# Relationships between Gene Trees and Species Trees[1]

*Pekka Pamilo*[2] *and Masatoshi Nei*

Center for Demographic and Population Genetics, The University of Texas Health Science Center at Houston

It is well known that a phylogenetic tree (gene tree) constructed from DNA sequences for a genetic locus does not necessarily agree with the tree that represents the actual evolutionary pathway of the species involved (species tree). One of the important factors that cause this difference is genetic polymorphism in the ancestral species. Under the assumption of neutral mutations, this problem can be studied by evaluating the probability ($P$) that a gene tree has the same topology as that of the species tree. When one gene (allele) is used from each of the species involved, the probability can be expressed as a simple function of $T_i \equiv t_i/(2N)$, where $t_i$ is the evolutionary time measured in generations for the $i$th internodal branch of the species tree and $N$ is the effective population size. When any of the $T_i$'s is $<1$, the probability $P$ becomes considerably $<1.0$. This probability cannot be substantially increased by increasing the number of alleles sampled from a locus. To increase the probability, one has to use DNA sequences from many different loci that have evolved independently of each other.

## Introduction

In the construction of phylogenetic trees, it is important to distinguish between a species (population) tree and a gene tree. The former refers to a tree of a group of species that reflects the actual evolutionary pathways, whereas the latter is a tree of a group of homologous (orthologous) genes each sampled from a different species (Tateno et al. 1982; Nei 1987). When there is allelic polymorphism within species, a tree constructed from DNA sequences for a given gene can be quite different from the species tree (Tajima 1983; Takahata and Nei 1985; Neigel and Avise 1986). This is particularly so when the time of divergence between different species is short.

Nei (1987) evaluated the probability that the topology of a gene tree is the same as that of the species tree for the case of three species, showing that the probability can be quite small depending on the population size and divergence times. When there are more than three species, the probability can be even smaller; but no studies seem to have been done on this problem. It is also important to know whether the probability can be increased by studying more than one allele at a locus from each species. The purpose of the present paper is to examine these two problems. The difference in topology between a gene tree and a species tree may also be introduced by sampling errors when the number of nucleotides examined is small. In the present

1. Key words: phylogenetic tree, gene tree, species tree, genetic polymorphism.

2. Current address: Department of Genetics, University of Helsinki, Arkadiankatu 7, SF-00100 Helsinki, Finland.

Address for correspondence and reprints: Dr. Masatoshi Nei, Center for Demographic and Population Genetics, Graduate School of Biomedical Sciences, The University of Texas Health Science Center at Houston, P.O. Box 20334, Houston, Texas 77225.

*Mol. Biol. Evol.* 5(5):568–583. 1988.

0737-4038/88/0505-0009$02.00

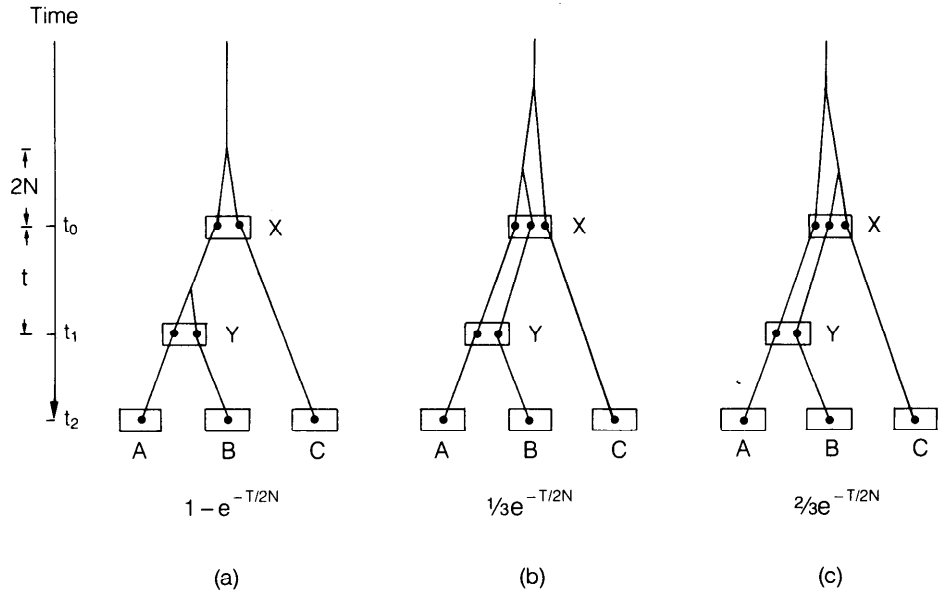


FIG. 1.—Relationships between the species and gene trees for the case of three species. A, B, and C represent three different species, and X and Y are the ancestral species that existed at the time of speciation ($t_0$ and $t_1$, respectively). $t = t_1 - t_0$. Dots in each box represent genes. The expected divergence time between two randomly chosen genes (alleles) from a population of effective size $N$ is $2N$ generations.

paper, however, we shall not consider this problem, since it has already been studied by Saitou and Nei (1986).

## Phylogenetic Trees Constructed from One Gene (Allele) from Each Species

In the following we evaluate the probability that a phylogenetic tree constructed from one gene (allele) from each species has the same topology as that of the species tree, under the assumption that both the topology and branch lengths of the species tree are known. We also assume that a gene (DNA sequence) is very long and that mutations (nucleotide substitutions) accumulate in proportion to evolutionary time (number of generations). Furthermore, we assume that the effective (diploid) population size ($N$) is the same for all species throughout the evolutionary time and that all mutations are neutral. These assumptions do not necessarily hold in nature, but they are useful for obtaining approximate answers to our questions.

### Case of Three Species

Let us first consider the simplest case of three species to illustrate our approach. Figure 1 shows three different types of gene trees for a given species tree. In tree (a), the three orthologous genes in species A, B, and C are derived from two different alleles in the ancestral species X. In this case, the phylogenetic tree for the three genes has the same topology as that of the species tree. In trees (b) and (c), the three genes are derived from three different alleles in the ancestral species X. However, the topology of gene tree (b) is identical with that of the species tree, whereas the topology of gene tree (c) is not. Theoretically, it is possible that all three genes in species A, B, and C

are derived from a single allele in species X at the time of the first population splitting, but the probability of occurrence of this event is very small (Tajima 1983; Takahata and Nei 1985). We therefore neglect this event. (Here we assume that as soon as a gene splits into two in a generation the two descendant genes (alleles) start to accumulate mutations.)

