Variances of the Average Numbers of Nucleotide Substitutions Within and Between Populations[1]

Masatoshi Nei and Li Jin

Center for Demographic and Population Genetics, Graduate School of Biomedical Sciences, The University of Texas Health Science Center at Houston

Statistical methods for computing the variances of nucleotide diversity within populations and of nucleotide divergence between populations are developed. Both variances are computed by finding the phylogenetic relationships of the DNA sequences studied through the unweighted pair-group method or some other tree-making method. The methods developed are applicable to both DNA sequence and restriction-site map data.

Introduction

One of the most useful measures of the extent of DNA polymorphism is nucleotide diversity (π), which is defined as the average number of either nucleotide differences or substitutions per site for a group of DNA sequences (alleles) sampled from a population (Nei 1987, chap. 10). During the past decade this quantity has been estimated for many different genes in various organisms. However, the variance of an estimate ($\hat{\pi}$) of this quantity has not been well studied. Nei and Tajima (1981) presented a formula for the sampling variance of $\hat{\pi}$ that is generated at the time of allelic sampling. This variance does not include the variance caused by the stochastic change of allele frequencies in the evolutionary process, so that it can be an underestimate of the total variance. Tajima (1983) studied the total variance of $\hat{\pi}$, including both stochastic and sampling variances, but his formulation depends on the following assumptions: (1) neutral mutations, (2) constant population size, (3) random sampling of genes, and (4) no errors in the estimation of nucleotide substitutions between genes sampled. In practice, these assumptions generally do not hold. Particularly the variance of an estimate of the number of nucleotide substitutions can be quite high when the number of nucleotides studied is small. Note also that DNA sequences are often determined not for a random sample of genes but for a set of different alleles that are deliberately chosen from different geographical areas.

It is not easy to estimate the total variance from single-locus data without making assumptions (1) and (2) above, because there are many different forms of selection that are usually unknown and because the pattern of change in population size varies from case to case. However, it is possible to compute the variance that is due to the estimation errors of nucleotide substitutions. This variance includes some but not all of the effects of stochastic changes of allele frequencies in the past, but it is important because it refers specifically to the *sampled sequences* whether these are random samples or not. This variance has recently been used to detect an enhanced rate of nonsynon-

1. Key words: nucleotide diversity, nucleotide divergence, restriction-site data.

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

Mol. Biol. Evol. 6(3):290–300. 1989.

0737-4038/89/0603-006$02.00

ymous (amino acid–altering) nucleotide substitution in the antigen-recognition site of the major histocompatibility complex (MHC) loci in humans and mice (Hughes and Nei 1988). By analogy, it is also possible to compute the corresponding variance of an estimate of nucleotide divergence between two populations (Nei and Li 1979). The purpose of the present paper is to present methods for computing these variances.

Nucleotide Diversity

DNA Sequences

Let us first indicate that there are two different ways of defining nucleotide diversity. One is the average of the proportion of different nucleotides between two sequences (p_{ij}) over all pairwise comparisons, and the other is the average of the number of nucleotide substitutions per site between two sequences (d_{ij}). In the computation of $\hat{\pi}$, d_{ij} can be estimated from p_{ij} by the following formula (Jukes and Cantor 1969):

$$\hat{d}_{ij} = -(3/4)\log_e(1 - 4p_{ij}/3)\,. \tag{1}$$

Although this formula depends on the assumption of equal rates of substitution among the four nucleotides, it is sufficient for our purpose, because d_{ij} is usually <0.1 and because, for this range of d_{ij}, this and other formulas give essentially the same result (Nei 1987, pp. 71–73).

In most eukaryotic genes, p_{ij} is so small that $\hat{d}_{ij}$ is nearly the same as p_{ij}. In such loci as MHC, however, this is not the case (e.g., see Hughes and Nei 1988), and $\hat{d}_{ij}$ may be substantially greater than p_{ij}. In general, $\hat{d}_{ij}$ is more appropriate for evolutionary studies than is p_{ij} because of its cumulative nature. In the following, we therefore use the following definition of $\hat{\pi}$:

$$\hat{\pi} = \sum_{i \neq j} \hat{d}_{ij}/[n(n-1)]\,, \tag{2}$$

where n is the number of DNA sequences sampled (not necessarily at random) and $\hat{d}_{ij}$ is the estimate of the number of nucleotide substitutions per site between sequences i and j. $\sum_{i \neq j}$ stands for summation for all pairwise comparisons.

