rate of nucleotide substitution per site per year (see, e.g., Nei and Tajima 1985).

The second model is Kimura's (1980) two-parameter model, where transitional nucleotide changes (A $\rightleftharpoons$ G and T $\rightleftharpoons$ C) are assumed to occur with a frequency different from that for transversional nucleotide changes (all other changes). This model is more appropriate to mitochondrial DNA (mtDNA) than the first model, because in mtDNA, transitional changes are known to occur at a much higher rate than transversional changes. In this model, $P_{ii}(t)$ and $P_{ij}(t)$ are given by

$$P_{ii}(t) = \tfrac{1}{4} + \tfrac{1}{4} \exp(-4\beta t) + \tfrac{1}{2} \exp[-2(\alpha + \beta)t] \quad (3a)$$

$$P_{ij}(t) = \tfrac{1}{4} + \tfrac{1}{4} \exp(-4\beta t) - \tfrac{1}{2} \exp[-2(\alpha + \beta)t] \text{(transition)} \quad (3b)$$

$$P_{ij}(t) = \tfrac{1}{4} - \tfrac{1}{4} \exp(-4\beta t) \quad \text{(transversion)} \quad (3c)$$

where α and β are the rates of transitional and transversional substitutions, respectively, and λ is given by $\alpha + 2\beta$. In applying this model, we assume $\alpha/\beta = 20$, following Brown et al.'s (1982) observation. Therefore, $\beta = (1/22)\lambda$ or $\alpha = (10/11)\lambda$. Thus, if we know λ, we can easily determine α and β.

It should be noted that the above two models are not always realistic, particularly when λT is large. The effect of violation of the underlying assumptions of the two models will be discussed later.

Using the above two models, we can now evaluate the probabilities of obtaining different nucleotide configurations. For example, the probability of having nucleotides i, j, and k at a given nucleotide site

Table 1. Nucleotide configurations for three species

Configuration	Species[a]			Probability (F_i)[b]		Observed no. of sites
	A	B	C	R = 0.2	R = 0.8	
C_1	i	i	i	0.870	0.897	m_1
C_2	i	i	j	0.054	0.083	m_2
C_3	i	j	i	0.036	0.009	m_3
C_4	j	i	i	0.036	0.009	m_4
C_5	i	j	k	0.004	0.001	m_5

[a] i, j, and k are three different nucleotides

[b] $\lambda T = 0.05$ and the one-parameter model is used. $R = t_t/T$ (see text)

for species A, B, and C in the evolutionary tree of Fig. 1 is

$$f(i, j, k) = \sum_{\ell} g_\ell [P_{\ell k}(t_1 + t_2) \cdot \sum_{m} \{P_{\ell m}(t_1) P_{mi}(t_2) P_{mj}(t_2)\}] \quad (4)$$

where g_ℓ is the probability of nucleotide ℓ occurring at this site. We assume that the nucleotide frequencies in DNA sequences are at equilibrium, so that $g_\ell = 0.25$ for all nucleotides in the above two models. Therefore, f(i, j, k) can be obtained if we specify t_1 and t_2.

We mentioned above that there are five different nucleotide configurations for three species (Table 1). The expected frequencies of the five configurations (F_i) can be obtained by summing the appropriate values of f(i, j, k). The values of F_i for $\lambda T = 0.05$ and the one-parameter model are given in Table 1. The expected frequencies of nucleotide configurations with more than three species can be obtained in the same way.

In practice, only a limited number of nucleotides are used for the construction of phylogenetic trees. Let m be the total number of nucleotides examined for a gene and assume that the rate of nucleotide substitution is the same for all sites. The probability of observing configuration 1 (C_1) at m_1 nucleotide sites, C_2 at m_2 sites, . . . , and C_k at m_k nucleotide sites ($m_1 + m_2 + \ldots + m_k = m$) is then given by

$$P = \frac{m!}{m_1!\, m_2! \ldots m_k!} \prod_{i=1}^{k} F_i^{m_i} \quad (5)$$

where k is the number of different possible configurations. We call a given set of nucleotide configurations among m nucleotide sites a compound nucleotide configuration.

Probability of Obtaining the Correct Tree

We are now in a position to compute the probability of obtaining the correct tree for a given number of

nucleotides examined. This probability depends not only on the tree-making method but also on the availability of outgroup species. Information on outgroup species is important for all the methods except UPGMA. In the present case, orangutans and gibbons can be used as outgroup species if DNA sequence data are available. In the following, we consider the cases of three species (no outgroup species), four species (one outgroup species), and five species (two outgroup species) separately.

Case 1. Three Species

The only tree-making method that gives a rooted tree is UPGMA. In this method, the topology and branch lengths are uniquely determined from distance values. Let d_{AB}, d_{AC}, and d_{BC} be the distances between species A and B, A and C, and B and C, respectively. In the following, we measure the distances in terms of nucleotide differences instead of the number of nucleotide substitutions, which may be estimated by Jukes and Cantor's (1969) method, because the probability of obtaining the correct topology is nearly the same for the two different distance measures (N. Saitou and M. Nei, unpublished). In this case, the genetic distances as measured by the number of nucleotide differences per DNA sequence can be expressed as

$$\begin{aligned} d_{AB} &= m_3 + m_4 + m_5 = m - m_1 - m_2 \\ d_{AC} &= m_2 + m_4 + m_5 = m - m_1 - m_3 \\ d_{BC} &= m_2 + m_3 + m_5 = m - m_1 - m_4 \end{aligned} \quad (6)$$

where m_i is the observed number of sites with the i-th configuration.

In UPGMA, the two species with the smallest distance between them are clustered first. This cluster is then regarded as a single species, and a new set of genetic distances is computed. The two species showing the smallest distance in this new distance matrix are again clustered. This process is continued until all the species are clustered into a single tree. This method produces a rooted rather than an unrooted tree. It is therefore clear that the condition for obtaining the correct tree for the case of three species in Fig. 1 is

$$d_{AB} < d_{AC} \quad \text{and} \quad d_{AB} < d_{BC} \quad (7)$$

This is equivalent to the condition $m_2 > m_3$ and $m_2 > m_4$. We have already derived a formula for computing the probability of each compound nucleotide configuration. Therefore, if we collect all probabilities of compound configurations satisfying the above condition, the probability of obtaining the correct tree is obtained.

The other distance matrix methods (Fitch and Margoliash's method, the distance-Wagner method, and the transformed distance method) are intended to construct unrooted trees [see Nei (in press) for the computational procedures for these methods]. Since there is only one unrooted topology for three species, we must place the root. The root can be placed at the midpoint of the pathway that connects the two most divergent species. In the case of three species, this procedure again gives condition (7) for obtaining the correct rooted tree. Therefore, the probability of obtaining the correct rooted tree is the same for all distance matrix methods. The maximum parsimony and the compatibility methods also obtain unrooted trees. Therefore, we can apply the same method of placing the root.

Let us now examine the relationship between the probability of obtaining the correct tree (P_c) and the number of nucleotides examined by using the two nucleotide substitution models mentioned earlier. When the number of nucleotides is relatively small ($m \leq 100$), it is possible to consider all possible combinations of $m_1, m_2, \ldots, m_k$ and to evaluate P_c using Eq. (5). When m is large, the number of different compound configurations becomes astronomical. Therefore, we used the following simulation method to evaluate P_c for $m > 100$: Using Eq. (5) and pseudorandom numbers, we first determined $m_1, m_2, \ldots, m_k$ for a given sequence length (one replication). We then computed the number of nucleotide differences (distances) for all pairs of species using Eq. (6). From these distances and condition (7), we determined whether or not the tree reconstructed was correct. This computation was repeated 10,000 times, and the relative frequency of obtaining the correct topology was used as an approximation to P_c.