The probabilities of occurrence of gene trees (a), (b), and (c) can be evaluated by using the theory of gene genealogy of Kingman (1982), Tajima (1983), and Takahata and Nei (1985). Tree (a) occurs only when the two genes in species A and B are derived from a gene that existed between the times of two speciation events $t_0$ and $t_1$. The probability of occurrence of this event is equal to the probability that the two alleles in species Y are derived from a gene that existed during the time period of $t = t_1 - t_0$ generations. For simplicity, we call this the probability ($P_{21}$) that the two alleles in species Y are derived from one allele in species X. $P_{21}$ is given by $1 - e^{-T}$, where $T = t/(2N)$ (Tajima 1983). On the other hand, the probability of occurrence of trees (b) and (c) is equal to the probability ($P_{22}$) that the two alleles in species Y are derived from two alleles in species X. This probability is given by $e^{-T}$. However, the two alleles entering into species A and B may be more closely related to each other than to the other allele. The probability of occurrence of this event is $\frac{1}{3}$, since the three alleles in X are randomly distributed into the three extant species. Therefore, the probabilities of occurrence of trees (b) and (c) are $e^{-T}/3$ and $2e^{-T}/3$, respectively.

It is now clear that the probability ($P$) that a gene tree has the same topology as that of the species tree is

$$P = P_{21}(T) + \tfrac{1}{3}P_{22}(T) = 1 - \tfrac{2}{3}e^{-T} \qquad (1)$$

(Nei 1987). The above equation indicates that $P$ is determined entirely by $T = t/(2N)$ and increases as $T$ increases. For example, $P$ is $\frac{1}{3}$ for $T = 0$ and 0.95 for $T = 2.6$ ($5.2N$ generations) [see fig. 2(a)]. Note that $5.2N$ generations correspond to ~1 Myr in human evolution if $N$ is $10^4$ and one generation is 20 years. Note also that the evolutionary time between $t_1$ and $t_2$ has no effect on the $P$ value.

## Case of Four Species

In the case of four species, there are two different topologies in which species A and B are more closely related to each other than to the other species (two unlabeled topologies) (fig. 3). In the following we denote the species by capital letters (A, B, C, . . .) and the genes sampled from them by lowercase letters (a, b, c, . . .). We also distinguish one topology from another by using parentheses. Thus, the topologies in figures 3(a) and 3(b) are denoted by (AB)(CD) and ((AB)C)D, respectively.

To evaluate the probability that a gene tree has the same topology as that of the species tree, we must consider two evolutionary periods, $T_1$ [$= t_1/(2N)$] and $T_2$ [$= t_2/(2N)$] in figure 3. Evaluation of this probability is accomplished by considering $P_{21}(T)$, $P_{22}(T)$, and other, similar quantities. In general, we must know the probability [$P_{ij}(T)$] that $i$ alleles in a generation are derived from $j$ alleles that existed $2TN$ generations ago. The $P_{ij}(T)$ can be obtained by the formulas given by Tavaré (1984), Watterson

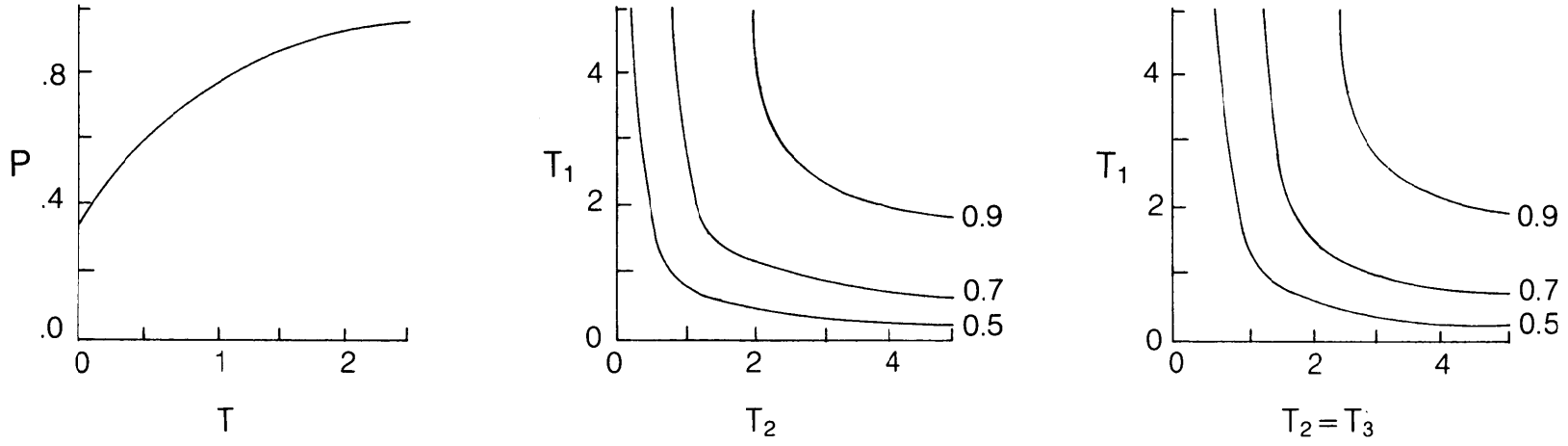


FIG. 2.—Relationships between the P value and internodal branch lengths (T, $T_1$, $T_2$, and $T_3$) of the species tree. Diagrams (b) and (c) show the relationships in terms of isoclines for the same P or $P_A$ value.

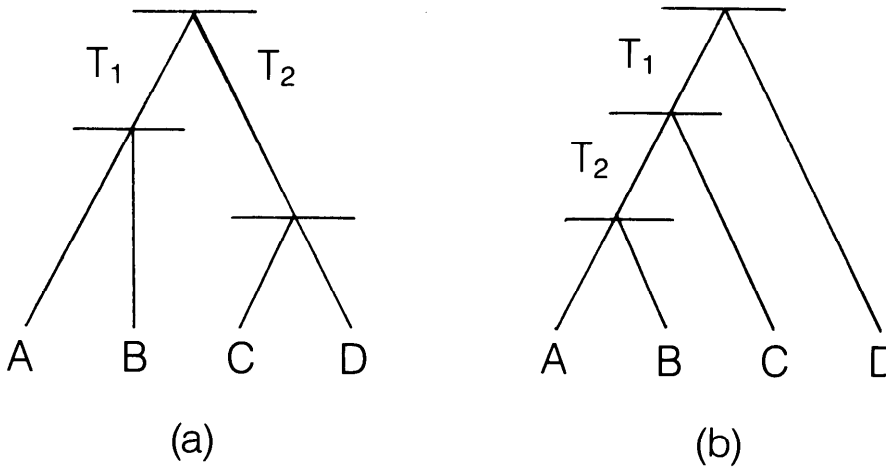


FIG. 3.—Two different types of species trees (unlabeled trees) that are possible for the case of four species.