The variance (V) of $\hat{\pi}$ defined in equation (2) is given by

$$V(\hat{\pi}) = \frac{1}{[n(n-1)]^2}\left[\sum_{i \neq j} V(\hat{d}_{ij}) + \sum_{i \neq j}\sum_{k \neq l} \mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl})\right]. \tag{3}$$

Therefore, if we know the variances and covariances in the bracket of the right-hand side of equation (3), $V(\hat{\pi})$ can be computed. $V(\hat{d}_{ij})$ is given by

$$V(\hat{d}_{ij}) = 9p_{ij}(1 - p_{ij})/[(3 - 4p_{ij})^2 m]\,, \tag{4}$$

(Kimura and Ohta 1972), where m is the number of nucleotides examined per sequence.

To compute $\mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl})$, it is necessary to know the phylogenetic relationship among all sequences sampled. If we assume no intragenic recombination, this relationship can be estimated from a phylogenetic tree constructed for the sequences.

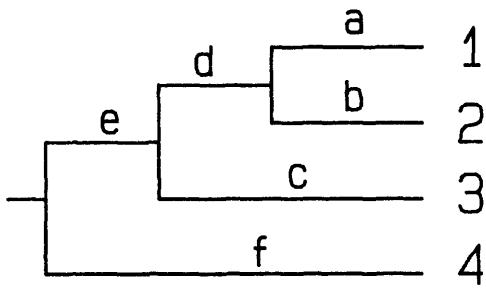

FIG. 1—Hypothetical phylogenetic tree. 1–4 Represent the DNA sequences sampled, whereas a–f denote branches.

There are several methods for constructing a tree from DNA sequence data (e.g., see Nei 1987, chap. 11), but most of them give similar estimates of the variance or standard error of $\hat{\pi}$ (see Discussion). We therefore suggest that the simple unweighted pair-group method (UPGMA) (Sneath and Sokal 1973) be used.

Suppose that a tree constructed from four DNA sequences is of the form given in figure 1. In a UPGMA tree, the lengths of the branches leading to two extant sequences are identical with each other. Thus, the sum of the lengths of branches a and d is identical with the length of branch c. Therefore, one can easily compute the expected distance between sequences i and j from the tree. We call this distance a *patristic distance* and denote it by b_{ij}. The b_{ij} is identical with $\hat{d}_{ij}$ if i and j are connected by only one node. Thus, $b_{12} = \hat{d}_{12}$. Otherwise, b_{ij} is not necessarily equal to $\hat{d}_{ij}$. However, they are usually quite similar to each other when molecular data are used (see the example given below).

In the following, we therefore assume that $b_{ij} = \hat{d}_{ij}$ and compute $V(\hat{d}_{ij})$ and $\mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl})$ under this assumption, since this simplifies the computation tremendously without affecting the final results appreciably [compare the results obtained by using UPGMA and the neighbor-joining (NJ) method, given later]. Under this assumption, $\mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl})$ can easily be computed. For example, $\mathrm{Cov}(\hat{d}_{12}, \hat{d}_{23})$ in figure 1 is obtained by noting that the only shared branch between $\hat{d}_{12}$ and $\hat{d}_{23}$ is b and that the nucleotide substitutions in other branches occur independently. Therefore, $\mathrm{Cov}(\hat{d}_{12}, \hat{d}_{23})$ is equal to the variance of branch b. The length of this branch can be estimated by UPGMA, and let us denote it by b. The p_{ij} value corresponding to this value is therefore given by

$$p = (3/4)[1 - e^{-4b/3}] \tag{5}$$

[see Nei 1987, eq. (5.2)]. $\mathrm{Cov}(\hat{d}_{12}, \hat{d}_{23})$ is then obtained by substituting the above p for p_{ij} in equation (4). Similarly, all other covariances can be obtained by using the phylogenetic relationships of sequences. When there is no shared branch between $\hat{d}_{ij}$ and $\hat{d}_{kl}$ as in the case of $\hat{d}_{12}$ and $\hat{d}_{34}$, the covariance between them will be zero. The principle of this procedure is similar to that of Nei et al.'s (1985) method for computing the variance of a branch point of a UPGMA tree, though their method is a little more complicated.

