Methods for Computing the Standard Errors of Branching Points in an Evolutionary Tree and Their Application to Molecular Data from Humans and Apes[1]

Masatoshi Nei, J. Claiborne Stephens, and Naruya Saitou
Center for Demographic and Population Genetics, University of Texas at Houston

Statistical methods for computing the standard errors of the branching points of an evolutionary tree are developed. These methods are for the unweighted pair-group method–determined (UPGMA) trees reconstructed from molecular data such as amino acid sequences, nucleotide sequences, restriction-sites data, and electrophoretic distances. They were applied to data for the human, chimpanzee, gorilla, orangutan, and gibbon species. Among the four different sets of data used, DNA sequences for an 895-nucleotide segment of mitochondrial DNA (Brown et al. 1982) gave the most reliable tree, whereas electrophoretic data (Bruce and Ayala 1979) gave the least reliable one. The DNA sequence data suggested that the chimpanzee is the closest and that the gorilla is the next closest to the human species. The orangutan and gibbon are more distantly related to man than is the gorilla. This topology of the tree is in agreement with that for the tree obtained from chromosomal studies and DNA-hybridization experiments. However, the difference between the branching point for the human and the chimpanzee species and that for the gorilla species and the human-chimpanzee group is not statistically significant. In addition to this analysis, various factors that affect the accuracy of an estimated tree are discussed.

Introduction

In recent years, an increasing number of authors have been using molecular data to construct evolutionary trees of species. There are several different methods for constructing evolutionary trees, but in all of them the accuracy of a reconstructed tree is quite low, unless the lengths (number of mutational changes) of all branches are sufficiently large (Peacock and Boulter 1975; Tateno et al. 1982; Nei et al. 1983). There are two types of errors involved in a reconstructed tree: (1) topological errors and (2) errors in the estimates of branch lengths. These two types of errors are intricately related, since the topology and branch lengths are usually estimated simultaneously. If we know the pattern of amino acid replacement or nucleotide substitution in evolution, the standard error (SE) of an estimate of evolutionary distance between a pair of species can be computed relatively easily (e.g., Kimura 1969; Kimura and Ohta 1972; Nei 1978; Nei and Tajima 1983). Some authors (e.g., Kumazaki et al. 1983) have substituted this error for the SE of a branching point. However, this does not give a correct value except under certain circumstances, and we need a general method for computing the SE of a branching point.

1. Key words: topological errors, UPGMA trees, molecular phylogeny.

Address for correspondence and reprints: Dr. Masatoshi Nei, Center for Demographic and Population Genetics, University of Texas at Houston, P.O. Box 20334, Houston, Texas 77225.

Mol. Biol. Evol. 2(1):66–85. 1985.

0737-4038/85/0201-0009$02.00

Evaluation of the SE is important because each branching point suggests that an important event of speciation or population splitting occurred there. In practice, however, topological errors are often introduced at the time of tree reconstruction, and an estimated topology may not represent the actual process of evolutionary changes of genes or organisms (Tateno et al. 1982; Nei et al. 1983).

In this paper we present methods for computing the SE values of branching points for a tree reconstructed by the UPGMA method (see Sneath and Sokal 1973). The UPGMA is known as a method of phenetic clustering, but Nei (1975) suggested that this method would give a good evolutionary tree when the expected rate of gene substitution is constant but the actual number of substitutions is subject to stochastic errors. Later, using computer simulation, Tateno et al. (1982) and Nei et al. (1983) showed that UPGMA is more efficient for getting the correct (true) tree than are several other alternative methods when a tree is reconstructed from genetic distances the expectations of which are proportional to evolutionary time. In this paper we consider four different types of molecular data: (1) amino acid–replacement data, (2) nucleotide substitution data, (3) restriction-sites data, and (4) electrophoretic data. We then apply the methods developed to study the reliability of the evolutionary trees reconstructed for the human and several ape species.

Theory

Amino Acid Replacement Data

Let us consider the evolutionary tree given in figure 1. In this figure, numerals 1, 2, 3, and 4 represent four different species, whereas numerals 5, 6, and 7 stand for branching points (or ancestral species). For simplicity, we assume that the rate of amino acid replacement for a protein is λ per year per amino acid site and is the same for all amino acids. We also assume that the evolutionary time considered is relatively short, so that parallel mutations in different evolutionary lineages or back mutations in the same lineage are negligible. (The effect of violation of these assumptions will be discussed later.) Under these assumptions, the expected number of amino acid replacements between a pair of species is given by $d = 2\lambda t$, where t is the time since divergence between the two species. If the total number of amino acids compared between the two species is n and the proportion of identical amino acids between them is i, then d is estimated by

$$d = -\log_e i \qquad (1)$$

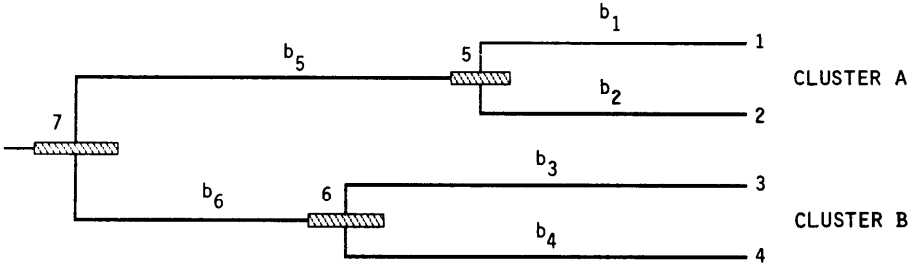

FIG. 1.—A hypothetical evolutionary tree. *1*, *2*, *3*, and *4* represent extant species, whereas *5*, *6*, and *7* represent branching points (or ancestral species). The hatched boxes represent the magnitude of SE.

with sampling variance

$$V(d) = \frac{1 - i}{in} \tag{2}$$

(e.g., see Nei [1975], p. 14).

It should be noted that if we knew the number of amino acid replacements for all amino acid sites, the variance of the number of amino acid replacements per site would have been $2\lambda t/n$ under the Poisson process. In practice, however, it is impossible to know these numbers, and we must estimate d by equation (1). Since equation (1) is based on incomplete information on amino acid replacements, equation (2) gives a variance larger than $2\lambda t/n$.

In a tree reconstructed by UPGMA, the branching point between two species (species 1 and 2 in fig. 1) is given by $b = d/2$ with variance $V(b) = V(d)/4$, where $b = b_1 = b_2$ in figure 1. Therefore, the SE of this branching point (5 in fig. 1) is given by the square root of V(b).

In a UPGMA tree, the distance between cluster A (species 1 and 2) and cluster B (species 3 and 4) in figure 1 is given by

$$d_{AB} = \frac{d_{13} + d_{14} + d_{23} + d_{24}}{4}, \tag{3}$$

where d_{jk} is the number of amino acid replacements between species j and k, and d_{12} and d_{34} are assumed to be smaller than d_{13}, d_{14}, d_{23}, and d_{24}. The d_{jk} value is estimated by the equation $d_{jk} = -\log_e i_{jk}$, where i_{jk} is the proportion of identical amino acids between species j and k. From equation (3), the variance of d_{AB} is given by

$$V(d_{AB}) = {}^1/_{16}[V(d_{13}) + V(d_{14}) + V(d_{23}) + V(d_{24}) + 2\,\mathrm{Cov}(d_{13}, d_{14}) + 2\,\mathrm{Cov}(d_{13}, d_{23}) + 2\,\mathrm{Cov}(d_{13}, d_{24}) + 2\,\mathrm{Cov}(d_{14}, d_{23}) + 2\,\mathrm{Cov}(d_{14}, d_{24}) + 2\,\mathrm{Cov}(d_{23}, d_{24})], \tag{4}$$

and the variance of branching point 7 in figure 1 is $V(b_{AB}) = V(d_{AB})/4$.