(1984), and Takahata and Nei (1985). In the present paper we need the following quantities:

$$P_{21}(T) = 1 - e^{-T},$$

$$P_{22}(T) = e^{-T},$$

$$P_{31}(T) = 1 - \tfrac{3}{2}e^{-T} + \tfrac{1}{2}e^{-3T},$$

$$P_{32}(T) = \tfrac{3}{2}(e^{-T} - e^{-3T}),$$

$$P_{33}(T) = e^{-3T},$$

$$P_{41}(T) = 1 - \tfrac{9}{5}e^{-T} + e^{-3T} - \tfrac{1}{5}e^{-6T},$$

$$P_{42}(T) = \tfrac{9}{5}e^{-T} - 3e^{-3T} + \tfrac{6}{5}e^{-6T},$$

$$P_{43}(T) = 2e^{-3T} - 2e^{-6T},$$

$$P_{44}(T) = e^{-6T}.$$

Let us now consider the probability that a gene tree has the same topology as that of the species tree, given that the latter is (AB)(CD). We denote this probability by $P[(AB)(CD)]$. The four genes in species A, B, C, and D may be derived from (1) two alleles, (2) three alleles, and (3) four alleles that existed in the common ancestral species. When they are derived from two alleles [case (1)], the gene tree always has the same topology as that of the species tree. The probability of occurrence of this case is given by $P_{21}(T_1)P_{21}(T_2)$ [see fig. 3(a)]. When they are derived from three alleles [case (2)], either the A-B species lineage or the C-D lineage will receive two of them. The probability of occurrence of the former event is $P_{22}(T_1)P_{21}(T_2)$, whereas

that of the latter is $P_{21}(T_1)P_{22}(T_2)$. In both cases the correct topology is obtained with a probability of $1/3$. On the other hand, the probability of occurrence of case (3) is $P_{22}(T_1)P_{22}(T_2)$. In this case, however, the correct topology is obtained only (1) when there are two pairs of alleles in which one member of a pair shows a closer relationship to the other member of that pair than to either member of the other pair (fig. 2 of Tajima [1983]) and (2) when the first pair enter the A-B or C-D lineage. The probability of the first event is $1/3$, as shown by Tajima (1983), and the probability of the second event is also $1/3$. Therefore, the correct topology is obtained with a probability of $1/9$.

Combining the above three cases, we obtain $P[(AB)(CD)]$ as follows:

$$P[(AB)(CD)] = P_{21}(T_1)P_{21}(T_2) + \tfrac{1}{3}P_{22}(T_1)P_{21}(T_2) + \tfrac{1}{3}P_{21}(T_1)P_{22}(T_2) + \tfrac{1}{9}P_{22}(T_1)P_{22}(T_2) = (1 - \tfrac{2}{3}e^{-T_1})(1 - \tfrac{2}{3}e^{-T_2}) . \quad (2)$$

The above equation is the product of two terms that are the $P$ values for three species [equation (1)]. This reflects the fact that a tree of four species can be decomposed into two trees of three species (see the case of five species).

Figure 2(b) shows the relationship among $T_1$, $T_2$, and $P$ in terms of isoclines. When one of $T_1$ and $T_2$ is small, $P$ cannot be increased substantially by increasing the other time-interval parameter. This indicates that if the species tree has a short internodal branch, a tree constructed from a single gene may be quite different from the species tree, irrespective of the length of the other internodal branch. The $P$ value is generally substantially lower in the case of four species than in the case of three species even if $T_1 = T_2 = T$. For example, the $P$ value for $T = 1$ is 0.75 in the case of three species but 0.57 in the case of four species. To have a $P$ value of 0.95, $T$ must be 2.6 ($5.2N$ generations) for the case of three species but 3.3 ($6.6N$ generations) for the case of four species.

The $P$ value for the case of species tree ((AB)C)D can be obtained in the same manner. In this case, there are five different possibilities (cases), in which a gene tree having the same topology as that of the species tree can be obtained (see fig. 4). Case (a) is the simplest one, in which all three ancestral species pass on one allele to each of their descendant species. The probability of occurrence of this case is obviously $P_{21}(T_1)P_{21}(T_2)$. In case (b), the first and third ancestral species pass on one allele to each of their descendant species, but the second ancestral species passes on two distinct alleles to the third ancestral species. The probability of occurrence of this event is $(1/3)P_{31}(T_1)P_{22}(T_2)$. The probabilities of all other events can be obtained in the same way. That is, the probabilities of occurrence of cases (c), (d), and (e) are $(1/3)P_{22}(T_1)P_{21}(T_2)$, $(1/9)P_{32}(T_1)P_{22}(T_2)$, and $(1/18)P_{33}(T_1)P_{22}(T_2)$, respectively. The probability that a gene tree has the same topology as that of the species tree ((AB)C)D is therefore given by

$$P[((AB)C)D] = P_{21}(T_1)P_{21}(T_2) + \tfrac{1}{3}P_{31}(T_1)P_{22}(T_2) + \tfrac{3}{3}P_{22}(T_1)P_{21}(T_2) + \tfrac{1}{9}P_{32}(T_1)P_{22}(T_2) + \tfrac{1}{18}P_{33}(T_1)P_{22}(T_2) = (1 - \tfrac{2}{3}e^{-T_1})(1 - \tfrac{2}{3}e^{-T_2}) - \tfrac{1}{9}e^{-(T_1 + T_2)}(1 - \tfrac{1}{2}e^{-2T_1}) . \quad (3)$$

This is somewhat different from equation (2), but as $T_1$ or $T_2$ or both increase, it approaches the latter. $P[(AB)(CD)]$ is always larger than $P[((AB)C)D]$. Therefore, equation (2) can be used as an upper-bound approximation for both topologies.