Despite the above simplification, the actual computation of $V(\hat{\pi})$ can be quite tedious when the number of sequences examined is large. We have therefore developed

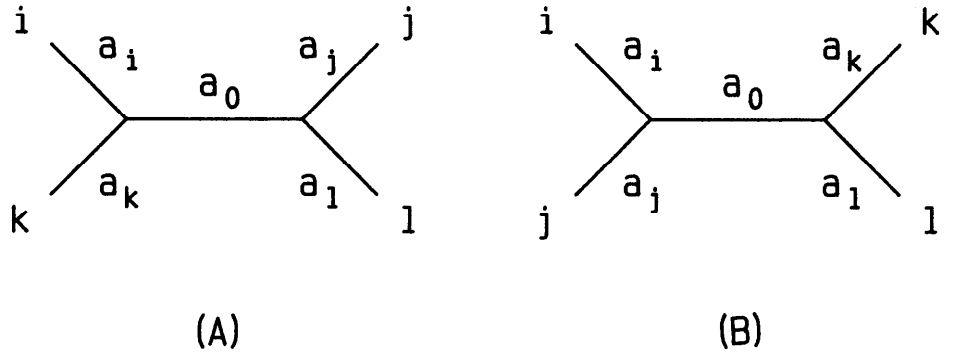

FIG. 2.—Diagrams showing two different unrooted trees for four DNA sequences i, j, k, and l. a_i = branch length.

the following algorithm, which is suited for computer computation. In this algorithm, both $V(\hat{d}_{ij})$ and $Cov(\hat{d}_{ij}, \hat{d}_{kl})$ can be computed by the same procedure.

Algorithm

Step 1

Construct a tree from a matrix of observed distances, $\hat{d}_{ij}$, by using UPGMA, and compute patristic distances, b_{ij}, for all sequence comparisons.

Step 2

To obtain $V(\hat{d}_{ij})$ or $Cov(\hat{d}_{ij}, \hat{d}_{kl})$, compute the following quantities:

$$\begin{aligned} b_1 &= b_{ij} + b_{kl}; \\ b_2 &= b_{ik} + b_{jl}; \\ b_3 &= b_{il} + b_{jk}. \end{aligned} \tag{6}$$

Step 3

Find the minimum value among b_1, b_2, and b_3 and denote it by b_m. Then compute

$$b = (b_1 - b_m)/2 \tag{7}$$

and the p value in equation (5). Here b represents the total length of the branches shared by sequences i,j and k,l. When there is no shared branch between i,j and k,l, $b = 0$.

Let us explain this principle by using figure 2. We first note that any set of four sequences can be represented by either diagram (A) or diagram (B). In diagram (A), $b_{ij} = a_i + a_j + a_o$, $b_{kl} = a_k + a_l + a_o$, $b_{ik} = a_i + a_k$, $b_{jl} = a_j + a_l$, $b_{il} = a_i + a_l + a_o$, and $b_{jk} = a_j + a_k + a_o$.

Therefore,

$$\begin{aligned} b_1 &= a_i + a_j + a_k + a_l + 2a_o; \\ b_2 &= a_i + a_j + a_k + a_l; \\ b_3 &= a_i + a_j + a_k + a_l + 2a_o. \end{aligned}$$

Thus, $b_m = b_2$, and $b = (b_1 - b_m)/2 = a_o$. This proves that b is equal to the length of the branch shared by sequences i,j and k,l. In the case of diagram (B), equation (6) gives

$$b_1 = a_i + a_j + a_k + a_l;$$

$$b_2 = a_i + a_j + a_k + a_l + 2a_o;$$

$$b_3 = a_i + a_j + a_k + a_l + 2a_o .$$

Therefore, $b_m = b_1$, and $b = 0$, showing that there is no shared branch between i, j and k, l.

Note that equation (7) holds even for computing $V(\hat{d}_{ij})$. In this case, $i = k$ and $j = l$. Therefore, b becomes $a_i + a_j + a_o$ in diagram (A) and $a_i + a_j$ in diagram (B). Furthermore, when $i = k$ and $j \neq l$, b becomes $a_i + a_o$ in diagram (A). This again represents the length of the branches shared by i,j and k,l.

