Sampling Errors in Phylogeny[1]

Naoyuki Takahata and Fumio Tajima
National Institute of Genetics (Japan)

The sampling variance of nucleotide diversity or branch length in a phylogenetic tree constructed by any distance method provides a criterion to judge whether a deduction or an inference made from data is statistically significant. However, computation of the sampling variance is usually tedious, particularly when the number of operational taxonomic units (OTUs) or DNA sequences is large, and must rely on computers. Recently, Nei and Jin (1989) have developed a computer algorithm, but it can be applied only to a simple substitution model. In this paper, we derive simple formulas for the minimum and maximum values of the sampling variance, which are independent of underlying substitution models. Application of these formulas demonstrates satisfactorily accurate estimates of the sampling variances and therefore their practical use.

Introduction

The sampling variance we are going to consider is that due to estimation error of nucleotide substitutions. In the method of Nei and Jin (1989; also see Nei et al. 1985), it does not matter whether a sample of DNA sequences is drawn randomly, but it does depend on how substitutions occur among the four nucleotides, and it is assumed that they change equally likely (Jukes and Cantor 1969). In reality, however, nucleotide changes do not necessarily occur at random, and this has led to the development of many elaborate substitution models (e.g., see Nei 1987, pp. 64–73, and references therein). If the method of Nei and his colleagues has difficulty in being accommodated to unequal substitution rates, its application is restricted virtually to the simplest substitution model. Obviously, it is not consistent to use an elaborate substitution model for converting the proportion of nucleotide differences per site (p) to the estimated number of substitutions per site and simultaneously to use the Jukes-Cantor model for computing the sampling variance. In the present paper, we would like to present a simple method that can be applied to any substitution model. Although it provides only the minimum and maximum values of the sampling variance, the computation involved is easy, and the range between the minimum and maximum variances is satisfactorily small. The procedure is similar to that of Nei et al. (1985), but the accuracy turns out to be better, for the reason given later.

Model and Analysis

In most nucleotide substitution models so far proposed (for review, see Kimura 1983, pp. 90–97; Nei 1987, pp. 64–73), it is assumed that substitutions follow (stationary) Markov processes. In other words, the interval between two successive sub-

[1] Key words: estimation error, nucleotide diversity, branch length, substitution model.

Address for correspondence and reprints: Naoyuki Takahata, National Institute of Genetics, Mishima 411, Japan.

Mol. Biol. Evol. 8(4):494–502. 1991.

0737-4038/91/0804-0007$02.00

stitutions per site is exponentially distributed whether substitutions occur at random or with some compositional bias among the four nucleotides. An immediate consequence of this is that the number of substitutions at the kth site (X_k) necessarily follows a Poisson distribution (Takahata, 1991; also see Tavaré 1986). Here we assume the Poisson to be appropriate. If we define

$$D = \frac{1}{n} \sum_{k=1}^{n} X_k$$

as the mean number of substitutions per site, taken over n nucleotide sites compared, then the value of D is a random variable, and the variance is

$$\mathrm{V}(D) = d/n\,, \tag{1}$$

where the lowercase d stands for the expectation of D.

It would be a simple matter to compute the sampling variance of nucleotide diversity or that at any node in a phylogeny if we could use equation (1) immediately. In practice, however, it is virtually impossible to know the actual number of substitutions per site, so the sampling variance (among sites) must be estimated from the same equation that provides the relationship between p and D (Kimura and Ohta 1972; Kimura 1980; Tajima and Nei 1984; Nei et al. 1985; Nei 1987, pp. 64–73). The equation for D for two DNA sequences may be written as

$$D = f(p)\,. \tag{2}$$

The sampling variance of D can then be estimated by

$$\mathrm{V}(D) = \frac{p(1-p)}{n}\left[\frac{df(p)}{dp}\right]^2 \tag{3}$$

(see Serfling 1981, p. 122). In the Jukes-Cantor model, $f(p) = -\frac{3}{4}\ln(1-\frac{4}{3}p)$, so that

$$\mathrm{V}(D) = \frac{9p(1-p)}{n(3-4p)^2} \tag{3a}$$

(Kimura and Ohta 1972). If the difference p is divided into more than one type of difference, such as transitions and transversions, an appropriate multinomial sampling distribution and partial derivatives must be used to derive the variance. For instance, in Kimura's (1980) model, where transition and transversion types of differences are distinguished as P and Q, respectively, the sampling variance is given by

$$\begin{aligned} \mathrm{V}(D) &= \frac{1}{n}[(a^2P + b^2Q) - (aP + bQ)^2]\,; \\ a &= \frac{1}{1-2P-Q}\,; \\ b &= \frac{1}{2}\left[\frac{1}{1-2P-Q} + \frac{1}{1-2Q}\right]. \end{aligned} \tag{3b}$$