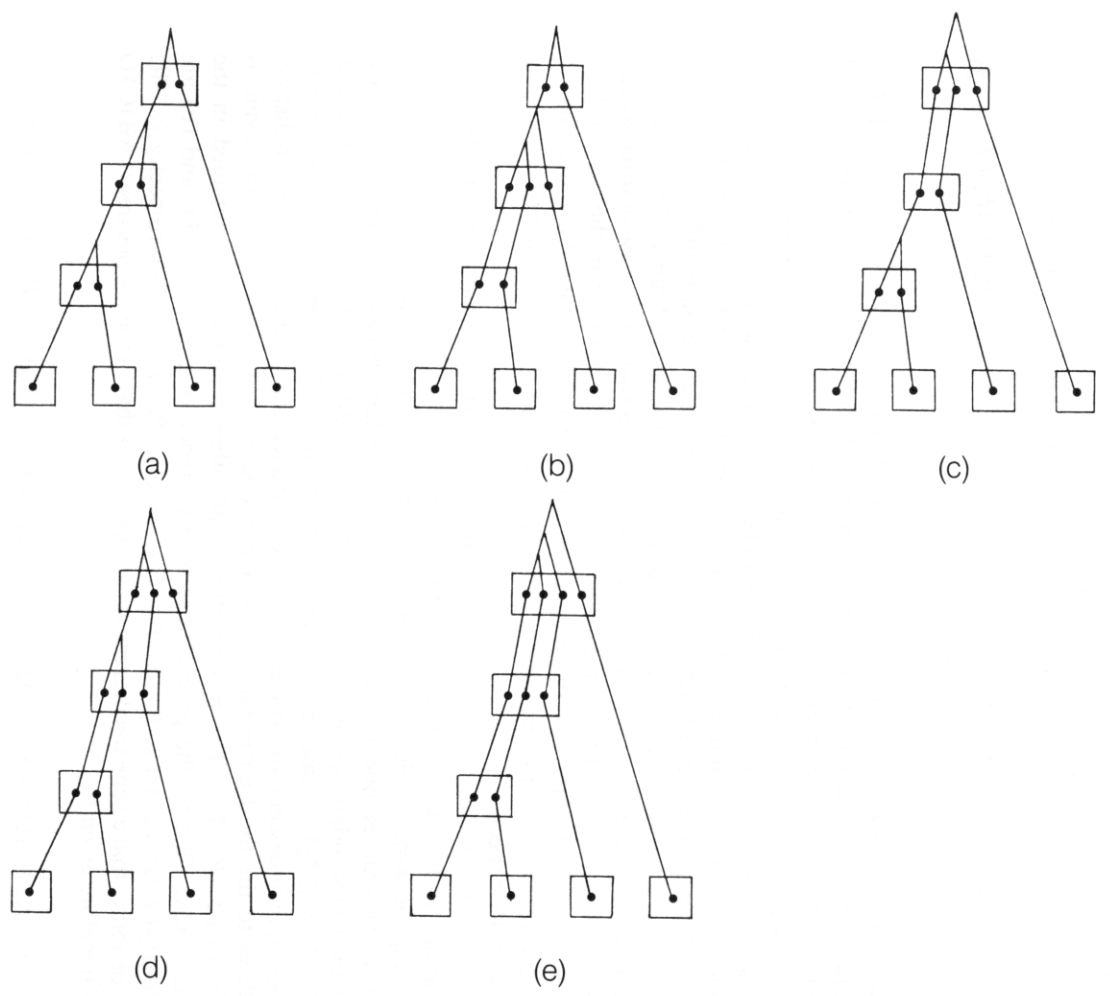


FIG. 4.—Five different ways of obtaining a gene tree that has the same topology as that of the species tree, ((AB)C)D.

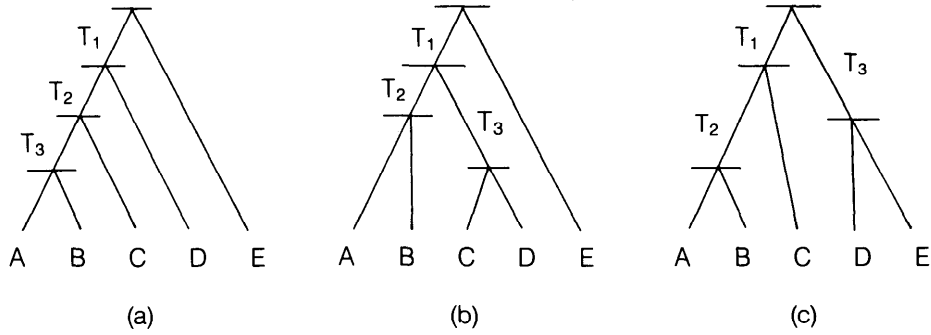


FIG. 5.—Three different types of species trees (unlabeled trees) that are possible for five species.

## Case of Five Species

In the case of five species, there are three unlabeled topologies (fig. 5), and we must evaluate the $P$ value for each of them. Evaluation of the $P$ value is tedious, because there are many different possibilities in which a given gene tree can be produced. For example, there are 15 different ways in which a gene tree identical with that in figure 5(a) can be produced. However, computation of the probability of occurrence of each event is straightforward; it can be computed by the same procedure as that for the cases of three and four species. The final results obtained are as follows:

$$P[(((AB)C)D)E] = (1 - \tfrac{2}{3}e^{-T_1})(1 - \tfrac{2}{3}e^{-T_2})(1 - \tfrac{2}{3}e^{-T_3}) - \tfrac{1}{9}e^{-T_2}[e^{-T_1} + e^{-T_3} - \tfrac{7}{6}e^{-(T_1 + T_3)} - \tfrac{1}{2}e^{-3T_1}(1 - \tfrac{1}{2}e^{-T_3}) - \tfrac{1}{2}e^{-(2T_2 + T_3)}(1 - \tfrac{1}{5}e^{-T_1} - \tfrac{1}{6}e^{-3T_1} - \tfrac{1}{30}e^{-6T_1})], \quad (4a)$$

$$P[((AB)(CD))E] = (1 - \tfrac{2}{3}e^{-T_1})(1 - \tfrac{2}{3}e^{-T_2})(1 - \tfrac{2}{3}e^{-T_3}) - \tfrac{1}{9}e^{-T_1}(1 - \tfrac{1}{2}e^{-2T_1})(e^{-T_2} + e^{-T_3} - \tfrac{4}{3}e^{-(T_2 + T_3)}) + \tfrac{1}{270}e^{-(T_1 + T_2 + T_3)}(4 - e^{-5T_1}), \quad (4b)$$

$$P[((AB)C(DE))] = (1 - \tfrac{2}{3}e^{-T_1})(1 - \tfrac{2}{3}e^{-T_2})(1 - \tfrac{2}{3}e^{-T_3}) - \tfrac{1}{9}e^{-(T_1 + T_2)} \times (1 - \tfrac{2}{3}e^{-T_3}) + \tfrac{1}{18}e^{-(3T_1 + T_2)}(1 - \tfrac{7}{10}e^{-T_3}), \quad (4c)$$

where $T_1$, $T_2$, and $T_3$ are the evolutionary time intervals as indicated in figure 5.

It should be noted that all of the above equations include the term

$$P_A = (1 - \tfrac{2}{3}e^{-T_1})(1 - \tfrac{2}{3}e^{-T_2})(1 - \tfrac{2}{3}e^{-T_3}) \quad (5)$$

and that other terms are negligibly small when at least two of $T_1$, $T_2$, and $T_3$ are sufficiently large. For example, the difference between equation (5) and any of equations (4a), (4b), and (4c) is $<0.041$ when $T_i > 0.5$ for all $i$'s and $<0.022$ when $T_i > 1$. This indicates that equation (5) can be used as an approximation of any of equations (4a), (4b), and (4c) unless two or three of the $T_i$'s are extremely small.