Step 4

Compute $\mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl})$ for all pairs of i,j and k,l, using

$$\mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl}) = 9p(1 - p)/[(3 - 4p)^2 m] . \tag{8}$$

For the case of $i = k$ and $j = l$, $\mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl})$ becomes equal to $V(\hat{d}_{ij})$.

Step 5

Compute $V(\hat{\pi})$ by using equation (3).

In the above algorithm there is no need of drawing a UPGMA tree. As long as b_{ij}'s are estimated, $V(\hat{\pi})$ can be computed automatically. We have developed a computer program for computing b_{ij}'s and $V(\hat{\pi})$, and a copy of it is available on request.

Restriction-Site Data

In the estimation of $\hat{\pi}$ the restriction-enzyme technique is often used, because it is much simpler than DNA sequencing. When restriction-site maps are available for each DNA sequence, estimates of $\hat{d}_{ij}$'s and their variances and covariances can easily be obtained. It is therefore possible to compute $\hat{\pi}$ by using essentially the same method as that given above. The only difference is that raw data are not nucleotide sequences but restriction-site maps.

Suppose that restriction-site maps produced by restriction enzymes with r recognition sites ($r = 4$, $^{16}/_3$, or 6 in most cases; see Nei 1987, p. 102) are available for n DNA sequences [e.g., mitochondrial DNAs (mtDNAs)]. Let m_i and m_j be the number of restriction sites in DNA sequences i and j, respectively, and let m_{ij} be the number of restriction sites shared by the two sequences. A maximum likelihood estimate of the number of nucleotide substitutions between the two sequences is then given by equation (1) with

$$p_{ij} = 1 - \hat{S}^{1/r} \tag{9}$$

or by

$$\hat{d}_{ij} = [-\log_e \hat{S}]/r \tag{10}$$

approximately, where $\hat{S} = 2m_{ij}/(m_i + m_j)$ (Nei and Li 1979). We determine the patristic distances (b_{ij}) from $\hat{d}_{ij}$'s by using a UPGMA tree as before and compute the length (b) of shared branches between sequences i,j and k,l. Once b is obtained, the expected value of $\hat{S}$ is given by

$$S = (1 - p)^r \tag{11}$$

with

$$p = (3/4)[1 - \exp(-4b/3)]$$

or by

$$S = e^{-rb} \tag{12}$$

approximately. $V(\hat{d}_{ij})$ or $\mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl})$ can then be computed by

$$\mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl}) = \frac{9(1 - p)^2(2 - S)(1 - S)}{2r^2\bar{m}(3 - 4p)^2 S} \tag{13}$$

or by

$$\mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl}) = (2 - S)(1 - S)/[2r^2\bar{m}S] \tag{14}$$

approximately, using the above S value (see Nei and Tajima 1983). Here $\bar{m}$ is the average of m_i over all sequences. $V(\hat{\pi})$ is then obtained by equation (3).

In the above formulation we assumed that all restriction enzymes used have the same r value. In practice, this is not always the case, and enzymes with different r values are often used. However, this poses no problem, since $\hat{d}_{ij}$ can again be computed by the maximum likelihood method (Nei and Tajima 1983). Furthermore, once the b value is estimated by the above method, $\mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl})$ can be computed by

$$\mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl}) = \frac{1}{\sum_r [1/\mathrm{Cov}_r(b)]}, \tag{15}$$

where $\mathrm{Cov}_r(b)$ is the value of equation (13) or equation (14) for a given value of r (see Nei et al. 1985 for details). In the computation of $\mathrm{Cov}_r(b)$, S should be computed by equation (11) or equation (12) for each r value separately.

Nucleotide Divergence

Suppose that n_X alleles and n_Y alleles are sequenced from populations X and Y, respectively, at a locus. The extent of nucleotide divergence (average number of net nucleotide substitutions per site) between the two populations can be measured by the following quantity (Nei and Li 1979).

$$\hat{d}_A = \hat{d}_{XY} - (\hat{d}_X + \hat{d}_Y)/2\,, \tag{16}$$

where $\hat{d}_X$ and $\hat{d}_Y$ are the $\hat{\pi}$ values in populations X and Y, respectively, whereas $\hat{d}_{XY}$ is the average number of nucleotide substitutions per site between X and Y.