The relationship between the probability of obtaining the correct tree and the number of nucleotides depends not only on the tree-making method but also on λT and the ratio $R = t_1/T$ in Fig. 1. Therefore, we first examined the relationships between P_c and R for various values of λT for the case $m = 100$, using the one-parameter model of nucleotide substitution. The results are presented in Fig. 2. As expected, P_c increases as R increases, but if λT is small, P_c increases very slowly and does not reach 1 even when $R = 1$. The reason for this is that when λT and m are small, there may be no nucleotide sites where substitution has occurred in the branch between nodes X and Y (see Fig. 1), in which case no trees can be constructed. When λT is much smaller than 1, the probability of there being no nucleotide differences in a sequence of m nucleotides is $P_{ii}^m(T) = [\frac{1}{4} + (\frac{3}{4})\exp(-4\lambda T/3)]^m$. When $R = 1$, the condition $d_{AB} < d_{AC}$ and $d_{AB} < d_{BC}$ reduces to $0 < d_{AC} (= d_{BC})$. This condition is violated if there is no substitution in either the A or the C branch. Therefore, the probability of obtaining the

correct tree for R = 1 is given by $1 - P_{ii}^{2m}(T)$. This value is 0.18 for $\lambda T = 0.001$ and 0.63 for $\lambda T = 0.005$, in agreement with Fig. 2.

Figure 2 shows that P_c generally increases as λT increases for a given value of R. In the case of R = 0.5, the P_c values for $\lambda T = 0.05$, 0.1, and 0.5 are 0.865, 0.957, and 0.997, respectively (curves D, E, and F in Fig. 2). If λT exceeds 0.5, however, P_c becomes smaller than for $\lambda T = 0.5$ for all values of R. In the case of humans, chimpanzees, and gorillas, R is probably less than 0.2 (Sibley and Ahlquist 1984 and Fig. 10). Therefore, P_c is smaller than 0.6 even if $\lambda T = 0.5$. This indicates that when the number of nucleotides examined is as small as 100, the probability of obtaining the correct topology cannot be very high, whatever the value of λT. Reanalyzing Brown et al.'s (1982) sequence data for mitochondrial DNA, Nei et al. (1985) estimated that the λT value for humans and gorillas is about 0.05. If this estimate is correct, the probability that the evolutionary tree constructed from about 100 nucleotides will be erroneous is more than 50%.

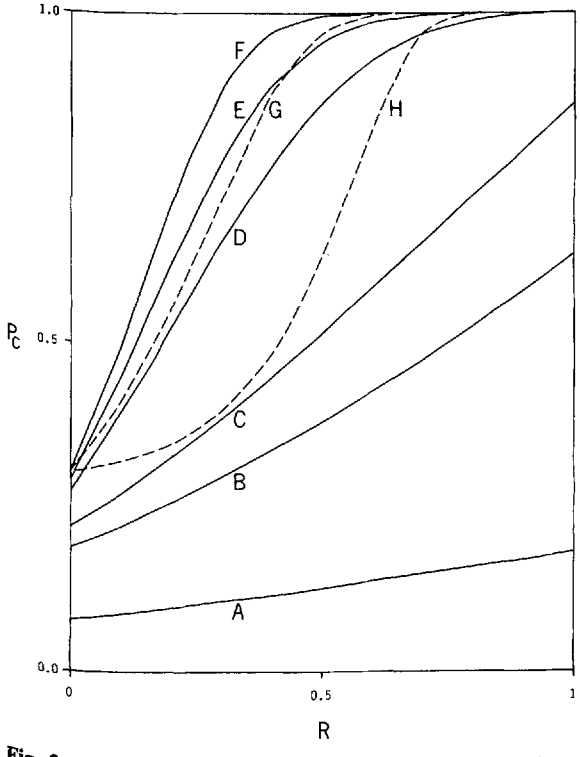

Fig. 2. Probability (P_c) of obtaining the correct topology for three species for various values of λT (A, 0.001; B, 0.005; C, 0.010; D, 0.050; E, 0.100; F, 0.500; G, 1.000; H, 2.000). The one-parameter model is used. The number (m) of nucleotides compared is 100

Figure 3 shows the relationships between P_c and m for various values of R for the case of $\lambda T = 0.05$. When R is as high as 0.8, P_c rapidly increases with increasing m. In this case, the number of nucleotides (m*) required for obtaining the correct tree with a probability of 0.95 is less than 100. This number increases rapidly as R decreases. The values of m* for R = 0.2, 0.15, and 0.1 are about 1100, 2000, and 4200, respectively. As mentioned earlier, the R value for the human-chimpanzee-gorilla divergence is likely to be smaller than 0.2. Therefore, a large number of nucleotides is required to resolve the branching order. Brown et al. (1982) sequenced 896 nucleotides from a portion of mitochondrial DNA. This number is apparently too small to resolve the problem of branching order.

The above conclusion was derived using the one-parameter model of nucleotide substitution. However, in the case of three species, essentially the same

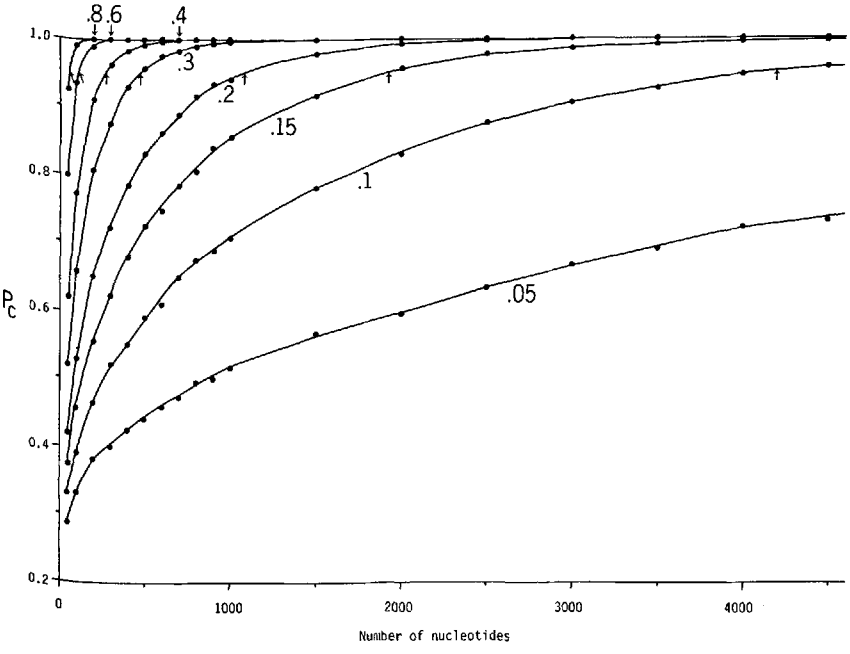

Fig. 3. Probabilities (P_c) of obtaining the correct topology for three species. Arrows indicate the number of nucleotides required for obtaining the correct topology with a probability of 0.95. Different curves represent different values of $R = t_1/T$ as marked. The one-parameter model is used

conclusion is obtained using the two-parameter model mentioned earlier (see discussion of UPGMA in the following section).

Case 2. Four Species

We consider the case where one outgroup species is known (D in Fig. 4). Our primary goal is still to find the branching order among species A, B, and C. When DNA sequence data are available from four species, there are 15 different types of nucleotide configurations (Table 2). The distance d_{ij} between species i and j can be expressed in terms of the observed numbers of these nucleotide configurations, as in the case of three species. The probabilities of these configurations for the two different models of nucleotide substitution are given in Table 2 for the case of tree A in Fig. 4.

The availability of outgroup species does not affect the tree constructed by UPGMA, because that method does not use information on outgroup species. In other tree-making methods, however, the availability of outgroup species substantially improves the accuracy of the reconstructed tree.