In any case, tediousness arises when we need to compute the variance of distance, which is defined appropriately from multiple sequences sampled from either a single or different species. In the former case, nucleotide diversity (π) for m DNA sequences sampled from a single species may be defined as

$$\pi = \frac{2}{m(m-1)} \sum_{i<j}^{m} D_{ij} , \tag{4}$$

where D_{ij} is the estimated number of nucleotide substitutions per site between the ith and jth sequences (Nei and Tajima 1981). The variance of π [$\mathrm{V}(\pi)$] is then estimated as

$$\mathrm{V}(\pi) = 4[m(m-1)]^{-2}[\sum_{i<j} \mathrm{V}(D_{ij}) + \sum_{i<j} \sum_{i'<j'} \mathrm{Cov}(D_{ij}, D_{i'j'})] \tag{5}$$

(Nei and Jin 1989).

When two different species are involved, it is often necessary to compute the mean distance between two different sets of DNA sequences. Suppose that there are two sets, A and B, which contain r and s sequences, respectively. We consider the mean (intercluster) distance between A and B (D_{AB}). This mean distance is defined as

$$D_{\mathrm{AB}} = \frac{1}{rs} \sum_{ij}^{rs} D_{ij} , \tag{6}$$

where D_{ij} is the distance between the ith sequence in A and the jth sequence in B. The variance of D_{AB} is given by equation (7) of Nei et al. (1985):

$$\mathrm{V}(D_{\mathrm{AB}}) = \frac{1}{(rs)^2} [\sum_{ij} \mathrm{V}(D_{ij}) + \sum_{i \neq i'} \sum_{j \neq j'} \mathrm{Cov}(D_{ij}, D_{i'j'})]. \tag{7}$$

In equations (5) and (7), $\mathrm{V}(D_{ij})$ can be computed from equation (3a), equation (3b), or similar equations, but computation becomes tedious when m or r and s are large, because the number of covariance terms is so large in this case.

However, as shown in the Appendix, it is easy to evaluate the range of $\mathrm{V}(\pi)$ and $\mathrm{V}(D)$. The minimum and maximum values of $\mathrm{V}(\pi)$ are, respectively,

$$\mathrm{V}_{\min}(\pi) = 4[m(m-1)]^{-2}\left[\frac{1}{n} \sum_{k} W_k^2 d_k + \sum_{i<j} \mathrm{V}(e_{ij})\right] \tag{8a}$$

and

$$\mathrm{V}_{\max}(\pi) = 4[m(m-1)]^{-2}\left[\frac{1}{n} \sum_{k} W_k^2 d_k + \frac{m(m-1)}{2} \sum_{i<j} \mathrm{V}(e_{ij})\right]. \tag{8b}$$

In the above, $\mathrm{V}(e_{ij}) = \mathrm{V}(D_{ij}) - {}^{1}/_{n} d_{ij}$, as given by equation (A5) in the Appendix.

This quantity accounts for the deviation of $V(D_{ij})$ from the sampling variance expected from the pure Poisson, and manipulation of $V(e_{ij})$ makes our procedure more accurate. W_k in equation (8) is the number of times in which the kth branch appears in all pairwise comparisons of m sequences. For instance, consider the internodal branch of length d_7 in figure 1. Since this branch is involved only when branch 1 or 2 is compared with the remaining branches and since there are four cases for each such comparison, the value of W_7 associated with this branch becomes 8. Similarly, we can easily evaluate the value for any branch. The values of W_k for the tree (including a broken branch) in figure 1 are $W_1 = W_2 = W_3 = W_4 = W_5 = W_{10} = W_6 = 5$, $W_7 = W_9 = 8$, and $W_8 = 9$ in all pairwise comparisons. The range of the sampling variance becomes

$$V_{\max}(\pi) - V_{\min}(\pi) = \left[1 - \frac{2}{m(m-1)}\right]\left[\sigma_W^2 - \frac{1}{n}d_W\right] \le \sigma_W^2 - \frac{1}{n}d_W\,,$$