The variance of $\hat{d}_A$ is therefore given by

$$V(\hat{d}_A) = V(\hat{d}_{XY}) + [V(\hat{d}_X) + V(\hat{d}_Y)]/4 - \mathrm{Cov}(\hat{d}_X, \hat{d}_{XY}) - \mathrm{Cov}(\hat{d}_Y, \hat{d}_{XY}) + \mathrm{Cov}(\hat{d}_X, \hat{d}_Y)/2 . \quad (17)$$

$V(\hat{d}_X)$ and $V(\hat{d}_Y)$ can be computed by equation (3). If we note that

$$\hat{d}_{XY} = \sum_{i=1}^{n_X} \sum_{k=1}^{n_Y} \hat{d}_{ik}/n_X n_Y , \quad (18)$$

then $V(\hat{d}_{XY})$ is given by

$$V(\hat{d}_{XY}) = [\sum_{i,k} \sum_{j,l} \mathrm{Cov}(\hat{d}_{ik}, \hat{d}_{jl})]/(n_X n_Y)^2 . \quad (19)$$

Here $\mathrm{Cov}(\hat{d}_{ik}, \hat{d}_{ik}) = V(\hat{d}_{ik})$. Similarly, $\mathrm{Cov}(\hat{d}_{XY}, \hat{d}_X)$ is obtained by

$$\mathrm{Cov}(\hat{d}_X, \hat{d}_{XY}) = [\sum_{i,j} \sum_{k,l} \mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl})]/[n_X^2(n_X - 1)n_Y] . \quad (20)$$

$\mathrm{Cov}(\hat{d}_Y, \hat{d}_{XY})$ is obtained in the same way, whereas $\mathrm{Cov}(\hat{d}_X, \hat{d}_Y)$ is

$$\mathrm{Cov}(\hat{d}_X, \hat{d}_Y) = [\sum_{i,j} \sum_{k,l} \mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{kl})]/[n_X(n_X - 1)n_Y(n_Y - 1)] . \quad (21)$$

Therefore, $V(\hat{d}_A)$ can be obtained by evaluating all covariances in equation (19), (20), and (21) and the variances and covariances in equations (3).

In this case a large number of covariances must be computed even for moderate sample sizes, but all covariances can be rapidly determined by the algorithm mentioned above. The method mentioned above can also be applied to restriction-site data, with the modification mentioned in the previous section.

Numerical Example

Since the computation of $V(\hat{\pi})$ and $V(\hat{d}_A)$ for nucleotide sequences is straightforward, let us consider an example for restriction-site data. Ferris et al. (1981) studied the restriction-site polymorphism of mtDNAs in chimpanzees. The values of m_i and m_{ij} for five mtDNAs (1–5) from common chimpanzees (*Pan troglodytes*) and for two mtDNAs (6 and 7) from pygmy chimpanzees (*P. panicus*) are presented in table 1. These data were obtained by using 13 enzymes with $r = 6$. The d_{ij} values can therefore be estimated by equation (1), with $p_{ij} = 1 - \hat{S}^{1/6}$. They are presented in table 1 (above the diagonal). For example, $\hat{d}_{12} = 0.0168$ since $\hat{S} = 2 \times 33/(38 + 35) = 0.904$ and $p_{ij} = 0.0167$. In this case, equation (10) also gives $\hat{d}_{ij} = 0.0168$.

The phylogenetic tree constructed by UPGMA is presented in figure 3(A). From this tree we can compute the b_{ij} values for all sequence comparisons, and they are given in table 2 (below the diagonal). Comparison of these values with $\hat{d}_{ij}$'s in table

Table 1

m_i, m_{ij}, and % d_{ij}'s ($100 \times d_{ij}$'s) for Seven mtDNAs from Chimpanzees

	Sequence						
Sequence	1 (1)	2 (2a)	3 (2b)	4 (2c)	5 (2d)	6 (3a)	7 (3b)
1	(38)	1.68	2.20	1.41	2.20	4.32	2.96
2	33	(35)	0.48	1.73	2.03	3.05	3.39
3	32	34	(35)	1.22	1.50	3.05	3.39
4	34	32	33	(36)	0.72	3.86	2.50
5	32	31	32	34	(35)	4.22	3.39
6	29	30	30	29	28	(37)	1.68
7	31	29	29	31	29	33	(36)

NOTE.—The values on and below the diagonal are m_i's and m_{ij}'s for 13 six-base enzymes, respectively. The values above the diagonal are 100 d_{ij}'s. Sequences 1–5 are from common chimpanzees, and the other sequences are from pygmy chimpanzees. The sequence designations in parentheses are those used by the original authors.