The variance $V(d_{jk})$ in equation (4) is given by equation (2), whereas the covariance between d_{13} and d_{14} is

$$\mathrm{Cov}(d_{13}, d_{14}) = \mathrm{Cov}(b_1 + b_5 + b_6 + b_3, b_1 + b_5 + b_6 + b_4) = V(d_{16}) = \frac{1 - i_{16}}{i_{16}n}, \tag{5}$$

because b_3 and b_4 are independent (see fig. 1). Here, d_{16} is $b_1 + b_5 + b_6$. In equation (5), i_{16} cannot be measured from actual data, since the ancestral species 6 does not exist. However, it can be estimated by

$$i_{16} = e^{-(b_1+b_5+b_6)} = e^{-d_{AB}+d_{34}/2}.$$

Similarly, all the other covariances can be obtained.

$$\mathrm{Cov}(d_{13}, d_{23}) = \frac{1 - i_{35}}{i_{35}n},$$

$$\mathrm{Cov}(d_{13}, d_{24}) = \mathrm{Cov}(d_{14}, d_{23}) = \frac{1 - i_{56}}{i_{56}n},$$

$$\mathrm{Cov}(d_{14}, d_{24}) = \frac{1 - i_{45}}{i_{45}n},$$

$$\mathrm{Cov}(d_{23}, d_{24}) = \frac{1 - i_{26}}{i_{26}n}.$$

These covariances can be computed from the values of d_{AB} and d_{jk}. For example,

$$i_{56} = e^{-d_{AB}+(d_{12}+d_{34})/2}.$$

Therefore, the SE of branching point 7 in figure 1 can be estimated.

It is now obvious that this procedure can be used for any branching point in a UPGMA tree irrespective of the number of species involved. In general, the UPGMA distance between two species clusters A and B is given by

$$d_{AB} = \frac{\sum_{jk} d_{jk}}{rs}, \tag{6}$$

where d_{jk} is the *intercluster* distance between the *j*th species in cluster A and the *k*th species in cluster B, and r and s are the numbers of species in clusters A and B, respectively. Therefore, the variance of d_{AB} is

$$V(d_{AB}) = \frac{\sum V(d_{jk}) + \sum \mathrm{Cov}(d_{jk}, d_{lm})}{(rs)^2}. \tag{7}$$

There are rs variances and rs(rs − 1) covariances of intercluster distances. ($\mathrm{Cov}[d_{jk}, d_{lm}]$ is equal to $\mathrm{Cov}[d_{lm}, d_{jk}]$.) The variances involved are directly obtainable from an equation equivalent to equation (2). The distance contributing to $\mathrm{Cov}(d_{jk}, d_{lm})$ is given by

$$d_{AB} - \frac{d_{(jl)} + d_{(km)}}{2}, \tag{8}$$

where $d_{(jl)}$ and $d_{(km)}$ are the *intracluster* distances between species j and l and species k and m, respectively. Therefore, we have

$$\mathrm{Cov}(d_{jk}, d_{lm}) = \frac{1 - i}{in}, \tag{9}$$

where

$$i = e^{-d_{AB}+(d_{(jl)}+d_{(km)})/2}. \tag{10}$$

Note that $d_{(jl)} = 0$ for $j = l$ and $d_{(km)} = 0$ for $k = m$.

Although the above theory is straightforward, the actual computation can be tedious, particularly when the number of species involved is large. However, the variances of branching points can be computed simultaneously with the estimation of UPGMA branch lengths. Such a computer program has been developed and may be obtained on request.

At this point, some readers might wonder how the above formulation is related to Chakraborty's (1977) study of the variance of a branch length. His study is different from ours on two points. First, he assumed a Poisson variance for d, so that it is smaller than our variance, as mentioned earlier. Second, he was interested in the variance of a branch length, rather than in the variance of a branching point. Since he assumed a Poisson variance for d, his expressions look simple. If we use the correct variance of d as given in equation (2), however, Chakraborty's method becomes more complicated than ours.

The variance given by equation (7) can be used for testing the statistical significance of the difference between two branching points in a tree, but some caution is necessary. When the two branching points to be tested belong to independent clusters (e.g., branching points 5 and 6 in fig. 1), the variance of the difference is simply the sum of the variances of the two branching points. However, when the two branching points are hierarchically related (e.g., 5 and 7 in fig. 1), there is a correlation between them. For example, the variance of the difference ($\delta = b_{AB} - b_{12}$) between points 7 and 5 in figure 1 is given by

$$V(\delta) = \frac{V(d_{AB}) + V(d_{12}) - 2\,\mathrm{Cov}(d_{AB}, d_{12})}{4}. \tag{11}$$

Therefore, if we use

$$V_U(\delta) = \frac{V(d_{AB}) + V(d_{12})}{4} \tag{12}$$

as the variance of δ, it will be an overestimate of the true variance, since $\mathrm{Cov}(d_{AB}, d_{12})$ is always positive. $\mathrm{Cov}(d_{AB}, d_{12})$ can easily be evaluated in the present case. That is,

$$\mathrm{Cov}(d_{AB}, d_{12}) = \mathrm{Cov}\left[\frac{d_{13} + d_{14} + d_{23} + d_{24}}{4}, d_{12}\right] = V(b_{12}) = \frac{V(d_{12})}{4}. \tag{13}$$

Therefore, we can use equation (11) for testing $b_{AB} - b_{12}$.

The above procedure can easily be extended to cases where both clusters A and B in figure 1 involve more than two species. A few examples are presented in the Appendix. When the number of species in cluster A is large, however, the computation can be quite tedious. In this case, it is convenient to note that the difference between $V(\delta)$ and $V_U(\delta)$ is usually small and that $V_U(\delta)$ can thus be used

for an approximate test. Since $V_U(\delta) \geq V(b_{AB} - b_{12})$, such a test will be a conservative one. In the Appendix, we also present a formula $[V_L(\delta)]$ for the lower bound of the variance of $b_{AB} - b_{12}$. This formula can be used for a liberal test if necessary. (See the numerical examples in the next section.)

In the above formulation, we assumed that the rate of amino acid replacement is the same for all amino acid sites. In practice, this assumption does not hold for most proteins (Fitch and Margoliash 1967*a*), and the actual number of amino acid replacements seems to follow the negative binomial rather than the Poisson distribution (Uzzell and Corbin 1971). Nevertheless, Nei and Chakraborty (1976) have shown that equation (1) holds approximately for most replacement patterns as long as d is <1. Of course, the variance of d will be larger than that given by equation (2), but the extent of underestimation of $V(d_{AB})$ given by equation (7) would not be serious as long as $d_{AB} < 1$. Furthermore, even if it is an underestimate, the variance of branching points will be useful in making a proper interpretation of estimated trees, as long as we understand it is an underestimate.