where we defined

$$\sigma_W^2 = \frac{2}{m(m-1)} \sum_{i<j} V(D_{ij})$$

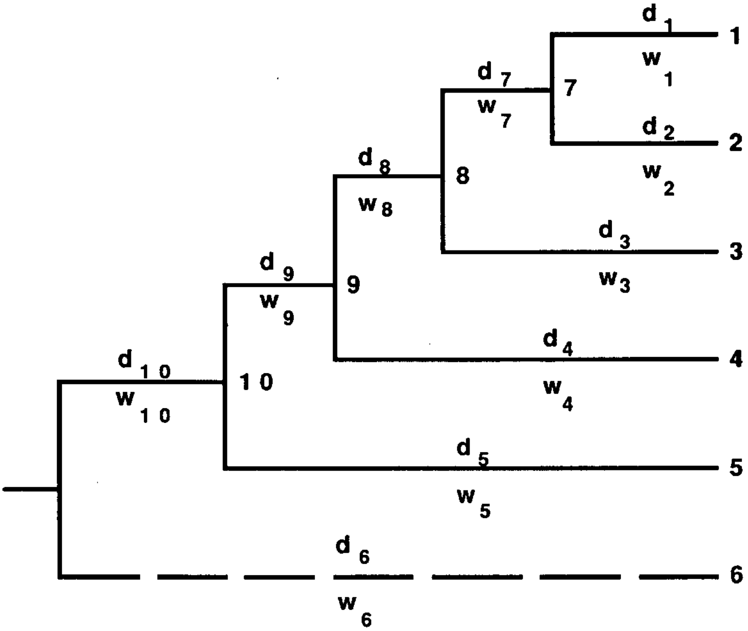

FIG. 1.—UPGMA tree for (solid lines only) five mtDNA sequences of primate (Brown et al. 1982) and (including the broken line) six mtDNA sequences of 2,316 nucleotides sampled from *Drosophila melanogaster* subgroup (Satta and Takahata 1990). In the latter example, $d_1 = d_2 = 0.0063$, $d_3 = 0.0142$, $d_4 = 0.0167$, $d_5 = 0.0183$, $d_6 = 0.0240$, $d_7 = 0.0079$, $d_8 = 0.0025$, $d_9 = 0.0015$, and $d_{10} = 0.0058$. d is the number of substitutions per site per year. Each W stands for the number of times that a particular branch appears in all pairwise comparisons when $V(\pi)$ and D_{AB} are computed.

and

$$d_W = \frac{2}{m(m-1)} \sum_{i<j} d_{ij} .$$

It is thus clear that, as m increases, the range becomes larger but its extent is bounded by the value of the rightmost term.

Similarly, the minimum and maximum variances of $\mathrm{V}(D_{\mathrm{AB}})$ can be obtained as, respectively,

$$\mathrm{V}_{\min}(D_{\mathrm{AB}}) = (rs)^{-2}\left[\sum_{ij} \mathrm{V}(D_{ij}) + \frac{rs-1}{n} \sum_{ij} d_{ij} - \frac{s^2}{n} \sum_{i<i'} d_{ii'} - \frac{r^2}{n} \sum_{j<j'} d_{jj'}\right] \quad (9a)$$

and

$$\mathrm{V}_{\max}(D_{\mathrm{AB}}) = (rs)^{-2}[rs \sum_{ij} \mathrm{V}(D_{ij}) - \frac{s^2}{n} \sum_{i<i'} d_{ii'} - \frac{r^2}{n} \sum_{j<j'} d_{jj'}] \quad (9b)$$

(see the Appendix). In the above, $d_{ii'}$ and $d_{jj'}$ are the expected intracluster distances in A and B, respectively. The computation of equation (9) is more straightforward than that of equation (8), and there is no need to count W_k. Note that the intra- and intercluster distances D are given by equation (2) and that the sampling variance $\mathrm{V}(D_{ij})$ is given by equation (3). For example, when sequences 1 and 2 belong to cluster A (i.e., $1, 2 \in \mathrm{A}$ & $r = 2$) and when sequences 3 and 4 belong to B (i.e., $3, 4 \in \mathrm{B}$ & $s = 2$), equations (9a) and (9b) simply become, respectively,

$$\mathrm{V}_{\min}(D_{\mathrm{AB}}) = \frac{1}{16}[\mathrm{V}(D_{13}) + \mathrm{V}(D_{14}) + \mathrm{V}(D_{23}) + \mathrm{V}(D_{24})] + \frac{3}{16n}(d_{13} + d_{14} + d_{23} + d_{24}) - \frac{1}{4n}(d_{12} + d_{34})$$

and

$$\mathrm{V}_{\max}(D_{\mathrm{AB}}) = \frac{1}{4}[\mathrm{V}(D_{13}) + \mathrm{V}(D_{14}) + \mathrm{V}(D_{23}) + \mathrm{V}(D_{24})] - \frac{1}{4n}(d_{12} + d_{34}) .$$