SOURCE.—Ferris et al. (1981).

1 indicates that $\hat{d}_{ij}$'s are generally similar to b_{ij}'s. The standardized discrepancy between $\hat{d}_{ij}$'s and b_{ij}'s, i.e.,

$$s = \left[\frac{\sum_{ij} (\hat{d}_{ij} - b_{ij})^2}{n(n-1)} \right]^{1/2}, \tag{22}$$

is 0.0042.

The b_{ij} values can now be used for computing $V(\hat{\pi})$ for common and pygmy chimpanzees. In the case of pygmy chimpanzees, there are only two DNA sequences, so that $\hat{\pi} = \hat{d}_{67} = 0.0168$ and $V(\hat{\pi}) = V(\hat{d}_{67}) = 0.00004535$. The standard error $[s(\hat{\pi})]$ of $\hat{\pi}$ is therefore 0.0067, which is about one-half of $\hat{\pi}$.

In the case of common chimpanzees, we have $\hat{\pi} = 0.0152$. To obtain $V(\hat{\pi})$, 10 variances of $\hat{d}_{ij}$ and 45 covariances of $\hat{d}_{ij}$ and $\hat{d}_{kl}$ must be computed. These variances and covariances are obtained by the algorithm mentioned above. For example, the variance of d_{15} is obtained first by computing p and S with $b_{15} = 0.0187$. Equations (9) and (11) give $p = 0.0185$ and $S = 0.8942$. Noting that $\bar{m} = 36$, we therefore have $V(\hat{d}_{15}) = 0.00005128$ from equation (13). The UPGMA tree in figure 2 shows that $\mathrm{Cov}(\hat{d}_{12},\hat{d}_{24})$ is obtained by considering the length ($b = 0.0081$) of the branches shared by the pathways between 1 and 2 and between 2 and 4. This branch length can also be obtained by our algorithm. In this case, we have $b_1 = b_{12} + b_{24} = 0.0187 + 0.0162 = 0.0349$, $b_2 = b_{12} + b_{24} = 0.0349$, and $b_3 = b_{14} + b_{22} = 0.0187$. Therefore, $b_m = 0.0187$, and $b = (0.0349 - 0.0187)/2 = 0.0081$. Using this b value, we obtain $\mathrm{Cov}(\hat{d}_{12},\hat{d}_{24}) = 0.0000202$. If we compute all variances of $\hat{d}_{ij}$ and covariances of $\hat{d}_{ij}$ and $\hat{d}_{kl}$, we finally obtain $V(\hat{\pi}) = 0.00001866$. Therefore, $s(\hat{\pi}) = 0.0043$.

The computation of $\hat{d}_{XY}$ and $\hat{d}_A$ (and of their standard errors) between common and pygmy chimpanzees can be computed by the same algorithm. They become $\hat{d}_{XY} = 0.0341 \pm 0.0083$ and $\hat{d}_A = 0.0181 \pm 0.0072$.

As mentioned above, the present example is based on restriction-site data only for enzymes with $r = 6$. Actually, Ferris et al. (1981) used two enzymes with $r = {}^{16}/_{3}$