Nucleotide Substitution Data

In the case of nucleotide substitution data, the relationship between the expected number of nucleotide substitutions per site ($2\lambda t$) and the proportion (i) of identical nucleotides between two DNA sequences that are compared is given by

$$d = -\tfrac{3}{4} \log_e[1 - \tfrac{4}{3}(1 - i)] \tag{14}$$

(Jukes and Cantor 1969), and the variance of d is

$$V(d) = \frac{i(1 - i)}{n[1 - 4(1 - i)/3]^2} \tag{15}$$

(Kimura and Ohta 1972). The variances of branching points can be obtained in the same way as is that in the case of amino acid substitution. However, note that the formula corresponding to equation (10) is given by

$$i = \tfrac{1}{4} + \tfrac{3}{4}e^{-4/3[d_{AB}-(d_{(jl)}+d_{(km)})/2]}. \tag{16}$$

Equation (14) is based on the assumption that nucleotide substitution occurs at random among the four nucleotides A, T, C, and G. However, this equation is known to hold quite well, even under nonrandom nucleotide substitution, as long as d is smaller than 0.5 (Kimura 1980, 1981; Takahata and Kimura 1981; Gojobori et al. 1982). Therefore, the above method for computing the variances of branching points seems to apply as long as $d < 0.5$. In the case of nonrandom nucleotide substitution with $d \geq 0.5$, a more elaborate method for estimating d is necessary. There are several such methods, but most of them are model-dependent, and the computation of the variance of d is complicated. The only method in which the variance can be computed relatively easily is Tajima and Nei's (1984). In this method, d is estimated by

$$d = -c \log_e\left[1 - \frac{(1 - i)}{c}\right],$$

where c is a function of the frequencies of the four nucleotides and nucleotide pairs between the two DNA sequences compared. In Jukes and Cantor's formula, c = ¾. Since this c can be estimated from data by Tajima and Nei's method, one can use the same method of computing the variance of branching points as that used in the case of c = ¾, simply replacing ¾ in equations (14), (15), and (16) by the estimate of c that is obtained.

Restriction-Sites Data

The evolutionary change of nucleotide sequence can also be studied by examining restriction cleavage site differences among different species or genes. When restriction enzymes with r nucleotides in the recognition sequence are used, the maximum likelihood estimate of the number of nucleotide differences per site is given by

$$d = \frac{[-\log_e S]}{r}, \tag{17}$$

where $S = 2m_{XY}/(m_X + m_Y)$ (Nei and Li 1979; Kaplan and Risko 1981; Nei and Tajima 1983). Here, m_X and m_Y are the numbers of restriction sites for DNA sequences X and Y, respectively, and m_{XY} is the number of restriction sites shared by the two sequences. The variance of d is given by

$$V(d) = \frac{(2 - S)(1 - S)}{2r^2\hat{m}S}, \tag{18}$$

where $\hat{m} = (m_X + m_Y)/2$ (Nei and Tajima 1983). Equations (17) and (18) are quite accurate as long as $d \leq 0.25$. When $d > 0.25$, Nei and Tajima's (1983) equations (21) and (23) should be used. However, as d increases, the reliability of the estimate of d gradually declines.

Since we have the formulae for computing the mean and variance of d, the SE values of branching points can be estimated as in the case of amino acid replacement data. The equation corresponding to equation (10) is

$$S = e^{-r[d_{AB}-(d_{(ji)}+d_{(km)})/2]}. \tag{19}$$

When restriction enzymes with various values of r are used, the maximum likelihood estimate of d can still be obtained by the methods of Kaplan and Langley (1979), Gotoh et al. (1979), Kaplan and Risko (1981), and Nei and Tajima (1983). The simplest method among these is Nei and Tajima's (1983) iteration method. The variance of the estimate of d thus obtained is given by

$$V(d) = \frac{1}{\sum_{i=1}^{n} [1/V_i(d_i)]}, \tag{20}$$

where n is the number of types of enzymes used, and $V_i(d)$ is the value of equation (18) for the *i*th type of enzymes when

$$S = e^{-r_i d}$$

is used. Here, r_i is the value of r for the *i*th type of restriction enzymes. The expected value of S for the *i*th type of enzymes can be obtained by replacing r by r_i in equation (19), and the corresponding variance can be obtained by equation (20). (See the numerical example given later in the text.)

Electrophoretic Distance

In the construction of evolutionary trees for closely related species or populations, electrophoretic data are often used. Although many distance measures are available for electrophoretic data, Nei's (1972) measure seems to be most appropriate for this purpose, since it is linearly related to evolutionary time, at least theoretically. By using computer simulation, Nei et al. (1983) have also shown that this distance measure, in combination with UPGMA, is most efficient in recovering the true tree among the several alternative distance measures examined. Farris (1981) criticized this measure for not permitting evolutionary path-length interpretation. However, this criticism is apparently based on his failure to distinguish between expected and observed distances (Nei et al. 1983; Felsenstein 1984).

Let x_i and y_i be the frequency of the *i*th allele at a locus in populations X and Y, respectively. The gene identities within and between populations are then computed by $j_X = \sum x_i^2$, $j_Y = \sum y_i^2$, and $j_{XY} = \sum x_i y_i$, where $\sum$ stands for the summation for all alleles at the locus. The average gene identities over all loci are simply the averages of these quantities, that is,

$$J_X = \frac{\sum_{k=1}^{n} j_{Xk}}{n}, \qquad J_Y = \frac{\sum_{k=1}^{n} j_{Yk}}{n}, \qquad \text{and} \qquad J_{XY} = \frac{\sum_{k=1}^{n} j_{XYk}}{n},$$

where k refers to the *k*th locus and n is the number of loci examined. Nei's (1972) standard genetic distance is then defined as $D = -\log_e I$, where $I = J_{XY}/\sqrt{J_X J_Y}$. If the time after divergence between populations X and Y is t years and the two populations are in equilibrium with respect to the effects of mutation, selection, and genetic drift, the expectation of I is given by $I = e^{-2\alpha t}$, where α is the rate of gene substitution per locus per year. Therefore, $D = 2\alpha t$.

The D value between a pair of species or populations and its variance can be estimated by the methods of Nei and Roychoudhury (1974) and Nei (1978), although these methods are more complicated than those for amino acid replacement or nucleotide substitution. However, it is not always easy to compute the variance of branching points, since the allele frequencies of ancestral populations such as those for branching points 5 and 6 in figure 1 are not known. The only cases in which this is not a problem are those in which (1) all populations examined have a relatively low average heterozygosity and are fixed for different alleles at some loci or (2) different strains in asexual haploid organisms such as *Escherichia coli* (e.g., Selander and Levin 1980) are studied. In these cases the variance of D between a pair of populations is given by

$$V(D) = \frac{1 - I}{In} \tag{21}$$

(Nei 1971), and all other procedures for obtaining the variance of branching points are the same as those for amino acid sequence data.

In practice, the applicability of equation (21) depends on the value of I relative to average heterozygosity (H). When I is <0.9 for all pairs of populations, and H is <0.2, equation (21) seems to give a quite reliable estimate, as long as many loci are examined. This is because in this case the genetic identity for a single locus ($I_j = j_{XY}/\sqrt{j_X j_Y}$) shows a distribution close to the binomial distribution, with I_j taking a value of 0 or 1 for a majority of loci. When I is >0.9 for most population pairs, however, equation (21) may give a serious overestimate of the true variance.