As mentioned, the maximum variance was also considered by Nei et al. (1985). A difference, however, exists between their formulation and ours. That is, while the present method takes full account of correlation produced by the underlying Poisson process, Nei et al.'s does not, and the last two negative terms in the above $\mathrm{V}_{\max}(D_{\mathrm{AB}})$ expression are ignored [see eq. (22) in Nei et al. 1985]. Obviously, neglecting such terms that are due to negative correlation overestimates $\mathrm{V}_{\max}(D_{\mathrm{AB}})$. As in the case of π, the range of the sampling variance of D_{AB} becomes slightly larger as s and r increase:

$$\mathrm{V}_{\max}(D_{\mathrm{AB}}) - \mathrm{V}_{\min}(D_{\mathrm{AB}}) = \left[1 - \frac{1}{rs}\right][\sigma^2_{\mathrm{AB}} - \frac{1}{n} d_{\mathrm{AB}}] ,$$

where

$$\sigma^2_{AB} = \frac{1}{rs} \sum_{ij} V(D_{ij})$$

and

$$d_{AB} = \frac{1}{rs} \sum_{ij} d_{ij} \, .$$

Numerical Examples

We apply equations (8a) and (8b) to six different mitochondrial DNA (mtDNA) sequences, each with 2,316 nucleotides, which were sampled from the *Drosophila melanogaster* subgroup (one from *D. melanogaster,* three from *D. simulans,* two from *D. mauritiana,* and one from *D. sechellia* sequences) (Satta and Takahata 1990). Here we are interested in π within this subgroup rather than from a single species. The distance matrix among the six sequences can be computed from table 1 of Satta and Takahata (1990) and generates a UPGMA tree of which the topology is identical with that shown in figure 1. If we use equation (3a), we estimate π as 0.0371 and estimate the standard error $\sqrt{V(\pi)}$ as 0.0025 (Nei and Jin 1989). On the other hand, equation (8) estimates $\sqrt{V_{\min}(\pi)}$ as 0.0024 and estimates $\sqrt{V_{\max}(\pi)}$ as 0.0026, with a set of W_k and d_k values for 10 branches of the UPGMA tree in figure 1. Thus the simple equations (8) give a narrow range of $V(\pi)$ value and allow one to compute an accurate sampling error of π in a simple way.

To show how to compute the standard errors of branch lengths, again consider the tree in figure 1 (solid lines only). According to Nei et al. (1985), we define branch lengths (b_i's; i = 7–10 in fig. 1) as $b_7 = \frac{1}{2}D_{12}$ [corresponding to D_{AB} with ($1 \in A$; $2 \in B$)], $b_8 = \frac{1}{4}(D_{13} + D_{23})$ [corresponding to D_{AB} with ($1, 2 \in A$; $3 \in B$)], $b_9 = \frac{1}{6}(D_{14} + D_{24} + D_{34})$ [corresponding to D_{AB} with ($1, 2, 3 \in A$; $4 \in B$)], and $b_{10} = \frac{1}{8}(D_{15} + D_{25} + D_{35} + D_{45})$ [corresponding to D_{AB} with ($1, 2, 3, 4 \in A$; $5 \in B$)]. From equation (9b), we have the maximum variances of b_i's as

$$V(b_7) = \frac{1}{4}V(D_{12}),\ V(b_8) = \frac{1}{8}[V(D_{13}) + V(D_{23})] - \frac{1}{16n} d_{12} \, ;$$

$$V(b_9) = \frac{1}{12}[V(D_{14}) + V(D_{24}) + V(D_{34})] - \frac{1}{36n}(d_{12} + d_{13} + d_{23}) \, ;$$

$$V(b_{10}) = \frac{1}{16}[V(D_{15}) + V(D_{25}) + V(D_{35}) + V(D_{45})] - \frac{1}{64n}(d_{12} + d_{13} + d_{14} + d_{23} + d_{24} + d_{34}) . \qquad (10)$$

Similarly, we can compute the minimum variances of b_i's from equation (9a).