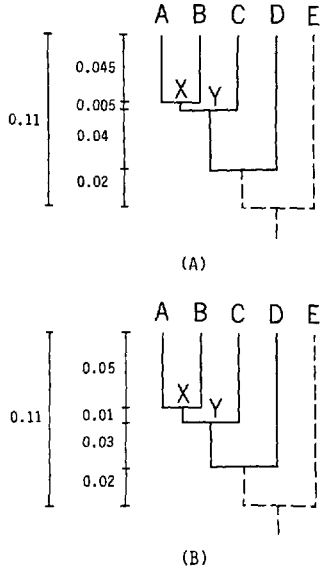

Fig. 4A, B. Two phylogenetic trees used for the cases of four and five species. A, B, C, any of human, chimpanzee, and gorilla; D, orangutan; E, gibbon

Transformed Distance Method. In this method, the original distance (d_{ij}) between species i and j is transformed into a new distance by the following equation, and the transformed distance (d'_{ij}) is used for tree making:

$$d'_{ij} = (d_{ij} - d_{iD} - d_{jD})/2 + c$$

where D refers to the outgroup species and c is a constant to prevent d_{ij} from becoming negative [see Farris (1977) or Nei (in press) for the details of the procedure]. The topology of a tree is then constructed from the values of d'_{AB}, d'_{AC}, and d'_{BC} using UPGMA. The condition for obtaining the correct tree is therefore given by the inequalities $d'_{AB} < d'_{AC}$ and $d'_{AB} < d'_{BC}$ or

$$d_{AB} + d_{CD} < d_{AC} + d_{BD} \qquad (8)$$

$$d_{AB} + d_{CD} < d_{AD} + d_{BC}$$

Table 2. Nucleotide configurations for four species

Configuration	Species[a]				Method[b]			Probability[c]		Observed no. of sites
	A	B	C	D	MP	TD	UP	One-parameter model	Two-parameter model	
C_1	i	i	j	j	+	+	+	0.0061	0.0093	m_1
C_2	i	j	i	j	+	+	+	0.0022	0.0052	m_2
C_3	i	j	j	i	+	+	+	0.0022	0.0052	m_3
C_4	i	i	j	k	−	+	+	0.0040	0.0011	m_4
C_5	i	j	i	k	−	+	+	0.0032	0.0008	m_5
C_6	j	i	i	k	−	+	+	0.0032	0.0008	m_6
C_7	j	k	i	i	−	+	−	0.0014	0.0004	m_7
C_8	j	i	k	i	−	+	−	0.0013	0.0003	m_8
C_9	i	j	k	i	−	+	−	0.0013	0.0003	m_9
C_{10}	i	i	j	i	−	−	+	0.0390	0.0391	m_{10}
C_{11}	i	j	i	i	−	−	+	0.0349	0.0349	m_{11}
C_{12}	j	i	i	i	−	−	+	0.0349	0.0349	m_{12}
C_{13}	i	i	i	j	−	−	−	0.1034	0.1006	m_{13}
C_{14}	i	j	k	ℓ	−	−	−	0.0002	0.0000	m_{14}
C_{15}	i	i	i	i	−	−	−	0.7626	0.7670	m_{15}

[a] i, j, k, and ℓ represent different nucleotides

[b] Pluses and minuses stand for the nucleotide configurations that are used and not used, respectively, under the conditions of the tree-making methods: MP, maximum parsimony method; TD, transformed distance method; UP, unweighted pair-group method

[c] Tree A of Fig. 4 is used

Distance-Wagner Method. In this method, the two most closely related species are clustered first, and a third species that is most closely related to this cluster is then joined with minimum branch lengths. This procedure is continued until all species are clustered into a single tree. In the case of the four species in Fig. 4, species D is usually the last to join the cluster because it is an outgroup species. Let us assume that D is indeed the last to join and consider the condition for obtaining the correct tree. It is clear from Fig. 5 that to obtain the correct tree, one must connect D to the branch between C and W. Therefore, we have

$$d_{D3,Z} < d_{D1,X} \quad \text{and} \quad d_{D3,Z} < d_{D2,Y} \tag{9}$$

where the $d_{i,j}$ represent the branch lengths of the tree in Fig. 5. According to Farris (1972), these branch lengths are estimated by

$$\begin{aligned} d_{D1,X} &= (d_{DW} + d_{AD} - d_{AW})/2 \\ d_{D2,Y} &= (d_{DW} + d_{BD} - d_{BW})/2 \\ d_{D3,Z} &= (d_{DW} + d_{CD} - d_{CW})/2 \end{aligned} \tag{10}$$

where

$$\begin{aligned} d_{AW} &= (d_{AB} + d_{AC} - d_{BC})/2 \\ d_{BW} &= (d_{AB} + d_{BC} - d_{AC})/2 \\ d_{CW} &= (d_{AC} + d_{BC} - d_{AB})/2 \end{aligned} \tag{11}$$

$$d_{DW} = \max(d_{AD} - d_{AW}, d_{BD} - d_{BW}, d_{CD} - d_{CW})$$

Here, max(a, b, c) denotes the maximum value among a, b, and c. Substituting Eqs. (10) into Eqs. (9), we have

$$\begin{aligned} d_{CD} - (d_{AC} + d_{BC} - d_{AB})/2 \\ < d_{AD} - (d_{AB} + d_{AC} - d_{BC})/2 \end{aligned} \tag{12}$$

$$\begin{aligned} d_{CD} - (d_{AC} + d_{BC} - d_{AB})/2 \\ < d_{BD} - (d_{AB} + d_{BC} - d_{AC})/2 \end{aligned}$$

which reduces to

$$\begin{aligned} d_{AB} + d_{CD} < d_{AC} + d_{BD} \text{ and} \\ d_{AB} + d_{CD} < d_{AD} + d_{BC} \end{aligned} \tag{13}$$

These inequalities are identical with Eqs. (8).

Although species D will usually be the last to join the cluster, any other species can be the last because of chance effects. However, it can be shown that the above condition holds for all cases. It can also be shown that the same condition holds for Tateno et al.'s (1982) and Faith's (1985) modifications of the distance-Wagner method.

Fitch and Margoliash's Method. This method is intended to choose the tree with the smallest percentage standard deviation of patristic (estimated) distances (p_{ij}) from observed distances (d_{ij}). Here, the patristic distance between species i and j is the sum of the lengths of all branches connecting species i and j in a tree. The percentage standard deviation is defined by

$$s_{FM} = \left[\frac{2}{n(n-1)} \sum_{i<j}^{n} \{(d_{ij} - p_{ij})/d_{ij}\}^2\right]^{1/2} \times 100 \tag{14}$$

where n is the number of species used and the p_{ij} are estimated by Fitch and Margoliash's (1967) "three groups" method. As is shown in the Appendix, the condition for obtaining the correct tree by this method is given by the inequalities

$$\begin{aligned} (d_{AC} - d_{AD} - d_{BC} + d_{BD})^2 \\ < (d_{AB} - d_{AD} - d_{BC} + d_{CD})^2 \end{aligned} \tag{15}$$

$$\begin{aligned} (d_{AC} - d_{AD} - d_{BC} + d_{BD})^2 \\ < (d_{AC} - d_{AB} - d_{CD} + d_{BD})^2 \end{aligned}$$

Tateno et al. (1982) used the following quantity to measure the deviation of patristic distances from observed distances:

$$s_0 = \left[\frac{2}{n(n-1)} \sum_{i<j}^{n} (d_{ij} - p_{ij})^2\right]^{1/2} \tag{16}$$

This measure gives a condition identical to (15) for the case of four species.

Maximum Parsimony Method. In this method, the number of nucleotide substitutions required to explain the evolutionary changes of the species considered is computed for all possible (or reasonable) topologies, and the topology with the smallest number of nucleotide substitutions is chosen. Only those nucleotide sites at which two different nucleotides exist in at least two different species are informative for this purpose. There are only three nucleotide

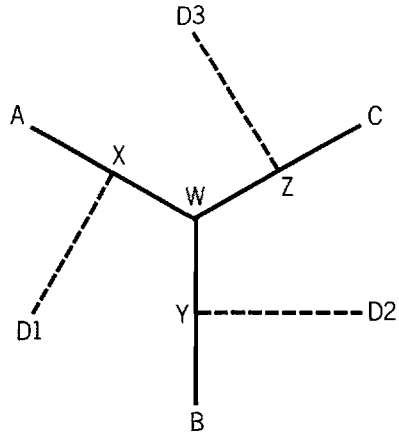

Fig. 5. Three possible ways (D1–D3) for species D to be connected to the unrooted tree of species A, B, and C

configurations (the first three configurations in Table 2) that are informative for the case of four species. Configuration 1 requires one nucleotide substitution if the topology in Fig. 4A is correct, whereas configurations 2 and 3 require two nucleotide substitutions. The latter two configurations favor different topologies where only one substitution is required.