Although it is difficult to develop a general method for computing the exact variance of a branching point, it is possible to get an upper bound or a maximum estimate of the variance. This maximum estimate is the average of the variances of all pairwise distances used for computing a branching point. For example, branching point 7 in figure 1 is given by half the average (D_{AB}) of D_{13}, D_{14}, D_{23}, and D_{24} in the case of electrophoretic data. The maximum variance of D_{AB} is then given by

$$V(D_{AB}) = \frac{V(D_{13}) + V(D_{14}) + V(D_{23}) + V(D_{24})}{4}. \quad (22)$$

This variance is obviously larger than the true variance given by an equation equivalent to equation (4) or equation (7). (Note that the divisor of the right hand side of equation (22) is rs = 4, whereas that of equation (4) is $(rs)^2 = 16$.) However, as long as we understand that equation (22) gives an upper bound of the true variance, we can use it for getting a rough idea of the reliability of branching points. The accuracy of $V(b_{AB}) = V(D_{AB})/4$ as an estimate of the variance of a branching point depends on tree structure. If most intracluster branches are short, the accuracy is quite high. However, if there are many long branches within each cluster, the variance may be seriously overestimated. In general, this method should be used only when equation (22) gives a smaller value than does equation (21). This is expected to occur when the I values are large, for example, >0.9.

Evolutionary Tree For the Human and Ape Species

In the past 10 yrs the genetic relationship of the human and ape species has been studied intensively by using various types of molecular data. In the following, we apply our methods to four different types of data.

Amino Acid Sequence Data

Although amino acid sequencing of proteins was started more than 20 yr ago, the sequence data for the human and ape species are still limited. The only data that can be used for our purpose are those for hemoglobins α and β, myoglobin, and fibrinopeptides A and B for the human, chimpanzee, gorilla, and orangutan species (Dayhoff 1972, 1973, 1978; Goodman et al. 1983). There are also data for two partial sequences (13 amino acids each) of the duplicate hemoglobin γ chains for these organisms (Huisman et al. 1973). The rate of amino acid substitution for fibrinopeptides is known to be significantly higher than that for hemoglobins. However, since the sequence differences among these four species are very small, we can pool the sequences together to compute the proportion of different amino acids (p = 1 − i) for each pair of these species, using a total of 496 amino acids for comparison. From these proportions, the number of amino acid replacements per site (d) and its SE can be computed by equations (1) and (2). The results obtained are presented in table 1. In the present case, $d = -\log_e(1 - p)$ is virtually identical

Table 1
Proportions of Different Amino Acids and Estimates of Amino Acid Replacements for Four Primate Species

	Human	Chimpanzee	Gorilla	Orangutan
Human		0.0020	0.0060	0.0202
Chimpanzee ...	0.0020 ± 0.0020		0.0081	0.0222
Gorilla	0.0061 ± 0.0035	0.0081 ± 0.0040		0.0222
Orangutan	0.0204 ± 0.0064	0.0224 ± 0.0068	0.0224 ± 0.0068	

NOTE.—Numbers above the diagonal are the proportion ($p = 1 - i$) of different amino acids for hemoglobins α, β, $^{A}\gamma$, $^{G}\gamma$; myoglobin; and fibrinopeptides A and B for the species pairings indicated. Numbers below the diagonal are numbers of amino acid replacements per site (±SE) for the species pairings indicated. The number of amino acids used for hemoglobins α, β, $^{A}\gamma$, and $^{G}\gamma$; myoglobin; and fibrinopeptides A and B are 141, 146, 13, 13, 153, and 30, respectively, for a total of 496. The data for hemoglobin $^{A}\gamma$ and $^{G}\gamma$ refer to a region (13 amino acids) sequenced by Huisman et al. (1973). The amino acid sequence for the remaining region is not known for all four of the species examined. The data on sequence differences for the orangutan hemoglobin-α and -β chains are those of Goodman et al. (1983). All other data were obtained from Dayhoff (1972, 1973, 1978).

with p, since p is very small. The evolutionary tree reconstructed from the d values by using UPGMA is given in figure 2. The branching point (a) between the human (H) and chimpanzee (C) species is $b_{HC} = 0.0020/2 = 0.0010$ (from table 1), and the SE is $0.0020/2 = 0.0010$, which is identical with b_{HC}, since b_{HC} is small. Similarly, the branching point (b) between the gorilla (G) and the human-chimpanzee lines is $b_{(HC)G} = (0.0061 + 0.0081)/4 = 0.0036$.

The variance $[V(b_{(HC)G})]$ of $b_{(HC)G}$ is $V(d_{(HC)G})/4$, where

$$V(d_{(HC)G}) = \frac{V(d_{HG}) + V(d_{CG}) + 2\,\mathrm{Cov}(d_{HG}, d_{CG})}{4}, \qquad (23)$$

from equation (7). We have $V(d_{HG}) = (0.0035)^2$ and $V(d_{CG}) = (0.0040)^2$ (from the SE values in table 1). To compute $\mathrm{Cov}(d_{HG}, d_{CG})$, we must know the expected distance between the gorilla species (G) and the branching point (a) between the human and chimpanzee species. It is given by $d = 2b_{(HC)G} - b_{HC} = 0.0062$. Therefore, the corresponding i value is $e^{-0.0062} = 0.9938$, from equation (10). Thus, $\mathrm{Cov}(d_{HG}, d_{CG}) = V(d)$ becomes 1.258×10^{-5}, from equation (9). Hence, the SE of $b_{(HC)G} = [V(d_{(HC)G})/4]^{1/2} = 0.0018$. The SE of branching point c can be obtained in

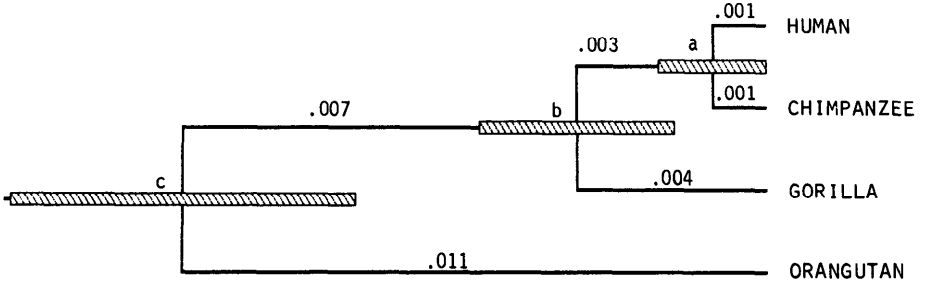

FIG. 2.—Evolutionary tree for four hominoid species, which was reconstructed from amino acid sequence data. The number given for each branch represents the branch length or the number of amino acid replacements per site. The hatched box represents 1 SE on each side of the mean branch length.

Table 2
Numbers of Nucleotide Substitutions per Site (d) for Five Primate Species

	Human	Chimpanzee	Gorilla	Orangutan
Chimpanzee ...	0.0939 ± 0.0107			
Gorilla	0.1106 ± 0.0118	0.1145 ± 0.0120		
Orangutan	0.1797 ± 0.0156	0.1940 ± 0.0163	0.1882 ± 0.0160	
Gibbon	0.2072 ± 0.0170	0.2175 ± 0.0175	0.2175 ± 0.0175	0.2160 ± 0.0174

NOTE.—Data from Brown et al. 1982. In the computation of d, 895 nucleotides were used because there was one nucleotide deletion in the segment of the orangutan mtDNA.

the same way, and it becomes 0.0032. These results are presented graphically in figure 2. The difference between branching points a and b can be tested by using the normal deviate (t). In the present case, t becomes $(0.0036 - 0.0010)/[0.0018^2 + 0.0010^2/2]^{1/2} = 1.34$, by using equation (A3), so that the difference is not statistically significant. If we had used the approximate variance, as calculable by equation (12), t would have been 1.26, which is not much different from the correct value. On the other hand, the difference between branching points b and c is significant at the 5% level, since we have $t = 0.0074/0.0035 = 2.10$ by using equation (A5). The same conclusion is obtained by using equation (12), which gives $t = 2.00$. A similar t value ($t = 2.15$) is also obtained by using the lower bound variance, $V_L(\delta)$, in equation (A7).