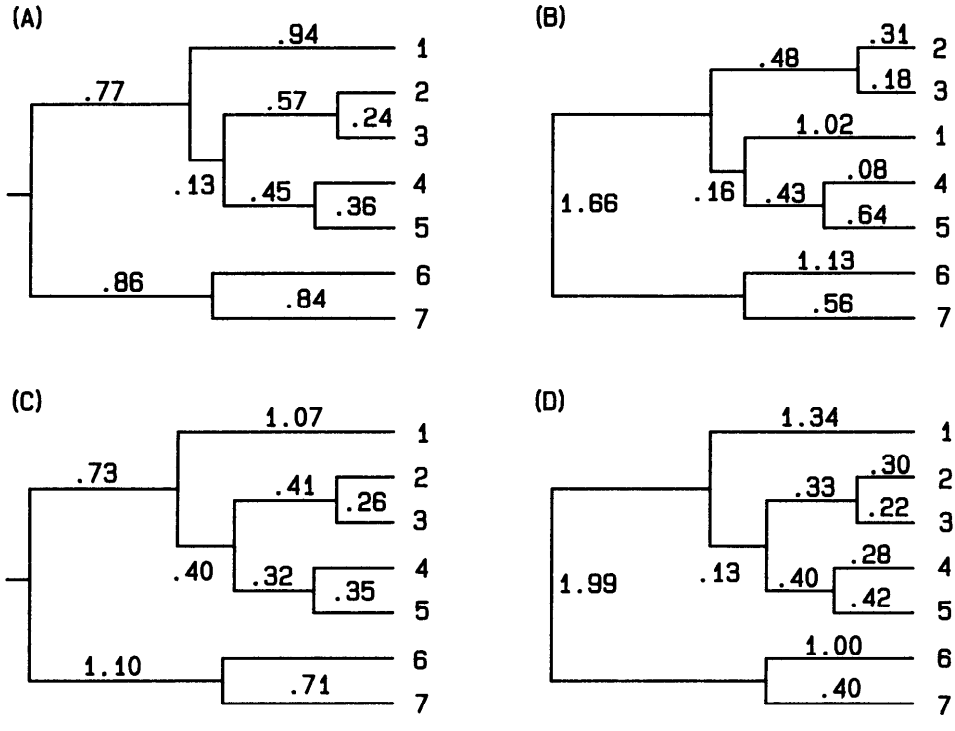

FIG. 3.—Evolutionary trees reconstructed for five mtDNA sequences (1–5) from common chimpanzees and for two mtDNA sequences (6 and 7) from pygmy chimpanzees. (A), Tree obtained by UPGMA from restriction-site data for enzymes with $r = 6$. (B), Tree obtained, by the NJ method, from data for enzymes with $r = 6$. (C), Tree obtained, by UPGMA, from data for three different types of enzymes. (D), Tree obtained, by the NJ method, from data for three different types of enzymes. Branch lengths are given in terms of % d ($100d$).

and one enzyme with $r = 4$, in addition to those with $r = 6$. Furthermore, there were three individuals having sequence 1, two individuals having sequence 3, and three individuals having sequence 4 in common chimpanzees, whereas there were two individuals having sequence 6 in pygmy chimpanzees. Therefore, the actual values of $\hat{\pi}$ and $\hat{d}_A$ (and their standard errors) should be computed by considering all these data (13 mtDNAs). This computation can be done quickly by our computer program. In the present case, we obtain $\hat{\pi} = 0.0144 \pm 0.0034$ for common chimpanzees and $\hat{\pi} = 0.0094 \pm 0.0034$ for pygmy chimpanzees. These values are smaller than those obtained above. By contrast, $\hat{d}_{XY}$ and $\hat{d}_A$ become 0.0383 ± 0.0078 and 0.0264 ± 0.0071, respectively. These are greater than the values reported elsewhere, as expected. The phylogenetic tree obtained is presented in figure 3(C).

Discussion

In the computation of $V(\hat{\pi})$ and $V(\hat{d}_A)$ we suggested that a UPGMA tree be used for finding the covariance of $\hat{d}_{ij}$ and $\hat{d}_{kl}$. Some readers might object to this suggestion, because UPGMA depends on the assumption of constant rate of nucleotide substitution and this assumption does not necessarily hold for real data. Actually, UPGMA need not be used. As long as the length of each branch is estimable, one can use any tree-making method, and the algorithm presented above directly applies. In

Table 2
% b_{ij}'s (100 × b_{ij}'s) Obtained from the UPGMA and NJ Trees

	SEQUENCE						
SEQUENCE	1	2	3	4	5	6	7
1		1.97	1.83	1.52	2.09	3.97	3.40
2	1.87		0.48	1.45	2.01	3.58	3.00
3	1.87	0.48		1.31	1.88	3.44	2.87
4	1.87	1.62	1.62		0.72	3.45	2.88
5	1.87	1.62	1.62	0.72		4.02	3.45
6	3.41	3.41	3.41	3.41	3.41		1.68
7	3.41	3.41	3.41	3.41	3.41	1.68	

NOTE.—The values below the diagonal are the b_{ij}'s obtained from the UPGMA tree, whereas the values above the diagonal are the b_{ij}'s obtained from the NJ tree (see fig. 2).

this case, however, some branch lengths may become negative. If this happens, we suggest that all negative values be set to zero under the assumption that they are due to sampling errors. If this procedure is adopted, most tree-making methods can be used for this purpose. In practice, however, the variances of $\hat{\pi}$ and of $\hat{d}_A$ that are obtained by different tree-making methods are quite similar.