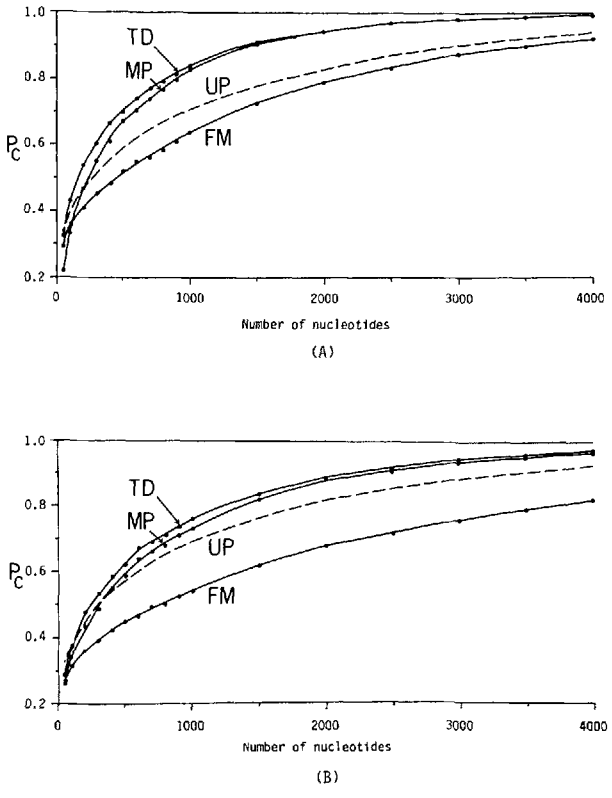

Fig. 6A, B. Probabilities (P_c) of obtaining the correct topology shown in Fig. 4A (four-species case): **A** one-parameter model; **B** two-parameter model. R = 0.1. UP, UPGMA; TD, transformed distance method; FM, Fitch and Margoliash's method; MP, maximum parsimony method

Therefore, to obtain the correct topology given in Fig. 4A, we must have

$$m_1 > m_2 \quad \text{and} \quad m_1 > m_3 \qquad (17)$$

where m_i is the observed number of the i-th configuration in Table 2.

Compatibility Method. Certain nucleotide configurations can be fitted to a given topology with the minimum number of substitutions (the number of variable nucleotides minus 1), whereas the other configurations require more than the minimum. The nucleotide sites with the first group of configurations are called compatible sites, whereas those with the second group are called incompatible sites. In the compatibility method, the topology with the largest number of compatible nucleotide sites is chosen as the final tree. Therefore, the condition for obtaining the correct topology with the compatibility method is identical with that for the maximum parsimony method in the case of four species, because when a site is incompatible only one additional substitution is necessary to make it compatible. This is true even for the case of five species. When $n \geq 6$, however, the conditions for the two methods are not identical.

Comparison of Different Methods. The relationships between the probability of obtaining the correct tree and the number of nucleotides examined for the UPGMA, transformed distance, Fitch–Margoliash, and maximum parsimony methods are given in Figs. 6 and 7. The relationships for the distance-Wagner and compatibility methods are identical with those of the transformed distance and maximum parsimony methods, respectively. The true phylogenetic trees for the cases of Figs. 6 and 7 are given by trees A and B in Fig. 4, respectively, excluding species E. The main difference between

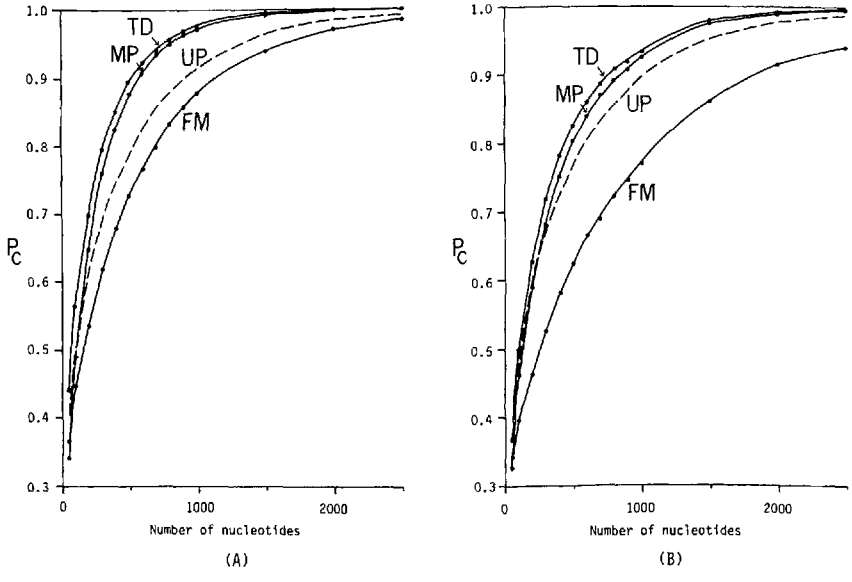

Fig. 7A, B. Probabilities P_c of obtaining the correct topology shown in Fig. 4B (four-species case): **A** one-parameter model; **B** two-parameter model. R = 0.17. Abbreviations as in Fig. 6

trees A and B is that the length between the branching points X and Y is longer in the latter than in the former.

Figure 6A shows the relationships for the one-parameter model of nucleotide substitution. The curve for UPGMA is identical with the one for R = 0.1 in Fig. 3. It is clear that the transformed distance method has a much higher probability (P_c) of obtaining the correct tree than does UPGMA. The maximum parsimony method also shows a higher value of P_c than does UPGMA except when $m < 200$. The P_c value for the maximum parsimony method is slightly lower than that for the transformed distance method when $m < 2000$. Note also that the maximum parsimony method shows the poorest performance when $m \leq 100$. This is probably due to the very small number of informative sites available for this case. The Fitch–Margoliash method shows the smallest value of P_c for $m > 100$. The inefficiency of this method is probably due to the low power of s_{FM} or s_0 in discriminating the correct tree from erroneous ones (Tateno et al. 1982).

The number of nucleotides required for obtaining the correct topology with a probability of 95% is given in Table 3 for the six tree-making methods considered. This number is 2100 for the transformed distance, distance-Wagner, maximum parsimony, and compatibility methods. UPGMA and the Fitch–Margoliash method require twice as many nucleotides.

The P_c values for the two-parameter model of nucleotide substitution are given in Fig. 6B. For all the tree-making methods, they are lower than those for the one-parameter model. This reduction in P_c is very small for UPGMA but is substantial for the other methods. It is caused by the higher frequency of backward and parallel mutations for the two-parameter model than for the one-parameter model. The number of nucleotides required (m*) is also greater for the two-parameter model (Table 3).

The probability of obtaining the correct topology for model tree B in Fig. 4 is substantially higher than that for tree A, as expected (Fig. 7). However, the relative probabilities for the six tree-making methods are nearly the same. The effect of the high frequency of transitional substitution is also similar for the two model trees. The numbers of nucleotides required are substantially smaller if tree B is the correct one (Table 3). That is, the branch length between X and Y has a strong influence on the number of nucleotides required.

Case 3. Five Species

Let us now consider the case where two outgroup species (D and E) are available (Fig. 4). As mentioned earlier, the P_c value for UPGMA is not affected by the availability of outgroup species. In other methods, however, the availability of two outgroup species increases the P_c value compared with the case of one outgroup species.