Nucleotide Sequence Data

Brown et al. (1982) sequenced a segment (896 nucleotides) of mitochondrial DNA (mtDNA) for the human, chimpanzee, gorilla, orangutan, and gibbon species. From these sequence data, we can estimate d and its SE by using equations (14) and (15), respectively. The results obtained are presented in table 2, and the UPGMA tree obtained from these distance data is given in figure 3. If we exclude the gibbon, the topology of this tree is identical with that of the tree derived from amino acid sequence data. The branching point (a) between the human and chimpanzee species is $b_{HC} = 0.0939/2 = 0.0470$ (from table 2). The SE of this b_{HC}

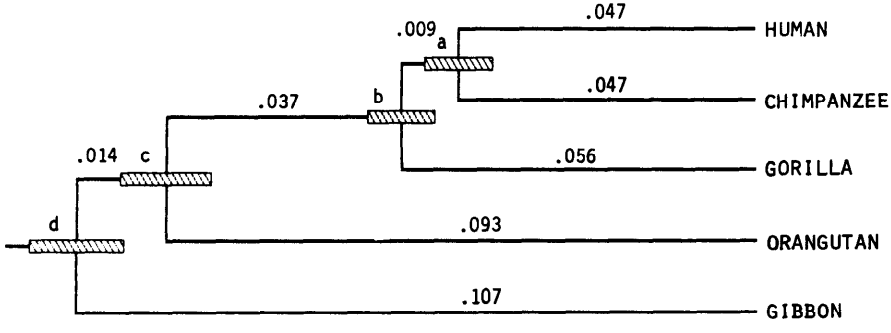

FIG. 3.—Evolutionary tree for five hominoid species, which was reconstructed from nucleotide-sequence data on a segment of mitochondrial DNA. The number given for each branch represents the branch length or the number of nucleotide substitutions per site. The hatched box represents 1 SE on each side of the mean branch length.

$= 0.0107/2 = 0.0054$. Similarly, the branching point (b) between the gorilla and the human-chimpanzee groups is $b_{(HC)G} = (0.1106 + 0.1145)/4 = 0.0563$.

The variance of $b_{(HC)G}$ can be obtained by using equation (23). In the present case, $V(d_{HG}) = (0.0118)^2$ and $V(d_{CG}) = (0.0120)^2$ (from table 2). On the other hand, the expected distance between the gorilla and branching point a is given by $d = 2b_{(HC)G} - b_{HC} = 0.0656$, so that the corresponding i value is $\frac{1}{4} + (3/4) \times \exp[-(4/3) \times 0.0656] = 0.9372$, from equation (16). Thus, $\mathrm{Cov}(d_{HG}, d_{CG}) = V(d)$ becomes 7.833×10^{-5}, from equation (15). Therefore, the SE of $b_{(HC)G} = 0.0052$. Similarly, the SE values of branching points c and d in figure 3 become 0.0071 and 0.0074, respectively. The differences between branching points a and b and those between branching points c and d in figure 3 are not statistically significant. However, the difference between branching points b and c is significant, t being 4.5. In this case, equation (12) gives $t = 4.2$, whereas equation (A7) gives $t = 4.6$.

Restriction-Sites Data

A phylogenetic tree for these five primate species can also be constructed from Ferris et al.'s (1981) restriction-sites data. Estimates of the numbers of nucleotide substitutions are obtained from the data in table 3 by using Nei and Tajima's (1983) equations (25) and (28). This table includes data on the number of restriction sites and the number of shared restriction sites for eighteen six-base enzymes and one four-base enzyme. Ferris et al. (1981) used two six-base enzymes with four recognition sequences (*Ava*I and *Hin*cII). Two of the four recognition sequences of each of these two enzymes were identical with those of two other six-base enzymes. For example, *Hin*cII recognizes the four sequences GTTGAC, GTCAAC, GTTAAC, and GTCGAC, but the sequences GTTAAC and GTCGAC are also recognized by *Hpa*I and *Sal*I, respectively. In the case of *Ava*I and *Hin*cII, we have therefore considered only those restriction sites that were not recognized by the other enzymes. We note that two mutational changes are required to transform one of the two remaining recognition sequences (e.g., GTTGAC and GTCAAC in *Hin*cII) to the other, so that the shared restriction sites for a pair of species (unique to these enzymes) must almost always have an identical sequence (e.g., GTTGAC or GTCAAC in *Hin*cII). Therefore, each of these two enzymes can be regarded as consisting of two different enzymes with $r = 6$. This makes the total number of enzymes with $r = 6$ equal to 20.

Table 3
Numbers of Restriction Sites (m) and Shared Restriction Sites (m_{XY}) for the Mitochondrial DNAs from Five Primate Species

	Human	Chimpanzee	Gorilla	Orangutan	Gibbon
Human	42 (6)				
Chimpanzee ...	19 (6)	42 (6)			
Gorilla	22 (4)	25 (6)	42 (6)		
Orangutan	20 (4)	15 (4)	16 (4)	36 (10)	
Gibbon	17 (4)	15 (4)	18 (4)	14 (5)	47 (7)

NOTE.—The figures in front of parentheses are the sums of m or m_{XY} for 18 6-base enzymes, whereas those in parentheses are the values of m or m_{XY} for one 4-base enzyme used. The values on and off the diagonal are m's and m_{XY}'s, respectively. Data from Ferris et al. (1981) were used.

Table 4
Numbers of Nucleotide Substitutions per Site for Five Primate Species

	Human	Chimpanzee	Gorilla	Orangutan
Chimpanzee ...	0.1142 ± 0.0209			
Gorilla	0.1071 ± 0.0198	0.0880 ± 0.0171		
Orangutan	0.1194 ± 0.0219	0.1613 ± 0.0283	0.1521 ± 0.0268	
Gibbon	0.1556 ± 0.0262	0.1735 ± 0.0289	0.1474 ± 0.0250	0.1729 ± 0.0291

NOTE.—Data from Ferris et al. (1981).

The estimates of the number of nucleotide substitutions obtained are presented in table 4. The estimates of substitutions between the human and chimpanzee species and between the human and gorilla species are virtually the same as those derived from the nucleotide sequence data, but the estimates for the other pairs of organisms are smaller than those derived from the sequence data, the difference between the two sets of data increasing with increasing d. This difference could be due to either inaccuracy of the estimates obtained from restriction-sites data or the difference in the rate of nucleotide substitution between the sequenced region and the entire region of mtDNA. We note that the SE values of d's derived from restriction-sites data are considerably larger than those derived from sequence data.

Figure 4 shows the UPGMA tree reconstructed from the d values in table 4. This tree has a topology different from that in figure 3. That is, among the human, chimpanzee, and gorilla species, the latter two cluster first in this tree, whereas in the tree in figure 3, the human and chimpanzee species make the first cluster. All other parts of the topology are the same for the two trees.