Let us illustrate this point by applying Saitou and Nei's (1987) NJ method to Ferris et al.'s restriction-site data, as in table 1. The NJ method is known to be quite efficient in recovering the correct tree even when substitution rate varies from branch to branch, yet the computer time required is quite short. Application of this method to the $\hat{d}_{ij}$ values in table 1 produces a tree, given in figure 3(B), that has a topology different from that of the UPGMA tree. Note also that the lengths of the branches leading to two extant sequences are no longer equal. The b_{ij} values obtained from this tree are presented in table 2 (above the diagonal). It is interesting to note that the b_{ij}'s obtained from this tree are generally more similar to $\hat{d}_{ij}$'s than to those obtained from the UPGMA tree. The standardized discrepancy is $s = 0.0031$ in this case. However, the standard errors of $\hat{\pi}$ and $\hat{d}_A$ obtained by this procedure are quite close to those obtained from the UPGMA tree. Thus, we have $s(\hat{\pi}) = 0.0044$ for common chimpanzees, $s(\hat{\pi}) = 0.0067$ for pygmy chimpanzees, and $s(\hat{d}_A) = 0.0072$. These values are virtually identical to those obtained by using UPGMA.

Using UPGMA, we previously computed $\hat{\pi}$'s and $\hat{d}_A$ for restriction-site data for 13 sequences including those obtained by enzymes with $r = {}^{16}/_{3}$ and $r = 4$. If we apply the NJ method to the same set of data, we obtain the tree presented in figure 3(D). This tree now has the same topology as that of the UPGMA tree. We also have $\hat{\pi} = 0.0144 \pm 0.0033$ for common chimpanzees, $\hat{\pi} = 0.0094 \pm 0.0034$ for pygmy chimpanzees, and $\hat{d}_A = 0.0264 \pm 0.0071$. These values are again virtually identical to those obtained by using UPGMA.

As mentioned earlier, the variances of $\hat{\pi}$ and $\hat{d}_A$ considered here refer to the sequences sampled without regard to their frequencies in the population. Therefore, they are different from those studied by Nei and Tajima (1981). However, when one wants to test the difference in $\hat{\pi}$ or $\hat{d}_A$ between different regions of genes or between synonymous and nonsynonymous substitutions for the same gene region, these are the variances that should be used (Hughes and Nei 1988). They are applicable whether there is selection or not, as long as the nucleotide substitutions in different sites are

more or less independent. They are of the same nature as the variance of the Jukes-Cantor estimate of nucleotide substitutions between two sequences [eq. (4)].

It should be noted that the variances considered here do not include the stochastic variance that is observed among different loci evolving independently. To evaluate the magnitude of this variance, one has to collect data from many loci and compute the interlocus variance as in the case of electrophoretic loci (Nei 1987, chap. 9). This variance is expected to be quite large, though few such data are available now.

Acknowledgments

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

LITERATURE CITED

FERRIS, S. D., W. M. BROWN, W. S. DAVIDSON, and A. C. WILSON. 1981. Extensive polymorphism in the mitochondrial DNA of apes. Proc. Natl. Acad. Sci. USA **78**:6319–6323.

HUGHES, A. L., and M. NEI. 1988. Pattern of nucleotide substitution at major histocompatibility class I loci reveals overdominant selection. Nature **335**:167–170.

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.

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

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

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

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.

NEI, M., and F. TAJIMA. 1981. DNA polymorphism detectable by restriction endonucleases. Genetics **97**:145–163.

———. 1983. Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics **105**:207–217.

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

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

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

WALTER M. FITCH, reviewing editor

Received August 17, 1988; revision received December 2, 1988