Table 3. The number (m*) of nucleotides required for obtaining the correct tree with a probability of 0.95

Tree-making method	Four species		Five species	
	One-parameter model	Two-parameter model	One-parameter model	Two-parameter model
	Tree A of Fig. 4			
TD = DW	2100	3100	1700	2600
MP = CP	2100	3300	1700	2700
UP[a]	4200	4700	4200	4700
FM	5000	8200	3000	4900
	Tree B of Fig. 4			
TD = DW	760	1200	640	890
MP = CP	790	1300	680	980
UP[a]	1400	1500	1400	1500
FM	1700	2800	990	1700

CP, compatibility method; DW, distance-Wagner method; other abbreviations as in Fig. 6

[a] These numbers are the same as those for three species

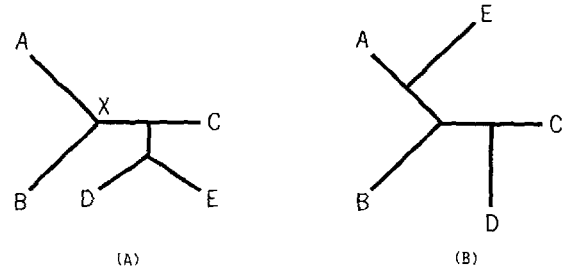

Fig. 8A, B. *Two possible topologies for five species*

Transformed Distance Method. When species D and E in Fig. 4 are known to be outgroup species, there are two ways of constructing a tree. One is to construct an unrooted tree using D or E as an outgroup species. In this case, the correct tree is obtained only when D and E make a cluster and this cluster is connected somewhere between species C and node X in Fig. 8A. The other method is to combine D and E as a single (composite) outgroup species and to construct a topology as in the case of four species. In this case, topologies in which species D and E are not clustered, as in Fig. 8B, do not occur. Thus, the second method is expected to be superior to the first in obtaining the correct tree. We shall therefore consider only the second method.

In the second method, the transformed distance between species i and j (i, j = A, B, C) is computed by

$$d'_{ij} = [d_{ij} - (d_{iD} + d_{iE})/2 - (d_{jD} + d_{jE})/2]/2 + c$$

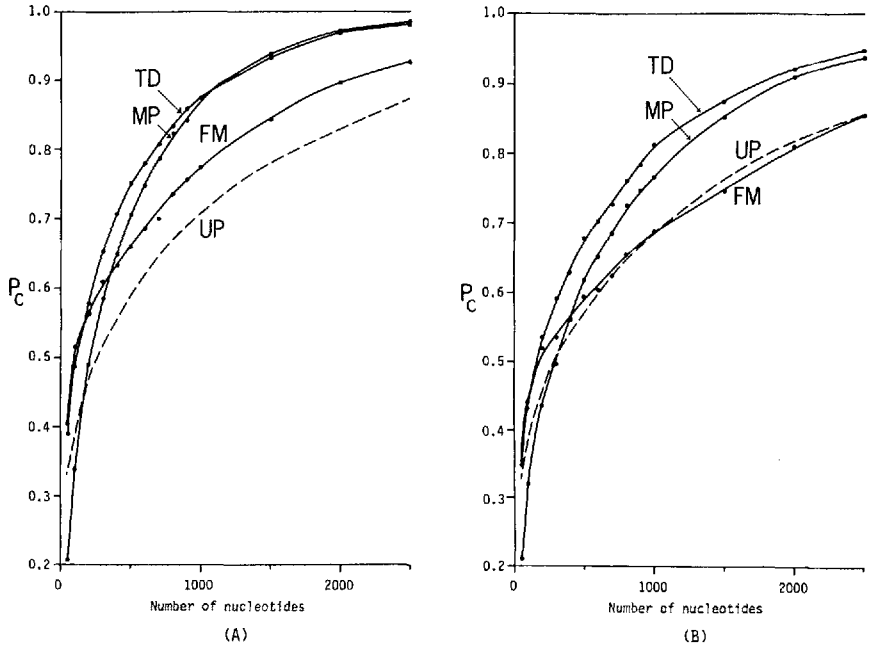

Fig. 9A, B. Probabilities (P_c) of obtaining the correct topology shown in Fig. 4A (five-species case): **A** one-parameter model; **B** two-parameter model. R = 0.1. Abbreviations as in Fig. 6

Therefore, the condition $d'_{AB} < d'_{AC}$ and $d'_{AB} < d'_{BC}$ can be written as

$$d_{AB} + (d_{CD} + d_{CE})/2 < d_{AC} + (d_{BD} + d_{BE})/2 \quad (18)$$

$$d_{AB} + (d_{CD} + d_{CE})/2 < d_{BC} + (d_{AD} + d_{AE})/2$$

These two inequalities are the condition for obtaining the correct tree.

Distance-Wagner Method. As with the transformed distance method, there are two different ways of constructing a topology. The first method is again inferior to the second method in obtaining the correct tree, though the details are somewhat different from those for the transformed distance method. The condition for obtaining the correct tree by the second method is identical with that for the transformed distance method. This is obvious because in the second method species D and E are combined and regarded as a single composite species. Here we consider only the second method.

Fitch and Margoliash's Method. There are again two ways of constructing a tree: One is to consider the five species separately, and the other is to combine species D and E. Unlike with the transformed distance and distance-Wagner methods, however, the first procedure is not inferior to the second. The second procedure becomes identical with that for four species if we replace d_{AD}, d_{BD}, and d_{CD} by $(d_{AD} + d_{AE})/2$, $(d_{BD} + d_{BE})/2$, and $(d_{CD} + d_{CE})/2$, respectively. Our computations have shown that the use of two outgroup species as a composite species increases the probability of obtaining the correct tree only slightly compared with the case of one outgroup species. We shall therefore disregard this case in the following.

The condition for obtaining the correct topology by the first tree-making procedure is somewhat complicated, but it is possible to evaluate the P_c value by the method given in the Appendix.

Maximum Parsimony and Compatibility Methods. There are 51 possible configurations for five species, and 25 of them are informative for purposes of phylogenetic inference. The frequencies of these configurations can be determined by using Eq. (5) and pseudorandom numbers for each value of m. Once the frequencies are determined, one can decide whether or not the topology of a tree obtained by the maximum parsimony method is correct. Here again, we have two ways of constructing a topology. The first one is to consider only three topologies where two outgroup species (D and E) are clustered. Since these two species are known to be outgroup species, this method is justified. The second method is the usual maximum parsimony procedure, where all 15 topologies are compared. Though the differences between the two methods are quite small when m is large, the first method gives better results when m is small. Therefore, we consider only the first method. As mentioned earlier, the compatibility method always gives the same tree as the maximum parsimony method in the case of five species.

Comparison of Different Methods. Figure 9 shows the relationships between the probability of obtaining the correct topology and the number of nucleotides examined for the six tree-making methods for

the case of tree A in Fig. 4. The results for the one-parameter model of nucleotide substitution are given in Fig. 9A, and those for the two-parameter model are in Fig. 9B. The relationship for UPGMA is identical with that for the case of three or four species. Comparison of Figs. 6A and 9A indicates that all other methods show better performance in obtaining the true tree with two outgroup species than with one outgroup species available. In particular, the P_c value for the Fitch–Margoliash method has increased substantially and now exceeds the value for UPGMA. When m = 50, the Fitch–Margoliash method shows the highest P_c value among all the methods examined. However, as m increases, the P_c values for all other methods except UPGMA rapidly increase, exceeding the value for the Fitch–Margoliash method. In general, the transformed distance and distance-Wagner methods again show the best performance. When $m > 1000$, however, the maximum parsimony and compatibility methods are slightly better.

When the rate of transitional nucleotide substitution is much higher than the rate of transversional substitution (the two-parameter model), the efficiency of the Fitch–Margoliash method declines considerably and becomes nearly equal to that of UPGMA. The P_c values for all other methods except UPGMA are also considerably lower than for the one-parameter model. It is also interesting that the maximum parsimony and compatibility methods are now inferior to the transformed distance and distance-Wagner methods for all m values examined.

The numbers of nucleotides required for obtaining the correct topology with a probability of 95% are given in Table 3. They are considerably smaller than those for four species, but at least 1700 nucleotides are still necessary whichever model of nucleotide substitution or tree-making method is used. In the case of mitochondrial DNA, where the two-parameter model is more appropriate, about 3000 nucleotides seem to be necessary. Table 3 also includes the numbers of nucleotides required for tree B in Fig. 4. As expected, they are considerably smaller than those for tree A. Even with tree B, however, about 900 nucleotides are required for mitochondrial DNA when the transformed distance or distance-Wagner method is used.

Discussion

In the present study, we have used two simple models of nucleotide substitution: the one-parameter and two-parameter models. Neither model appears to be very realistic when long-term evolution is considered (Gojobori et al. 1982; Aquadro et al. 1984; Li et al. 1984). However, when the number of nucleotide substitutions is relatively small, as between humans and apes, the effect of the deviation from the two models is probably small. A much more serious effect may be generated by the well-known unequal rates of substitution at different nucleotide sites. If there are sites where the substitution rate is unusually high, the reliability of a reconstructed tree declines because there could be many backward and parallel substitutions. Variation in the rate of substitution among evolutionary lineages would also reduce the reliability of a reconstructed tree. Therefore, the numbers of nucleotides required for obtaining a correct tree presented in this paper should be regarded as minima.