The branching point (a) between the chimpanzee and the gorilla and its SE can be obtained directly from table 4. That is, $b_{CG} = 0.0440 \pm 0.0085$. The branching point (b) between the human and the chimpanzee-gorilla group is $b_{H(CG)} = (0.1142 + 0.1071)/4 = 0.0553$. The variance of this branching point can be

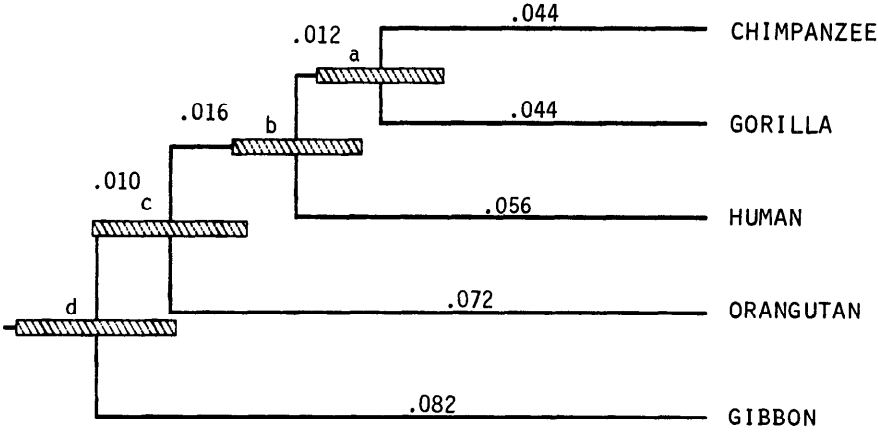

FIG. 4.—Evolutionary tree for five hominoid species, which was reconstructed from restriction sites data for mitochondrial DNA. The number given for each branch represents the branch length (the number of nucleotide substitutions per site). The hatched box represents 1 SE on each side of the mean branch length.

obtained by an equation similar to equation (23). From table 4, $V(d_{HC}) = (0.0209)^2$ and $V(d_{HG}) = (0.0199)^2$. We also have $Cov(d_{HC}, d_{HG}) = V(d)$, where $d = 2b_{H(CG)} - b_{CG} = 0.0667$. The value of V(d) can be computed by $[\sum_r V_r(d)^{-1}]^{-1}$, where $V_r(d)$ is the variance of d for r-base enzymes given by $(2 - S)(1 - S)/[2r^2\bar{m}S]$ (Nei and Tajima 1983). Here, $S = e^{-rd}$, and $\bar{m}$ is the mean number of restriction sites. Nei and Tajima (1983) defined $\bar{m}$ as the mean for the two species to be compared, but in the present case, $\bar{m}$ should be defined as the mean for all species being investigated, since we are computing the expected variance of d for the entire set of data. This mean can be obtained from the data in table 3, and it becomes 42 for the six-base enzymes and 7 for the four-base enzyme. Therefore, we have $V_6(d) = 0.00022$, and $V_4(d) = 0.0017$. Thus, $V(d) = 0.00019$, and $V(d_{H(CG)}) = 0.00030$. Hence, the SE of $b_{H(CG)} = 0.0087$. The SE values of all other branching points can be obtained in the same way. The results obtained are presented graphically in figure 4. It can be seen that the SE values of the branching points are considerably larger than those of the tree obtained from nucleotide sequencing (fig. 3). The difference between branching points a and b is again statistically nonsignificant. In the present case, even the difference between branching points b and c is not significant.

Electrophoretic Data

The phylogenetic relationship of the human and ape species was also studied by Bruce and Ayala (1979) by using electrophoresis. They examined the electrophoretic variation of 23 protein loci for man, two species of chimpanzee (*Pan troglodytes* and *P. paniscus*), the gorilla, two subspecies of orangutan (*Pongo pygmaeus abelii* from Sumatra and *P. p. pygmaeus* from Borneo), two species of gibbon (*Hylobates lar* and *H. concolor*), and the siamang. Bruce and Ayala did not use their own data on the human species. Instead, they used data obtained by other workers. It is therefore difficult to reconstruct their computation of genetic distances between the human and other primate species. In the following, we have eliminated from our analysis all data concerning the human species. Data concerning the siamang have also been eliminated, since they were derived from only one individual. Although they studied 23 protein loci, Bruce and Ayala found a few species in which gene frequency data could not be obtained for a few loci. We therefore eliminated the Borneo orangutan and used the 20 loci that were shared by the remaining six species. Estimates of genetic distances and their SE values were computed by using Nei's (1978) method. The results obtained are presented in table 5. In this table,

Table 5
Genetic Distances and Average Heterozygosities for Six Species of Apes

	Chimpanzee	Pygmy Chimpanzee	Gorilla	Orangutan	Lar Gibbon	Concolor Gibbon
Chimpanzee	0.011	0.107	0.475	0.269	0.668	0.806
Pygmy chimpanzee	0.077 (0.075)	0	0.419	0.140	0.661	0.799
Gorilla	0.176 (0.174)	0.164 (0.161)	0.054	0.546	0.510	0.666
Orangutan	0.124 (0.124)	0.084 (0.087)	0.192 (0.191)	0.055	0.634	0.633
Lar gibbon	0.220 (0.218)	0.219 (0.216)	0.184 (0.182)	0.210 (0.210)	0.052	0.133
Concolor gibbon	0.255 (0.249)	0.254 (0.247)	0.223 (0.217)	0.215 (0.210)	0.082 (0.084)	0

NOTE.—Figures above the diagonal are genetic distances, those on the diagonal are average heterozygosities, and those below the diagonal are the SE's of genetic distances. The SE's of genetic distances in parentheses are those obtained by equation (21). Data from Bruce and Ayala (1979) are used.

the SE values obtained by equation (21) are given in parentheses. It is clear that these SE values are very close to those obtained by Nei's more accurate method. This is because I is <0.9 for all pairs of species, and the average heterozygosity is low for all species (table 5). In computing the SE values of branching points, we can therefore use the same method as that for amino acid sequence data.

The UPGMA tree reconstructed from the genetic distances listed in table 5 is given in figure 5. The topology of this tree is different from that of all previous trees, the orangutan being closer to the chimpanzee than to the gorilla. In the present case, the SE values of branching points a, e, and b become 0.037, 0.042, and 0.050, respectively. The SE values of the other branching points are shown graphically in figure 5. It is seen that the phylogenetic tree constructed from electrophoretic data is even less reliable than that obtained from restriction sites data. Although the orangutan clusters with the chimpanzee before it joins with the gorilla, the SE of the branching point (c) between the gorilla and the chimpanzee-orangutan group is so large that the distance between b and c is not statistically significant. The branching point between the gibbon and the other apes also has a large SE.

This low reliability of electrophoretic data is partly due to the small number of loci used. In a computer simulation, Nei et al. (1983) have shown that when the number of loci used is <30, the topology of a reconstructed tree is subject to a large stochastic error. The accuracy of a reconstructed tree also depends on the detectability of protein differences by electrophoresis. The higher the detectability, the higher the reliability. It should be noted that in Bruce and Ayala's experiment, this detectability was not particularly high. Previously, King and Wilson (1975) had studied the genetic distance between the human and chimpanzee species and obtained D = 0.62, which is nearly two times higher than the estimate (0.39) obtained by Bruce and Ayala.