Equation (5) indicates that $P_A$ is determined by three independent components,

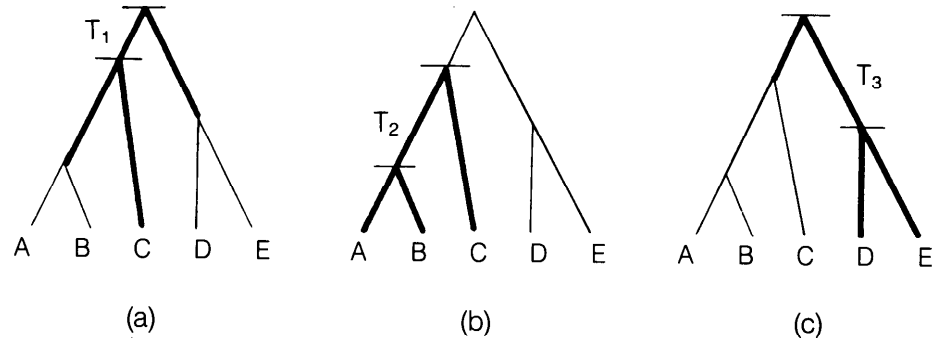


FIG. 6.—Decomposition of the species tree for five species into three trees of three "species." Note that three-component trees are not really independent.

each of which represents the $P$ value for the case of three species. This is because a tree of five species can be decomposed into three trees of three species though the component trees are not really independent (see fig. 6). Although equation (5) is an approximate formula, it represents a good property for studying the relationship between gene trees and species trees. As in the case of four species, it can be used as an upper-bound approximation for $P$. Figure 2(c) gives the $P_A$ values as a function of $T_1$, $T_2$, and $T_3$, where $T_2 = T_3$ is assumed. The isoclines for $P_A$ have a shape similar to that for the case of four species, but to achieve the same $P_A$ value larger values of $T_1$, $T_2$, and $T_3$ are required. For example, the $P_A$ for $T_1 = T_2 = T_3 = T = 1$ is now 0.43, and the $T$ value required for $P_A = 0.95$ is 3.7 ($7.4N$ generations). The isoclines in figure 2(c) remain the same when $T_1$ is interchanged with $T_2$ or $T_3$.

### General Case

When there are more than five species, computation of the exact $P$ value becomes very complicated. However, the upper bound of the $P$ value can easily be obtained. A tree with $n$ species has $n - 2$ internodal branches and thus can be decomposed into $n - 2$ units of three species. As we have seen above, each unit gives a nearly independent contribution to the $P$ value when $T_i$ is large. Therefore, the upper bound is given by

$$P_A = \prod_{i=1}^{n-2} (1 - \tfrac{2}{3}e^{-T_i}) . \qquad (6)$$

We have not evaluated the difference between this value and the exact value for the case of $n > 5$. However, even this upper bound can be quite small when $n$ is large, and to get a rough idea of the accuracy of a gene tree the above equation seems to be useful.

## Phylogenetic Trees Constructed from Two Alleles at a Locus from Each Species

When two alleles are sampled from each species at a locus, the number of possible gene trees becomes very large even for a small number of species involved. In the following we consider only the case of three species.

When two alleles are sampled from each of three species, the six alleles may have been derived from two, three, four, five, or six different alleles in the common ancestral

**Table 1**
**Eighteen Different Ways of Producing Gene Trees Caused by Polymorphisms in Ancestral Species X and Y of Figure 1, and Their Probabilities[a]**

| Ancestral Polymorphism[b] | | | Consistency between Gene Tree and Species Tree | | |
|---|---|---|---|---|---|
| (AB);C | A;B | Probability | Consistent | Inconsistent | Unresolvable |
| 1;1 | 1;1 | $P_{21}(T_1)P^2_{21}(T_2)P_{21}(T_1 + T_2)$ | 1 | ... | ... |
| 2;1 | 1;1 | $P_{21}(T_1)P^2_{21}(T_2)P_{22}(T_1 + T_2)$ | 1 | ... | ... |
| 1;2 | 1;1 | $P_{22}(T_1)P^2_{21}(T_2)P_{21}(T_1 + T_2)$ | $1/3$ | $2/3$ | ... |
| 2;2 | 1;1 | $P_{22}(T_1)P^2_{21}(T_2)P_{22}(T_1 + T_2)$ | $2/9$ | $5/9$ | $2/9$ |
| 1;1 | 1;2 | $2P_{31}(T_1)P_{21}(T_2)P_{22}(T_2)P_{21}(T_1 + T_2)$ | 1 | ... | ... |
| 2;1 | 1;2 | $2P_{31}(T_1)P_{21}(T_2)P_{22}(T_2)P_{22}(T_1 + T_2)$ | 1 | ... | ... |
| 1;2 | 1;2 | $2P_{32}(T_1)P_{21}(T_2)P_{22}(T_2)P_{21}(T_1 + T_2)$ | $5/9$ | $2/9$ | $2/9$ |
| 2;2 | 2;2 | $2P_{32}(T_1)P_{21}(T_2)P_{22}(T_2)P_{22}(T_1 + T_2)$ | $17/27$ | $5/27$ | $5/27$ |
| 1;3 | 1;2 | $2P_{33}(T_1)P_{21}(T_2)P_{22}(T_2)P_{21}(T_1 + T_2)$ | $3/18$ | $9/18$ | $6/18$ |
| 2;3 | 1;2 | $2P_{33}(T_1)P_{21}(T_2)P_{22}(T_2)P_{22}(T_1 + T_2)$ | $19/180$ | $85/180$ | $76/180$ |
| 1;1 | 2;2 | $P_{41}(T_1)P^2_{22}(T_2)P_{21}(T_1 + T_2)$ | 1 | ... | ... |
| 2;1 | 2;2 | $P_{41}(T_1)P^2_{22}(T_2)P_{22}(T_1 + T_2)$ | 1 | ... | ... |
| 1;2 | 2;2 | $P_{42}(T_1)P^2_{22}(T_2)P_{21}(T_1 + T_2)$ | $19/27$ | $2/27$ | $6/27$ |
| 2;2 | 2;2 | $P_{42}(T_1)P^2_{22}(T_2)P_{22}(T_1 + T_2)$ | $65/81$ | $5/81$ | $11/81$ |
| 1;3 | 2;2 | $P_{43}(T_1)P^2_{22}(T_2)P_{21}(T_1 + T_2)$ | $11/54$ | $9/54$ | $34/54$ |
| 2;3 | 2;2 | $P_{43}(T_1)P^2_{22}(T_2)P_{22}(T_1 + T_2)$ | $36/90$ | $14/90$ | $40/90$ |
| 1;4 | 2;2 | $P_{44}(T_1)P^2_{22}(T_2)P_{21}(T_1 + T_2)$ | $1/10$ | $3/10$ | $6/10$ |
| 2;4 | 2;2 | $P_{44}(T_1)P^2_{22}(T_2)P_{22}(T_1 + T_2)$ | $49/675$ | $242/675$ | $384/675$ |

NOTE.—The species tree considered is (AB)C with $T_1$ beween the two speciation events and $T_2$ since divergence between species A and B (time between $t_1$ and $t_2$ divided by $2N$ in fig. 1).

[a] Two alleles are assumed to have been sampled from each of the three species.

[b] $i;j$ represents the case where $i$ alleles enter into the first species [(AB) or A] and in which $j$ alleles enter into the second species (C or B).

species X (see fig. 1). The ancestral species Y also may have contributed two, three, or four alleles to its two descendant species. Therefore, there are 18 different ways of producing gene trees even if only these two factors are considered (table 1). If we take into account the ways in which these alleles can enter into their descendant species, the total number of different gene trees becomes >350.

When two alleles are sampled from each species, it is not always easy to determine whether a gene tree is consistent with the species tree, because the two alleles from each species often show different evolutionary relationships among different species (see fig. 7). However, it is possible to compute the average number of nucleotide differences between genes for each pair of species under the assumption that the number is proportional to evolutionary time. We can then use this number as a measure of interspecific genetic distance and construct a gene tree. This gene tree will be called a composite tree. The topology of the composite gene tree can then be compared with that of the species tree.