We have shown that a large number of nucleotides is required for resolving the branching order among three closely related species in the absence of outgroup species and that all methods considered are equally efficient in obtaining the correct topology under the conditions we considered. In the presence of outgroup species, however, the transformed distance, distance-Wagner, maximum parsimony, and compatibility methods are more efficient in recovering the correct tree than UPGMA and the Fitch–Margoliash method. We have also shown that the availability of two outgroup species improves the accuracy of tree reconstruction except with UPGMA. The Fitch–Margoliash method shows the poorest performance when only one outgroup species is used but becomes better than UPGMA when two outgroup species are used.

It should be noted that this conclusion depends on the extent of DNA divergence among the species studied as well as on the shape of the tree. In this paper, we have considered a case similar to the divergence of the mitochondrial DNAs (mtDNAs) from humans, chimpanzees, gorillas, orangutans, and gibbons. Therefore, our conclusion may not apply to cases where the extent of DNA divergence and the tree shape are quite different. However, the mathematical theory developed here can be used for any case. Note also that our primary objective in this paper was to resolve the branching order among three closely related species. For this purpose, we considered the case where one or two outgroup species are available. Our problem is therefore different from the construction of a tree for four or five species without outgroup species. Furthermore, the relative merits of different tree-making methods depend on the number of species as well as on the tree shape. Therefore, the conclusions obtained from this study should not be extrapolated to other cases without caution. For example, the Fitch–Margoliash method showed a rather poor performance for our case, but its relative merit may increase as the number of species used increases.

Brown et al. (1982) conducted a parsimony analysis for their mtDNA data (896 bp) for humans and apes and showed that the most parsimonious tree is CG–H–OB (topology 2), where C, G, H, O, and B stand for chimpanzees, gorillas, humans, orangutans, and gibbons. Nei et al. (1985) analyzed Brown et al.'s data using UPGMA and obtained a different tree (HC–G–OB; topology 1). Recently, Hixson and Brown (1986) sequenced about 950 bp of the small rRNA gene region of mtDNA for pygmy chimpanzees, common chimpanzees, gorillas, and orangutans. It would be interesting to combine these two data sets and reconstruct the tree for humans, chimpanzees, gorillas, and orangutans using the six tree-making methods considered here. For this purpose, we can use a combined sequence of 1834 bp, excluding insertions and deletions.

Table 4 shows the numbers of nucleotide differences and the estimates of nucleotide substitutions per site (evolutionary distances) obtained by Jukes and Cantor's (1969) formula (see also Kimura and Ohta 1972). With this set of distances, all four distance matrix methods (UPGMA, transformed distance, distance-Wagner, and Fitch–Margoliash) give the same topology: HC–GO. The maximum parsimony and compatibility methods also support this topology (see Table 5). The UPGMA tree with the standard errors of the branching points obtained by Nei et al.'s (1985) method is given in Fig. 10.

Table 4. Numbers of nucleotide differences (below diagonal) and evolutionary distances (above diagonal) for two regions of mtDNA from humans, chimpanzees, gorillas, and orangutans[a]

	Human	Chimp.	Gorilla	Orang.
Human	—	0.063 ±0.006	0.072 ±0.007	0.134 ±0.009
Chimp.	114 (35)	—	0.077 ±0.007	0.141 ±0.009
Gorilla	126 (34)	134 (39)	—	0.140 ±0.009
Orang.	225 (82)	237 (84)	234 (85)	—

[a] 1834 nucleotides from Brown et al. (1982) and Hixson and Brown (1986) were used. The numbers of nucleotide differences in parentheses are those for the small rRNA gene region (a sequence of 939 bp) of Hixson and Brown (1986). The human sequence for this region is from Anderson et al. (1981)

The results of our parsimony analysis are different from those of Brown et al. (1982). We used about twice as many nucleotides as they did. So one might think that our conclusion is more reliable than Brown et al.'s. Unfortunately, the rate of nucleotide substitution in the small rRNA region is considerably lower than in the other region, and this results in a smaller number of informative sites in this region than in the other region. Theoretically, if one uses two outgroup species, the accuracy of the tree obtained improves compared with the case of one outgroup species, as shown in this paper. In the case of Brown et al.'s (1982) data (region 1 only), however, the difference in the minimum number of nucleotide substitutions required between topologies 1 and 2 is −2 in Brown et al.'s analysis (two outgroup species) and 4 in our analysis (one outgroup species). This is inconsistent with the theoretical expectation, but probably due to chance effects.

While the parsimony analysis does not give a consistent result, all four distance matrix methods give the same topology (topology 1) for both Brown et al.'s data (Nei et al. 1985; Nei, in press) and the combined data. This strengthens support for topology 1, but the branch length between X and Y in the UPGMA tree in Fig. 10 is still not statistically significant even with the combined data. It seems that, as predicted from the present study, many more nucleotides have to be examined if we are to settle this problem. Koop et al. (1986) compared about 2100 nucleotides (excluding insertions and deletions) of η-globin genes of humans, chimpanzees, gorillas, and orangutans. The nucleotide divergence ($2\lambda T$) among humans, chimpanzees, and gorillas is approximately 0.017, which is only one-sixth the value for Brown et al.'s (1982) mtDNA data. Because of this low rate of nucleotide substitution, η-globin gene data are less informative for resolving the human–chimpanzee–gorilla divergence.

In the present paper, we have considered the effect of stochastic errors of nucleotide substitution on a reconstructed tree, ignoring the effect of genetic polymorphism that might have existed at the time

Table 5. Numbers of nucleotide changes required to explain three topologies for humans (H), chimpanzees (C), gorillas (G), and orangutans (O)[a]

	Region 1			Region 2			Regions 1 & 2		
Topology	TI	TV	Total	TI	TV	Total	TI	TV	Total
HC–GO	58	3	61	25	2	27	83	5	88
CG–HO	59	6	65	26	2	28	85	8	93
HG–CO	63	6	69	26	1	27	89	7	96

[a] Region 1, data of Brown et al. (1982); region 2, data of Hixson and Brown (1986). TI, transitional substitutions; TV, transversional substitutions. Only phylogenetically informative configurations of nucleotides are considered

of speciation. In the presence of polymorphism, a tree reconstructed from a single gene from each species may be different from the species tree, even if a large number of nucleotides are used (Tateno et al. 1982; Takahata and Nei 1985). In the case of three species, the probability of obtaining erroneous topologies is high when the difference between the times of the first and second speciations is small (Fig. 1). Considering neutral mutations, Nei (1986) has shown that the number of generations (t_1) required for obtaining the species tree with a probability of 1 − E is given by $t_1 = -2N \log_e(3E/2)$, since $E = \frac{2}{3} \exp(-t_1/2N)$. N is the long-term effective population size. Therefore, if we want to make E as small as 0.05, t_1 must be 5.2 N generations. In hominoids, N could be about 10^4 (Nei and Graur 1984). If this is the case, $t_1 = 5.2 \times 10^4$ generations for E = 0.05, or about 800,000 years if one generation consists of 15 years. Therefore, unless the interval between the first and second speciation events is longer than one million years, examination of a single gene from each species is unlikely to give a clear-cut resolution of the branching order among humans, chimpanzees, and gorillas.

Recently, Ueda et al. (1985) reported the existence of a truncated pseudogene for immunoglobulin C_ϵ in humans and gorillas but not in chimpanzees, the other apes, and the Old World monkeys. This finding is consistent with the topology HG–CO. However, Hixson and Brown (1986) discovered a one-base deletion shared by chimpanzees and gorillas, which is consistent with the topology CG–HO. Obviously, these two topologies are contradictory; the results appear to be due to polymorphism at the time of divergence among humans, chimpanzees, and gorillas.