Discussion

In the present paper, we have assumed that the *expected* number of gene substitutions is proportional to evolutionary time, although the actual number may

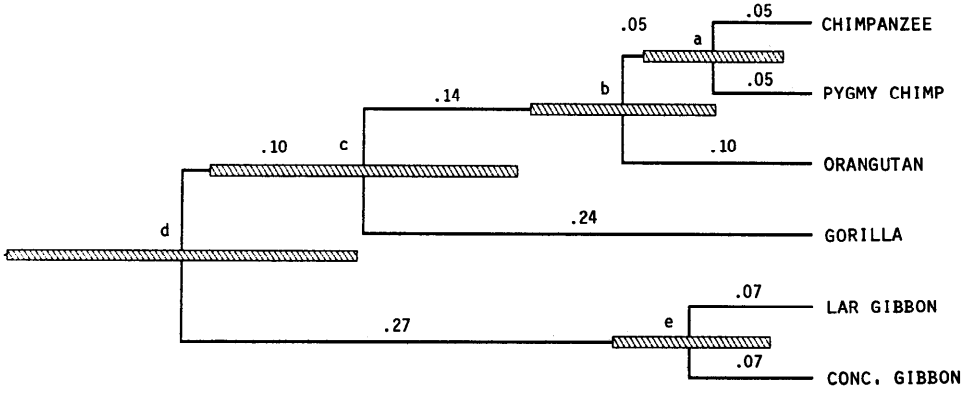

FIG. 5.—Evolutionary tree for six hominoid species, which was reconstructed from electrophoretic data. The number given for each branch represents the branch length (genetic distance). The hatched box represents 1 SE on each side of the mean branch length.

deviate from the expected number because of stochastic errors. This assumption is necessary for estimating the topology and branch lengths of a UPGMA tree. In practice, of course, it would not always hold, and if it does not, a reconstructed tree may deviate substantially from the true tree. Deviations of reconstructed trees from the true trees also occur because of stochastic errors in the process of gene substitution, unless the number of amino acids or nucleotides examined is very large (Tateno et al. 1982; Nei et al. 1983). Therefore, caution should be exercised in the interpretation of a tree reconstructed from molecular data.

Some tree-making methods, such as Fitch and Margoliash's (1967*b*) and Farris's (1972) parsimony methods, are designed to take care of unequal rates of gene substitution in different branches. In these methods, however, even stochastic errors are considered to be a reflection of unequal rates of gene substitutions, and this introduces another source of error into the process of tree making. Indeed, in the presence of large stochastic errors, Fitch and Margoliash's and Farris's methods are not necessarily better than UPGMA in recovering the true tree, even if the substitution rate varies with branch to some extent (Tateno et al. 1982). Furthermore, in parsimony methods, it is not easy to compute the SE values of branching points since there is no way to compute the expectation and variance of gene substitutions.

We have computed the SE values of branching points for four different types of molecular data and found that the SE values relative to the means or the coefficients of variation are smallest for nucleotide sequence data and largest for electrophoretic data. The smaller coefficients of variation for nucleotide sequence data than for amino acid sequences are mainly due to two factors: (1) the number of nucleotides examined is larger than the number of amino acids examined and (2) the d values for nucleotide sequence data are larger than those for amino acid replacement data. The coefficients of variation for nucleotide sequence data are also substantially smaller than those for restriction sites data. This is largely because the number of nucleotides assayed is larger in the former than in the latter. In general, the average number of nucleotides assayed by restriction enzymes having r nucleotides per DNA sequence is given by $r\bar{m}$. In the present case, $\bar{m} = 42$ for $r = 6$, and $\bar{m} = 7$ for $r = 4$. Therefore, it becomes $6 \times 42 + 4 \times 7 = 280$. This is approximately one-third of the number of nucleotides assayed (895) in nucleotide sequencing.

As is clear from figure 5, electrophoretic data show the largest coefficients of variation of d (=D) among the four types of data used here. This is partly because the number of loci examined (20) is small in the present case. However, even if one uses 80 loci, the coefficients of variation would be only approximately half of those given in figure 5 and still considerably higher than those for nucleotide sequences of mtDNA. Therefore, electrophoretic data seem to be less informative than mtDNA data, unless a very large number of loci are examined. However, electrophoretic data have one advantage over mtDNA data, namely, that they give an average evolutionary change of genes for many loci, whereas mtDNA represents a small piece of DNA, which is maternally inherited without gene recombination. The evolutionary tree reconstructed from a single piece of DNA without recombination is subject to larger stochastic errors than that reconstructed from many independently evolving genes.

In this paper, we reconstructed the evolutionary tree of the human and several ape species using four different types of data. The topology of the tree reconstructed is not the same for all types of data. The most reliable tree among the four that were reconstructed seems to be that obtained from nucleotide sequence data. In this

tree, the gibbon and the orangutan separate from the human, chimpanzee, and gorilla species significantly earlier than the latter three species diverge. Among the latter three species, the gorilla diverges from the human earlier than the chimpanzee does, but the difference between the two branching points is not statistically significant. Therefore, we cannot rule out the possibility that the three species diverged at nearly the same time (Sarich and Wilson 1967). The results derived from protein data are essentially the same as those derived from nucleotide sequence data. It should also be noted that this topology is in agreement with that of the trees reconstructed by both chromosomal studies (Yunis and Prakash 1982) and DNA hybridization (Sibley and Ahlquist 1984). However, the topology of the parsimony tree obtained by Brown et al. (1982) is different from ours, even though the same set of data was used. In a statistical analysis of the parsimony tree reconstructed by Ferris et al. (1981), Templeton (1983) concluded that the topology in figure 4 is significantly better than that in figure 3. However, his conclusion is not justified, since the parsimony method he used introduces many statistical biases when it is applied to restriction sites data (Nei and Tajima 1984).

Acknowledgements

This study was supported by research grants from the National Institutes of Health and the National Science Foundation.

APPENDIX

Since it is not easy to develop a general algorithm to compute the variance of the difference between two branching points that are hierarchically related, we herewith illustrate the computation in a consideration of three typical cases. In general, we consider the case in which the two species clusters A and B join at branching point b_{AB} and, within cluster A, the two subclusters A1 and A2 join at branching point b_{A1A2}. We then evaluate the variance of the difference between the two branching points ($\delta = b_{AB} - b_{A1A2}$). For mathematical convenience, we consider the variance $[V(\delta_d)]$ of $\delta_d \equiv 2\delta = d_{AB} - d_{A1A2}$. Obviously, $V(\delta) = V(\delta_d)/4$. In general,

$$V(\delta_d) = V(d_{AB}) + V(d_{A1A2}) - 2\,\mathrm{Cov}(d_{AB}, d_{A1A2}). \tag{A1}$$

We already know how to compute $V(d_{AB})$ and $V(d_{A1A2})$. So, we consider the computation of $\mathrm{Cov}(d_{AB}, d_{A1A2})$ only.

Case 1: A1 and A2 each have one species, and B has s species. When s = 2, this is identical with the case represented in figure 1. In this case, we have

$$\mathrm{Cov}(d_{AB}, d_{A1A2}) = \mathrm{Cov}\left[\sum_{i=1}^{s} \frac{d_{iA1} + d_{iA2}}{2s}, d_{A1A2}\right] = s\,\frac{V(b_{\alpha A1}) + V(b_{\alpha A2})}{2s} = \frac{V(d_{A1A2})}{4}, \tag{A2}$$

where α is the branching point between A1 and A2, and $b_{\alpha A1}$ and $b_{\alpha A2}$ refer to the branch lengths between α and A1 and between α and A2, respectively. Therefore,

$$V(\delta_d) = V(d_{AB}) + \frac{V(d_{A1A2})}{2}. \tag{A3}$$

Case 2: A1, A2, and B contain two, one, and s species, respectively.

$$\begin{aligned}\mathrm{Cov}(d_{AB}, d_{A1A2}) &= \mathrm{Cov}[\sum_{i=1}^{s} (d_{iA1(1)} + d_{iA1(2)} + d_{iA2})/(3s), (d_{A1(1)A2} + d_{A1(2)A2})/2] \\ &= s[V(b_{\alpha A1(1)}) + V(b_{\alpha\beta}) + V(b_{\alpha A2}) + V(b_{\alpha\beta}) + V(b_{\alpha A1(2)}) + V(b_{\alpha A2})]/(6s) \\ &= [V(d_{A1A2}) + 2V(b_{\alpha\beta})]/6, \end{aligned} \tag{A4}$$

where α is the branching point between A1 and A2, and β is that between the two species in A1, that is, A1(1) and A1(2). Hence, $V(\delta_d)$ is given by

$$V(\delta_d) = V(d_{AB}) + \frac{2[V(d_{A1A2}) - V(b_{\alpha\beta})]}{3}. \tag{A5}$$

$V(b_{\alpha\beta})$ can be obtained by equation (10) or its equivalent formula.