Some composite gene trees clearly show the same topology as that of the species tree, even if the two alleles sampled from the same species may not share the most recent common ancestral gene [fig. 7(a)–7(c)]. Trees (d) and (e) in figure 7 give a composite tree whose topology is consistent with that of the species tree. (Composite

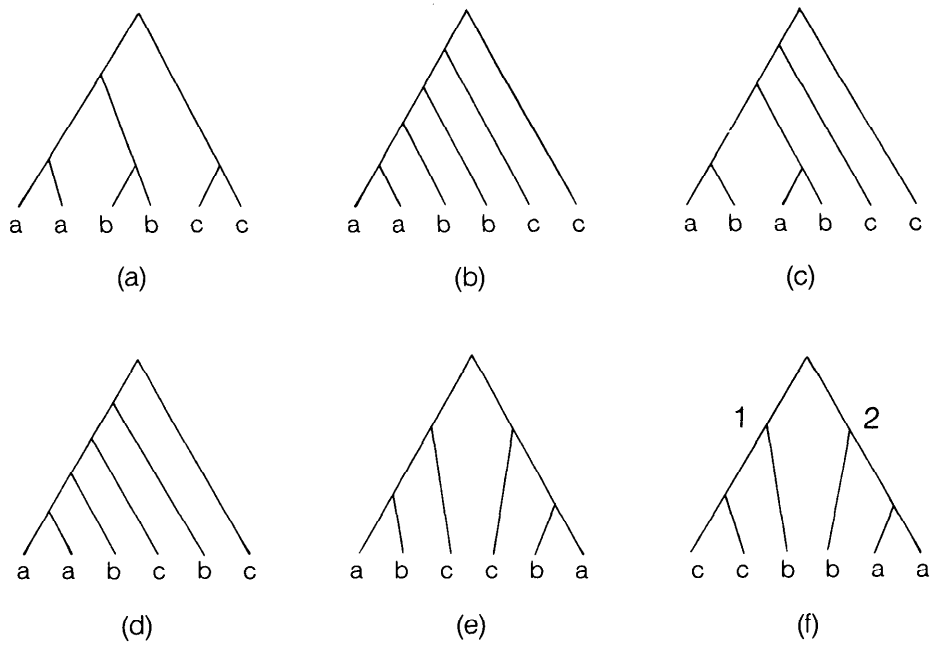


FIG. 7.—Some examples of gene trees for the case in which two alleles are sampled at a locus from each of three species. Letters a, b, and c represent the alleles sampled from species A, B, and C, respectively. The species tree is the same as that of fig. 1. Numbers 1 and 2 represent branching events 1 and 2 mentioned in text.

trees are not drawn.) In the case of tree (f), the conclusion depends on the times of occurrence of the branching events 1 and 2 of the gene tree. If there were three genes at the time of the first species splitting, event 1 must have preceded event 2. In this case, the expected genetic distance between A and B is less than that between A and C or between B and C, so that the composite tree is consistent with the species tree. When there were five or six alleles at the time of the first species splitting, however, it is impossible to know which event, 1 or 2, took place earlier. (A similar situation occurs for some cases of four ancestral alleles.) In the following, we classify this case as unresolvable.

The probabilities of occurrence of the 18 different possible cases with respect to the polymorphism in the ancestral species can be computed by using appropriate $P_{ij}$'s. The results obtained are presented in table 1. In this table, $T_1$ is equal to $T$ in figure 1, and $T_2$ is the time between $t_1$ and $t_2$ divided by $2N$. The probability of occurrence of each case depends on $T_2$ as well as on $T_1$ because two alleles are sampled from each species. The probabilities of occurrence of consistent, inconsistent, and unresolvable composite gene trees for each of the above 18 possible cases are also presented in table 1. Therefore, if $T_1$ and $T_2$ are specified, the overall probability ($P$) that a gene tree has the same topology as that of the species tree can be computed. The results obtained are presented in table 2.

Table 2 shows that when $T_1 = T_2 = 0$ the $P$ value is very small and that most trees are "unresolvable." This occurs because there is no real species tree in this case, and the polymorphic alleles in the ancestral species are merely distributed into three groups (species). When $T_1$ remains 0, the $P$ value increases with increasing $T_2$ and eventually reaches 0.333, which is the $P$ value for the case of a single allele sampled

**Table 2**
**Probabilities of Occurrence of Consistent (P), Inconsistent (Q), and Unresolvable (R) Gene Trees for the Case of Two Alleles Sampled from Each Species, in Comparison with Those for the Case of One Allele Sampled**

| | | Two Alleles Sampled | | | One Allele Sampled | |
|---|---|---|---|---|---|---|
| $T_1$ | $T_2$ | P | Q | R | P | Q |
| 0 | 0 | 0.073 | 0.359 | 0.569 | 0.333 | 0.667 |
| | 1 | 0.196 | 0.521 | 0.283 | | |
| | 2 | 0.277 | 0.609 | 0.114 | | |
| | 3 | 0.312 | 0.645 | 0.043 | | |
| | 4 | 0.325 | 0.659 | 0.016 | | |
| | 5 | 0.330 | 0.664 | 0.006 | | |
| 0.5 | 0 | 0.606 | 0.104 | 0.290 | 0.596 | 0.404 |
| | 1 | 0.571 | 0.279 | 0.151 | | |
| | 2 | 0.584 | 0.356 | 0.062 | | |
| | 3 | 0.590 | 0.386 | 0.024 | | |
| | 4 | 0.594 | 0.398 | 0.009 | | |
| | 5 | 0.595 | 0.402 | 0.003 | | |
| 1 | 0 | 0.795 | 0.052 | 0.153 | 0.755 | 0.245 |
| | 1 | 0.754 | 0.163 | 0.083 | | |
| | 3 | 0.753 | 0.233 | 0.014 | | |
| | 5 | 0.755 | 0.244 | 0.002 | | |
| 2 | 0 | 0.929 | 0.018 | 0.053 | 0.910 | 0.090 |
| | 1 | 0.912 | 0.059 | 0.029 | | |
| | 3 | 0.910 | 0.086 | 0.005 | | |
| | 5 | 0.910 | 0.090 | 0.001 | | |
| 3 | 0 | 0.974 | 0.007 | 0.020 | 0.967 | 0.033 |
| | 1 | 0.968 | 0.022 | 0.010 | | |
| | 5 | 0.967 | 0.033 | 0.000 | | |
| 4 | 0 | 0.990 | 0.002 | 0.007 | 0.988 | 0.012 |
| | 1 | 0.988 | 0.008 | 0.004 | | |
| | 5 | 0.988 | 0.012 | 0.000 | | |
| 5 | 0 | 0.996 | 0.001 | 0.003 | 0.996 | 0.004 |
| | 5 | 0.996 | 0.004 | 0.000 | | |

NOTE.—The species tree considered is (AB)C with $T_1$ between the two speciation events and $T_2$ since divergence between species A and B (time between $t_1$ and $t_2$ divided by $2N$ in fig. 1).

from each species. When $T_1 = 0.5$, the $P$ value is quite high (0.606) even for $T_2 = 0$ but does not increase with increasing $T_2$.