To avoid this type of problem, we must study many independent genes (loci). The number of genes required for obtaining the correct *species* tree with a probability of 95% may be obtained in the following way: In the case of three species, there are three possible topologies, and only one of them is the correct one. Under the assumption of no selection and constant population size, the probability that a gene tree has the same topology as the species tree (topology 1) is $G_1 = 1 - \frac{2}{3} \exp(-t_1/2N)$ (see Nei 1986). The probabilities of obtaining the other two topologies (topologies 2 and 3) are $G_2 = G_3 = \frac{1}{3} \exp(-t_1/2N)$ (see Nei 1986). Therefore, if we study k independent genes, the probability that p genes show topology 1, q genes show topology 2, and r genes show topology 3 is

$$Q(p, q, r) = \frac{k!}{p!\, q!\, r!} G_1^p G_2^q G_3^r \qquad (19)$$

Suppose that the topology supported by the largest number of genes is regarded as the correct one. Under this procedure, we obtain the correct topology only when p is greater than both q and r. It is therefore possible to compute the probability (Q_c) of obtaining the correct topology by summing all Q(p, q, r) satisfying p > q and p > r for a given value of k. The Q_c for various values of $t_1/2N$ and k are given in Table 6. When $t_1/2N$ is equal to or larger than 2, Q_c is quite high even for a small number of genes used. When $t_1/2N$ is smaller than 0.5, however, a large number of genes is necessary to obtain the species tree with a high probability. In Table 6, Q_c for k = 2 is smaller than that for k = 1. This is because when k = 2 there are cases with p = q = 1 or p = r = 1 and these cases are not included. The lower values of Q_c for k = 4 than for k = 3 occur for the same reason.

Equation (19) allows us to compute the number of genes or loci (k*) required for obtaining the correct species tree with a probability of 95% when t_1 and N are given. It is 3 for $t_1/2N = 2$, 5 for $t_1/2N = 1.5$, 7 for $t_1/2N = 1$, and 14 for $t_1/2N = 0.5$.

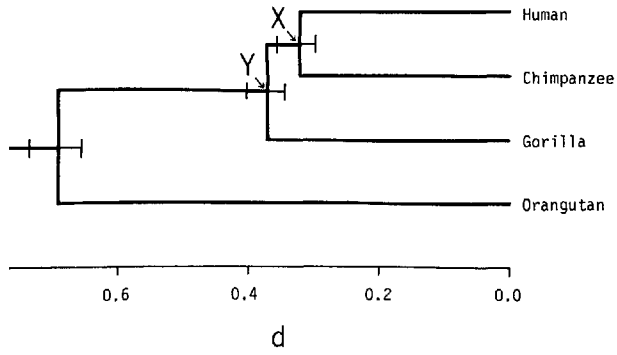

Fig. 10. Phylogenetic tree of mitochondrial DNAs from humans, chimpanzees, gorillas, and orangutans. This topology is supported by all the tree-making methods considered here. The branch lengths were estimated by UPGMA. The standard error of each branching point was obtained by Nei et al.'s (1985) method

Table 6. Probabilities of obtaining the correct species tree from data for k genes

	k						
$t_1/2N$	1	2	3	4	5	6	7
4	0.988	0.976	1.000	1.000	1.000	1.000	1.000
2	0.910	0.828	0.977	0.967	0.994	0.996	0.999
1.5	0.851	0.725	0.940	0.916	0.974	0.984	0.992
1	0.755	0.570	0.849	0.798	0.901	0.933	0.954
0.5	0.596	0.355	0.642	0.555	0.675	0.745	0.776

Acknowledgments. We thank Drs. P. Smouse, P. Pamilo, N. Takahata, and J.C. Stephens for their comments on earlier versions of the manuscript. This work was supported by grants from the National Science Foundation and National Institutes of Health to M. Nei.

Appendix: Condition for Obtaining the Correct Tree with Fitch and Margoliash's (1967) Method

Four Species. To compute S_{FM} in Eq. (14) for a given tree, we must have estimates of the branch lengths of the tree (a, b, c, d, and e in Fig. A1A). In the Fitch–Margoliash (FM) method, these lengths are estimated in the following way: We can start either from the

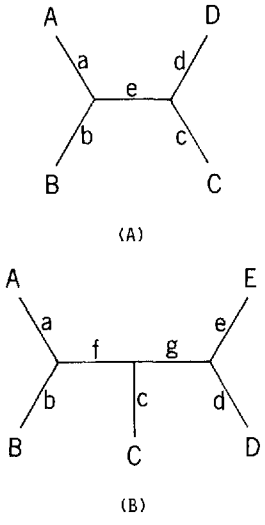

Fig. A1A, B. Trees for four species (**A**) and five species (**B**). a, b, . . . , g are branch lengths

species pair A and B or from the pair C and D; the resulting estimates are the same. Therefore, let us start from A and B. First the distance between species A and another "species" X (C and D) is computed by $d_{AX} = (d_{AC} + d_{AD})/2$. Similarly, the distance between B and X is $d_{BX} = (d_{BC} + d_{BD})/2$. Branch lengths a and b are then estimated by

$$a = (d_{AB} + d_{AX} - d_{BX})/2 \tag{A1}$$

$$b = (d_{AB} + d_{BX} - d_{AX})/2 \tag{A2}$$

Species A and B are now combined and regarded as a single composite species. The distances between this composite species and the other species are computed by $d_{(AB)C} = (d_{AC} + d_{BC})/2$ and $d_{(AB)D} = (d_{AD} + d_{BD})/2$. One can then estimate branch lengths c, d, and e by

$$c = [d_{CD} + d_{(AB)C} - d_{(AB)D}]/2 \tag{A3}$$

$$d = [d_{CD} + d_{(AB)D} - d_{(AB)C}]/2 \tag{A4}$$

$$e = [d_{(AB)C} + d_{(AB)D} - d_{CD}]/2 - d_{AB}/2 \tag{A5}$$

If we use matrix algebra, the above branch-length estimates can be written as

$$\mathbf{A} = \tfrac{1}{4}\mathbf{MD} \tag{A6}$$

where **A** and **D** are the column vectors of a, b, c, d, e, and d_{AB}, d_{AC}, d_{AD}, d_{BC}, d_{BD}, d_{CD}, respectively, and **M** is the matrix

$$\mathbf{M} = \begin{bmatrix} 2 & 1 & 1 & -1 & -1 & 0 \\ 2 & -1 & -1 & 1 & 1 & 0 \\ 0 & 1 & -1 & 1 & -1 & 2 \\ 0 & -1 & 1 & -1 & 1 & 2 \\ -2 & 1 & 1 & 1 & 1 & -2 \end{bmatrix} \tag{A7}$$

Once the branch lengths are obtained, the patristic distances p_{ij} are given by $p_{AB} = a + b$, $p_{AC} = a + c + e$, $p_{AD} = a + d + e$, $p_{BC} = b + c + e$, $p_{BD} = b + d + e$, and $p_{CD} = c + d$. Therefore, one can compute the percentage standard deviation s_{FM} in Eq. (14). In the present case, $(d_{AB} - p_{AB})^2 = (d_{CD} - p_{CD})^2 = 0$ and $(d_{ij} - p_{ij})^2 = (d_{AC} - d_{AD} - d_{BC} + d_{BD})^2/16$ for the other pairs of species (i and j). Thus we have

$$s_{FM} = [(d_{AC} - d_{AD} - d_{BC} + d_{BD})^2 \times C]^{1/2} \times 100 \tag{A8}$$

where

$$C = \frac{1}{8n(n-1)}[d_{AC}^{-2} + d_{AD}^{-2} + d_{BC}^{-2} + d_{BD}^{-2}]$$

and n is the number of species involved. It can be shown that the s_{FM} for the other two toplogies are $[(d_{AB} - d_{AD} - d_{BC} + d_{CD})^2 \times C]^{1/2} \times 100$ and $[(d_{AC} - d_{AB} - d_{CD} + d_{BD})^2 \times C]^{1/2} \times 100$. Therefore the condition for obtaining the correct topology is given by (15).