Case 3: A1 and A2 both contain two species, and B has s species.

$$\begin{aligned}\mathrm{Cov}(d_{AB}, d_{A1A2}) = \mathrm{Cov}[\sum_{i=1}^{s} (d_{iA1(1)} + d_{iA1(2)} + d_{iA2(1)} + d_{iA2(2)})/(4s), (d_{A1(1)A2(1)} \\ + d_{A1(1)A2(2)} + d_{A1(2)A2(1)} + d_{A1(2)A2(2)})/4] = \frac{V(d_{A1A2})}{8} + \frac{V(b_{\alpha\beta}) + V(b_{\alpha\gamma})}{4}, \end{aligned} \tag{A6}$$

where γ is the branching point between two species in A2, that is, A2(1) and A2(2). Therefore, we can compute $V(\delta_d)$ and $V(\delta)$.

The above procedure can be extended to any type of branching pattern, but it is quite complicated if many species are involved. In this case, one may use the approximate formula $\mathrm{Cov}(d_{AB}, d_{A1A2}) = V(d_{A1A2})/4$. This is exact for case 1 but gives a slight overestimate of the true covariance for the other cases. It can be derived by assuming that the variances of $b_{\alpha\beta}$, $b_{\alpha\gamma}$, etc., are equal to the variance of $b_{\alpha A1}$ or $b_{\alpha A2}$. This provides a lower bound of the variance of δ, that is,

$$V_L(\delta) = \frac{[V(d_{AB}) + V(d_{A1A2})/2]}{4}. \tag{A7}$$

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.

BRUCE, E. J., and F. J. AYALA. 1979. Phylogenetic relationships between man and the apes: electrophoretic evidence. Evolution **33**:1040–1056.

CHAKRABORTY, R. 1977. Estimation of time of divergence from phylogenetic studies. Can J. Genet. Cytol. **19**:217–223.

DAYHOFF, M. O. 1972. Atlas of protein sequence and structure. Vol. **5**. National Biomedical Research Foundation, Silver Spring, Md.

———. 1973. Atlas of protein sequence and structure. Vol. **5**. Suppl. 1. National Biomedical Research Foundation, Silver Spring, Md.

———. 1978. Atlas of protein sequence and structure. Vol. **5**. Suppl. 3. National Biomedical Research Foundation, Silver Spring, Md.

FARRIS, J. S. 1972. Estimating phylogenetic trees from distance matrices. Am. Nat. **106**:645–668.

———. 1981. Distance data in phylogenetic analysis. Pp. 3–23 *in* V. FUNK and D. BROOK, eds. Advances in cladistics. New York Botanical Garden, Bronx, New York.

FELSENSTEIN, J. 1984. Distance methods for inferring phylogenies: a justification. Evolution **38**:16–24.

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.

FITCH, W. M., and E. MARGOLIASH. 1967*a*. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochem. Genet. **1**:65–71.

———. 1967*b*. Construction of phylogenetic trees. Science **155**:279–284.

GOJOBORI, T., K. ISHII, and M. NEI. 1982. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J. Mol. Evol. **18**:414–423.

GOODMAN, M., G. BRAUNITZER, A. STANGL, and B. SCHRANK. 1983. Evidence on human origins from haemoglobins of African apes. Nature **303**:546–548.

GOTOH, O., J.-I. HAYASHI, H. YONEKAWA, and Y. TAGASHIRA. 1979. An improved method for estimating sequence divergence between related DNAs from changes in restriction endonuclease cleavage sites. J. Mol. Evol. **14**:301–310.

HUISMAN, T. H. J., W. A. SCHROEDER, M. E. KEELING, N. GENGOZIAN, A. MILLER, A. R. BRODIE, J. R. SHELTON, J. B. SHELTON, and G. APELL. 1973. Search for nonallelic structural genes for γ-chains of fetal hemoglobin in some primates. Biochem. Genet. **10**:309–318.

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

KAPLAN, N., and C. H. LANGLEY. 1979. A new estimate of sequence divergence of mitochondrial DNA using restriction endonuclease mapping. J. Mol. Evol. **13**:295–304.

KAPLAN, N., and K. RISKO. 1981. An improved method for estimating sequence divergence of DNA using restriction endonuclease mappings. J. Mol. Evol. **17**:156–162.

KIMURA, M. 1969. The rate of molecular evolution considered from the standpoint of population genetics. Proc. Natl. Acad. Sci. USA **63**:1181–1188.

———. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

———. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78**:454–458.

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

KING, M. C., and A. C. WILSON. 1975. Evolution at two levels in humans and chimpanzees. Science **188**:107–116.

KUMAZAKI, T., H. HORI, and S. OSAWA. 1983. Phylogeny of protozoa deduced from 5S rRNA sequences. J. Mol. Evol. **19**:411–419.

NEI, M. 1971. Interspecific gene differences and evolutionary time estimated from electrophoretic data on protein identity. Am. Nat. **105**:385–398.

———. 1972. Genetic distance between populations. Am. Nat. **106**:283–292.

———. 1975. Molecular population genetics and evolution. North-Holland, New York.

———. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics **89**:583–590.

NEI, M., and R. CHAKRABORTY. 1976. Empirical relationship between the number of nucleotide substitutions and interspecific identity of amino acid sequences in some proteins. J. Mol. Evol. 7:313–323.

NEI, M., and W.-H. LI. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76**:5269–5273.

NEI, M., and A. K. ROYCHOUDHURY. 1974. Sampling variances of heterozygosity and genetic distance. Genetics **76**:379–390.

NEI, M., and F. TAJIMA. 1983. Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics **105**:207–217.

———. 1984. Evolutionary change of restriction cleavage sites and phylogeny inference for human and apes. Unpublished manuscript.

NEI, M., F. TAJIMA, and Y. TATENO. 1983. Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J. Mol. Evol. **19**:153–170.

PEACOCK, D., and D. BOULTER. 1975. Use of amino acid sequence data in phylogeny and evaluation of methods using computer simulation. J. Mol. Biol. **95**:513–527.

SARICH, V. M., and A. C. WILSON. 1967. Immunological time scale for homonoid evolution. Science **158**:1200–1204.

SELANDER, R. K., and B. R. LEVIN. 1980. Genetic diversity and structure in *Escherichia coli* populations. Science **210**:545–547.

SIBLEY, C. G., and J. E. AHLQUIST. 1984. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J. Mol. Evol. **20**:2–16.

SNEATH, P. H. A., and R. R. SOKAL. 1973. Numerical taxonomy. Freeman, San Francisco.

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

TAKAHATA, N., and M. KIMURA. 1981. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics **98**:641–657.

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.

TEMPLETON, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and apes. Evolution **37**:221–244.

UZZELL, T., and K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. Science **172**:1089–1096.

YUNIS, J. J., and O. PRAKASH, 1982. The origin of man: a chromosomal pictorial legacy. Science **215**:1525–1530.

ROBERT K. SELANDER, reviewing editor

Received March 29, 1984; revision received June 26, 1984.