When $T_1 \geq 0.5$, the $P$ value for $T_2 = 0$ is higher than that for $T_2 = \infty$. Therefore, there is some benefit for studying two alleles rather than one from each species, but the benefit is quite small. Furthermore, as $T_2$ increases, $P$ declines below the value for $T_2 = \infty$ and then again starts to increase to the level for the case of single alleles studied. In the case of $T_1 \geq 2$, $P$ is generally higher than that for the case of single alleles sampled for most $T_2$ values. However, the benefit of studying two alleles from each species is again small. In general, the $P$ value is determined mainly by $T_1$ and is hardly affected by $T_2$ except when $T_1$ is very small.

## Some Applications

The above study indicates that the topological error introduced by sequence polymorphism in ancestral species is substantial when the evolutionary time considered is short and when the effective population size is large. To get some idea about the actual magnitude of errors, let us consider two examples from hominoid and human evolution.

The first example is the phylogeny for humans, chimpanzees, gorillas, orangutans, and gibbons. This phylogeny has recently been studied by a number of authors (e.g., Ferris et al. 1981; Brown et al. 1982; Goodman et al. 1984; Sibley and Ahlquist 1984; Nei et al. 1985), but no concensus has been obtained. Analyzing Brown et al.'s (1982) data on partial sequences (896 bp) of mitochondrial DNA (mtDNA), Nei (1985) obtained the topology of the form given in figure 5(a) when humans, chimpanzees, gorillas, orangutans, and gibbons are represented by A, B, C, D, and E, respectively. He also estimated that the evolutionary times corresponding to $t_1$ ($\equiv 2NT_1$), $t_2$, and $t_3$ in figure 5(a) are $2.0 \times 10^6$, $5.2 \times 10^6$, and $1.2 \times 10^6$ years, respectively. These estimates actually refer to the evolutionary times of the gene tree estimated, but let us assume that they also refer to the evolutionary times of the species tree. The generation time and the effective population size in early hominoid evolution seem to have been $\sim$15 years and $\sim 10^4$ years, respectively (Nei and Graur 1984). (The effective size for mtDNA is about one-quarter of that for nuclear genes, but let us assume that $N = 10^4$ for mtDNA for simplicity.) If we use these estimates, we obtain $T_1 = 2.0 \times 10^6/(2 \times 10^4 \times 15) = 6.7$, $T_2 = 17.3$, and $T_3 = 4.0$. The approximate probability that a gene tree has the same topology as that of the species tree then becomes 0.987, from equation (5). This probability is quite high. However, our estimates of $t_i$ and $N$ could be wrong. Particularly, $t_3$ could be much smaller than Nei's estimate (see Sarich and Wilson 1967). If $T_3 = 1.0$ but $T_1$ and $T_2$ remain the same, $P$ becomes 0.754. Therefore, the effect of ancestral polymorphism could be substantial.

The second example is the phylogeny of the three major races of man, caucasoids, negroids, and mongoloids. This phylogeny has also been controversial in recent years (e.g., see Cavalli-Sforza and Bodmer 1971; Nei and Roychoudhury 1974, 1982; Wainscoat et al. 1986). Using gene frequency data from 62 protein loci, Nei and Roychoudhury (1982) obtained a (population) phylogeny similar to that in figure 1 when caucasoids, mongoloids, and negroids are represented by A, B, and C, respectively. They also estimated that the time of divergence between negroids and the caucasoid-mongoloid line is 115,000 years ago, whereas the time of divergence between caucasoids and mongoloids is 41,000 years ago. If we assume that the generation time and the long-term effective population size for these three groups were 20 years and $10^4$, respectively, the $T$ in figure 1 becomes $74{,}000/(2 \times 10^4 \times 20) = 0.185$. Therefore, $P$ is 0.446. (If we use $N = 4 \times 10^4$, $P$ becomes 0.363.) Note that when $T$ is small, the $P$ value cannot be increased substantially by increasing the number of alleles sampled. Therefore, information on nucleotide sequences from a single locus would not resolve the problem of the phylogeny of the three major races of man.

This conclusion seems to be important in the interpretation of results from recent studies on human mtDNA. mtDNA contains $\sim$16,000 nucleotides but is inherited as a single entity without recombination, so that it is equivalent to a single gene in the present paper. A number of authors (e.g., Johnson et al. 1983; Nei 1985; Cann et al. 1987) have obtained an mtDNA phylogeny consistent with that obtained from

protein loci. However, the present study indicates that this is not a strong support of the phylogeny because the accuracy of a gene tree is low. This view is also supported by the large standard error of the estimate of net nucleotide substitutions between populations (Nei 1985). Recently, using restriction-site data for the β-globin gene complex, Wainscoat et al. (1986) also obtained a gene tree that is consistent with the protein phylogeny. Here again, however, the accuracy of the tree is quite low. To obtain a more reliable tree, one must use DNA sequences from many loci that have evolved independently of each other.

## Discussion

We have seen that the probability of a gene tree having the same topology as the species tree is quite small when any of the $T_i$'s is relatively small and that this probability cannot be increased substantially by increasing the number of alleles sampled at a locus. However, if there are DNA sequence data for many different loci that have evolved independently, the species tree can be inferred from gene trees more accurately. Let us now study this problem, following Saitou and Nei (1986).

We consider the case of three species and assume that one gene (DNA sequence) from each of $r$ independent loci is studied for each of the three species. There are three different possible gene trees for each locus, i.e., (ab)c, (ac)b, and (bc)a. Let us denote these trees by α, β, and γ, respectively, and assume that the topology of the species tree is (AB)C. When $r$ loci are examined, $i$, $j$, and $k$ loci may show gene trees α, β, and γ, respectively ($i + j + k = r$). Inference of the species tree is made by using $i$, $j$, and $k$, and the correct topology is obtained when $i > j$ and $i > k$. The probability ($P_T$) of occurrence of this event can be evaluated by using the following trinomial distribution:

$$P(i, j, k) = \frac{r!}{i!j!k!} P^i Q_1^j Q_2^k, \tag{7}$$

where $P$, $Q_1$, and $Q_2$ are the probabilities of occurrence of trees α, β, and γ, respectively. We know that $P = 1 - (2/3)\exp(-T)$, whereas $Q_1$ and $Q_2$ are both $(1/3)\exp(-T)$. Therefore, $P_T$ is given by the sum of equation (7) for all cases in which $i > j$ and $i > k$. Saitou and Nei (1986) evaluated $P_T$ for several values of $T$ and $r$. When $T < 1$, $P_T$ increases rather slowly with increasing $r$. Those authors have also shown that the number of loci required for obtaining the correct species tree with a probability of 95% is three for $T = 2$, five for $T = 1.5$, seven for $T = 1$, and 14 for $T = 0.5$.