Five Species. Let us consider the tree given in Fig. A1B. As in the case of four species, we can start from the pair A and B or the pair D and E, and the result will be the same. Let us start from the pair A and B. The branch lengths a and b are then estimated by Eqs. (A1) and (A2), respectively, if we redefine d_{AX} and d_{BX} as

$$d_{AX} = (d_{AC} + d_{AD} + d_{AE})/3$$

$$d_{BX} = (d_{BC} + d_{BD} + d_{BE})/3$$

Furthermore, c and f can be estimated by

$$c = [d_{(AB)C} + d_{CX} - d_{(AB)X}]/2$$

$$f = [d_{(AB)C} + d_{(AB)X} - d_{CX}]/2 - d_{AB}/2$$

where

$$d_{(AB)C} = (d_{AC} + d_{BC})/2$$

$$d_{CX} = (d_{CD} + d_{CE})/2$$

$$d_{(AB)X} = (d_{AD} + d_{BD} + d_{AE} + d_{BE})/4$$

We also have

$$d = [d_{DE} + d_{(ABC)D} - d_{(ABC)E}]/2$$

$$e = [d_{DE} + d_{(ABC)E} - d_{(ABC)D}]/2$$

$$g = [d_{(ABC)D} + d_{(ABC)E} - d_{DE}]/2 - (a + b + c + 2f)/3$$

where

$$d_{(ABC)D} = (d_{AD} + d_{BD} + d_{CD})/3$$

$$d_{(ABC)E} = (d_{AE} + d_{BE} + d_{CE})/3$$

If we use matrix algebra, the above equations for branch-length estimates can be written as

$$\mathbf{A} = (1/24)\mathbf{MD} \tag{A9}$$

where **A** and **D** are the column vectors of a, b, c, d, e, f, g, and d_{AB}, d_{AC}, d_{AD}, d_{AE}, d_{BC}, d_{BD}, d_{BE}, d_{CD}, d_{CE}, d_{DE}, respectively, and **M** is

$$\mathbf{M} = \begin{bmatrix} 12 & 4 & 4 & 4 & -4 & -4 & -4 & 0 & 0 & 0 \\ 12 & -4 & -4 & -4 & 4 & 4 & 4 & 0 & 0 & 0 \\ 0 & 6 & -3 & -3 & 6 & -3 & -3 & 6 & 6 & 0 \\ 0 & 0 & 4 & -4 & 0 & 4 & -4 & 4 & -4 & 12 \\ 0 & 0 & -4 & 4 & 0 & -4 & 4 & -4 & 4 & 12 \\ -12 & 6 & 3 & 3 & 6 & 3 & 3 & -6 & -6 & 0 \\ 0 & -6 & 3 & 3 & -6 & 3 & 3 & 6 & 6 & -12 \end{bmatrix}$$

The differences between the patristic and observed distances can now be written as

$$\begin{bmatrix} d_{AC} - p_{AC} \\ d_{AD} - p_{AD} \\ d_{AE} - p_{AE} \\ d_{BC} - p_{BC} \\ d_{BD} - p_{BD} \\ d_{BE} - p_{BE} \\ d_{CD} - p_{CD} \\ d_{CE} - p_{CE} \end{bmatrix} = \frac{1}{12} \times \begin{bmatrix} -4 & 2 & 2 & 4 & -2 & -2 & 0 & 0 \\ 2 & -5 & 3 & -2 & 3 & -1 & 2 & -2 \\ 2 & 3 & -5 & -2 & -1 & 3 & -2 & 2 \\ 4 & -2 & -2 & -4 & 2 & 2 & 0 & 0 \\ -2 & 3 & -1 & 2 & -5 & 3 & 2 & -2 \\ -2 & -1 & 3 & 2 & 3 & -5 & -2 & 2 \\ 0 & 2 & -2 & 0 & 2 & -2 & -4 & 4 \\ 0 & -2 & 2 & 0 & -2 & 2 & 4 & -4 \end{bmatrix} \times \begin{bmatrix} d_{AC} \\ d_{AD} \\ d_{AE} \\ d_{BC} \\ d_{BD} \\ d_{BE} \\ d_{DC} \\ d_{CE} \end{bmatrix} \tag{A10}$$

and $d_{AB} = p_{AB}$ and $d_{DE} = p_{DE}$. Therefore, we can compute s_{FM} using Eq. (14).

Since we know that species D and E are outgroup species, we consider only the tree where D and E are clustered. However, this cluster may be connected to any of the branches A–W, B–W, and C–W of Fig. 5. Only when the (D, E) cluster is connected to the C–W branch do we obtain the correct topology. In our computation of the probability of obtaining the correct topology, we chose the tree that gave the smallest value of s_{FM} for each set of distance values (each replication) and examined whether or not the tree constructed was correct. The relative frequency of obtaining the correct tree is taken as the probability P_c.

References

Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465

Aquadro CF, Kaplan N, Risko KJ (1984) An analysis of the dynamics of mammalian mitochondrial DNA sequence evolution. Mol Biol Evol 1:423–434

Bianchi NO, Bianchi MS, Cleaver JE, Wolff S (1985) The pattern of restriction enzyme-induced banding in the chromosomes of chimpanzee, gorilla, and orangutan and its evolutionary significance. J Mol Evol 22:323–333

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

Eck RV, Dayhoff MO (1966) Inferences from protein sequence comparisons. In: Dayhoff MO (ed) Atlas of protein sequence and structure 1966. National Biomedical Research Foundation, Silver Spring, Maryland, pp 161–169

Faith DP (1985) Distance methods and the approximation of most-parsimonious trees. Syst Zool 34:312–325

Farris JS (1972) Estimating phylogenetic trees from distance matrices. Am Nat 106:645–668

Farris JS (1977) On the phenetic approach to vertebrate classification. In: Hecht MK, Goody PC, Hecht BM (eds) Major patterns in vertebrate evolution. Plenum, New York, pp 823–850

Ferris SD, Wilson AC, Brown WM (1981) Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. Proc Natl Acad Sci USA 78:2432–2436

Fitch WM (1977) On the problem of discovering the most parsimonious tree. Am Nat 111:223–257

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 155:279–284

Gojobori T, Li W-H, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369

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

Hixson JE, Brown WM (1986) A comparison of the small ribosomal RNA genes from the mitochondrial DNA of the great apes and humans: sequence, structure, evolution, and phylogenetic implications. Mol Biol Evol 3:1–18

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

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, Ohta T (1972) On the stochastic model for estimation of mutational distance between homologous proteins. J Mol Evol 2:87–90

Klotz LC, Blanken RL (1981) A practical method for calculating evolutionary trees from sequence data. J Theor Biol 91:261–272

Koop BF, Goodman M, Xu P, Chan K, Slightom JL (1986) Primate η-globin DNA sequences and man's place among the great apes. Nature 319:234–238

Le Quesne WJ (1969) A method of selection of characters in numerical taxonomy. Syst Zool 18:201–205

Li W-H (1981) Simple method for constructing phylogenetic trees from distance matrices. Proc Natl Acad Sci USA 78: 1085–1089

Li W-H, Wu C-I, Luo C-C (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21: 58–71

Nei M (1986) Stochastic errors in DNA evolution and molecular phylogeny. In: Gershowitz H (ed) Evolutionary perspectives and the new genetics. Alan R Liss, New York, in press

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

Nei M, Graur D (1984) Extent of protein polymorphism and the neutral mutation theory. Evol Biol 17:73–118

Nei M, Tajima F (1985) Evolutionary change of restriction cleavage sites and phylogenetic inference for man and apes. Mol Biol Evol 2:189–205

Nei M, Stephens JC, Saitou N (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

Sibley CG, Ahlquist JE (1984) The phylogeny of the hominoid primates, as indicated by DNA–DNA hybridization. J Mol Evol 20:2–15

Sneath PHA, Sokal RR (1973) Numerical taxonomy. WH Freeman, San Francisco, pp 230–234

Takahata N, Nei M (1985) Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344

Tateno Y, Nei M, Tajima F (1982) Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J Mol Evol 18:387–404

Templeton AR (1983) Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution 37:221–244

Ueda S, Takenaka O, Honjo T (1985) A truncated immunoglobulin ϵ pseudogene is found in gorilla and man but not in chimpanzee. Proc Natl Acad Sci USA 82:3712–3715

Received March 17, 1986/Revised June 25, 1986