A similar computation can be made for the case of a larger number of species. However, the computation becomes extremely complicated as the number of species increases. This is because the number of possible gene trees rapidly increases with increasing number of species. For example, the number of possible gene trees for the case of four species is 15, and we must establish the probability of occurrence of each of these trees. Once this probability is obtained, we can evaluate $P_T$ by considering a multinomial distribution with 14 independent variables.

Because of the tedious computation involved, we have evaluated the $P_T$ value only for the case of four species. In this case, there are two unlabeled species topologies, (AB)(CD) and ((AB)C)D, as mentioned earlier. We computed the probabilities of occurrence of the 15 gene trees for each of these species topologies and examined the probability of obtaining the correct species tree from data for $r$ independent loci. In

**Table 3**
**Probabilities of Obtaining Correct Species Trees ($P_T$) from Data for $r$ Independent Loci**

| | r | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $T_1(=T_2)$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| | | | | A. (AB)(CD)[a] | | | | | |
| 1.0 ..... | 0.60 | 0.73 | 0.78 | 0.82 | 0.86 | 0.90 | 0.92 | 0.94 | 0.95 |
| 1.5 ..... | 0.81 | 0.89 | 0.92 | 0.95 | 0.97 | 0.98 | 0.99 | 0.99 | 1.00 |
| 2.0 ..... | 0.92 | 0.96 | 0.98 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | | | B. ((AB)C)D[a] | | | | | |
| 1.0 ..... | 0.58 | 0.70 | 0.75 | 0.80 | 0.84 | 0.88 | 0.90 | 0.92 | 0.94 |
| 1.5 ..... | 0.81 | 0.88 | 0.92 | 0.95 | 0.97 | 0.98 | 0.99 | 0.99 | 1.00 |
| 2.0 ..... | 0.92 | 0.96 | 0.98 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

[a] Topology of the species tree.

this computation, we assumed that the most common tree among the $r$ gene trees produced is inferred as the species tree. The $P_T$ value was then computed by counting the cases where the inference of the species tree is correct. When there was no most common gene tree, we assumed that the species tree was not obtained.

The results for some $T_1$ and $T_2$ values are presented in table 3. When $T_1 = T_2$, the two different species trees (topologies) have little effect on the $P_T$ value as long as $T_1 = T_2 \geq 1$. The number of loci required for obtaining the species tree with a probability of 95% is only slightly larger than that for the case of three species when $T_1 = T_2 \geq 1.5$ but substantially larger when $T_1 = T_2 \leq 1$.

As was mentioned in the Introduction, this paper is not intended to examine the effect of the number of nucleotides examined on the topology of an estimated tree. When the number of nucleotides examined is small, an erroneous topology may be obtained even if the expected gene tree happens to be identical with the species tree. When the species examined are relatively closely related, the number of nucleotides ($m^*$) required for obtaining the correct species tree with a probability of 95% is substantial. For example, Saitou and Nei (1986) have shown that in order to have a reliable tree for mtDNAs of humans, chimpanzees, gorillas, orangutans, and gibbons, at least 2,600–2,700 nucleotides must be examined. This number is considerably larger than the number currently available (Brown et al. 1982; Hixon and Brown 1986). In the case of nuclear genes, an even larger number of nucleotides are necessary, since the rate of nucleotide substitution is ~10 times lower in nuclear genes than in mtDNA. It is therefore important to keep in mind that the probability of an estimated gene tree having the same topology as that of the species tree is considerably lower than the values given in the present paper when relatively short DNA sequences are used for constructing gene trees.

**Acknowledgment**

We thank Naoyuki Takahata for his comments on an earlier version of this paper. This study was supported by research grants from the National Institutes of Health and the National Science Foundation.

LITERATURE CITED

BROWN, W. M., E. M. PRAGER, A. WANG, and A. C. WILSON. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18**:225–239.

CANN, R. L., M. STONEKING, and A. C. WILSON. 1987. Mitochondrial DNA and human evolution. Nature **325**:31–36.

CAVALLI-SFORZA, L. L., and W. F. BODMER. 1971. The genetics of human populations. W. H. Freeman, San Francisco.

FERRIS, S. D., A. C. WILSON, and W. M. BROWN. 1981. Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. Proc. Natl. Acad. Sci. USA **78**:2432–2436.

GOODMAN, M., B. F. KOOP, J. CZELUSNIAK, M. L. WEISS, and J. L. SLIGHTOM. 1984. The η-globin gene: its long evolutionary history in the β-globin gene family of mammals. J. Mol. Biol. **180**:803–823.

HIXSON, J. E., and W. M. BROWN. 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.

JOHNSON, M. J., D. C. WALLACE, S. D. FERRIS, M. C. RATTAZZI, and L. L. CAVALLI-SFORZA. 1983. Radiation of human mitochondrial DNA types analyzed by restriction endonuclease cleavage patterns. J. Mol. Evol. **19**:255–271.

KINGMAN, J. F. C. 1982. On the genealogy of large populations. J. Appl. Probab. **19A**:27–43.

NEI, M. 1985. Human evolution at the molecular level. Pp. 41–64 *in* T. OHTA and K. AOKI, eds., Population genetics and molecular evolution. Japan Scientific Societies Press, Tokyo.

———. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

NEI, M., and D. GRAUR. 1984. Extent of protein polymorphism and the neutral mutation theory. Evol. Biol. **17**:73–118.

NEI, M., and A. K. ROYCHOUDHURY. 1974. Genetic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. Am. J. Hum. Genet. **26**:421–443.

———. 1982. Genetic relationship and evolution of human races. Evol. Biol. **14**:1–59.

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

NEIGEL, J. E., and A. C. AVISE. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. Pp. 515–534 *in* S. KARLIN and E. NEVO, eds., Evolutionary processes and theory. Academic Press, New York.

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

SARICH, V. M., and A. C. WILSON. 1967. Immunological time scale for hominid evolution. Science **158**:1200–1203.

SIBLEY, C. G., and J. E. AHLQUIST. 1984. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J. Mol. Evol. **20**:2–15.

TAJIMA, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437–460.

TAKAHATA, N., and M. NEI. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110**:325–344.

TATENO, Y., M. NEI, and F. TAJIMA. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J. Mol. Evol. **18**:387–404.

TAVARÉ, S. 1984. Line-of-descent and genealogical processes, and their application in population genetics models. Theor. Popul. Biol. **26**:119–164.

WAINSCOAT, J. S., A. V. S. HILL, A. L. BOYCE, J. FLINT, M. HERNANDEZ, S. L. THEIN, J. M. OLD, J. R. LYNCH, A. G. FALUSI, D. J. WEATHERALL, and J. B. CLEGG. 1986. Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. Nature **319**:491–493.

WATTERSON, G. A. 1984. Lines of descent and the coalescent. Theor. Popul. Biol. **26**:77–92.

WALTER M. FITCH, reviewing editor

Received October 20, 1987; revision received May 6, 1988