As an example of the sampling variance of branch length in a phylogenetic tree, consider the UPGMA tree in figure 1 (solid lines only), which was obtained from Brown et al.'s (1982) mtDNA sequence data (895 bp) for human (OTU 1), chimpanzee (OTU 2), gorilla (OTU 3), orangutan (OTU 4), and gibbon (OTU 5) (see

Nei 1987, pp. 294). On the basis of the Jukes-Cantor model, Nei et al. (1985) estimated the b_i's as $b_7 = 0.0470$, $b_8 = 0.0563$, $b_9 = 0.0937$, and $b_{10} = 0.1073$ and estimated the sampling errors $\sigma = \sqrt{V(b)}$ as $\sigma_7 = 0.0054$, $\sigma_8 = 0.0052$, $\sigma_9 = 0.0071$, and $\sigma_{10} = 0.0074$. We can compare them with the minimum and maximum values for the four branches; for the four branches, these maximum and minimum values are, respectively, $\sigma_7 = 0.0054$ and 0.0054 for branch 7, $\sigma_8 = 0.0052$ and 0.0054 for branch 8, $\sigma_9 = 0.0068$ and 0.0073 for branch 9, and $\sigma_{10} = 0.0070$ and 0.0077 for branch 10. Although there are some differences between the minimum and maximum values, they are quite narrow, and, as n increases, they become even smaller. The main source of the sampling variance is the limited number of nucleotides compared, rather than the sample size.

Computation of $V_{\min}$ and $V_{\max}$ is simple enough to do with a small calculator. Furthermore, for large values of n the difference between $V_{\min}$ and $V_{\max}$ is so small that they provide useful information. To be conservative, however, one may use $V_{\max}$. It is also clear that equations (8) and (9) can be applied directly to any substitution model. The main reason that we have used the Jukes-Cantor model in the two examples is to compare $V_{\min}$ and $V_{\max}$ with the exact variance. This comparison is possible only under the Jukes-Cantor model.

APPENDIX

Since the derivation of equation (8) is essentially the same as that of equation (9), we derive equation (9) only. Let D_k be the actual number of nucleotide substitutions per site that occurred in the kth branch. Then

$$D_{ij} = \sum_k w_{ijk} D_k + e_{ij} , \tag{A1}$$

where e_{ij} is the error caused by an estimation method and where $w_{ijk} = 1$ if the kth branch connects the ith and jth sequences and where $w_{ijk} = 0$ otherwise. Since D_k follows a Poisson distribution,

$$V(D_{ij}) = \frac{1}{n} \sum_k w_{ijk} d_k + V(e_{ij}) , \tag{A2}$$

where d_k is the expectation of D_k and where $V(e_{ij})$ is the variance of e_{ij}. The covariance between D_{ij} and $D_{i'j'}$ is given by

$$\mathrm{Cov}(D_{ij}, D_{i'j'}) = \frac{1}{n} \sum_k w_{ijk} w_{i'j'k} d_k + \mathrm{Cov}(e_{ij}, e_{i'j'}) . \tag{A3}$$

Substituting equations (A2) and (A3) for the variances and covariances in equation (7), we have

$$V(D_{AB}) = \frac{1}{(rs)^2} [\frac{1}{n} \sum_k W_k^2 d_k + \sum_{ij} V(e_{ij}) + \sum_{ij \neq i'j'} \mathrm{Cov}(e_{ij}, e_{i'j'})] , \tag{A4}$$

where $W_k = \sum_{ij} w_{ijk}$. $V(e_{ij})$ can be estimated as

$$V(e_{ij}) = V(D_{ij}) - \frac{1}{n} \mathrm{d}_{ij} , \tag{A5}$$

since the expectation of D_{ij} is $d_{ij} = \sum_k w_{ijk} d_k$.

Although $\mathrm{Cov}(e_{ij}, e_{i'j'})$ cannot be obtained, we can evaluate $\mathrm{V}_{\min}(D_{\mathrm{AB}})$ and $\mathrm{V}_{\max}(D_{\mathrm{AB}})$, by assuming that $\mathrm{Cov}(e_{ij}, e_{i'j'}) = 0$ and that $\mathrm{Cov}(e_{ij}, e_{i'j'}) = \mathrm{V}(e_{ij})$, respectively; they become

$$\mathrm{V}_{\min}(D_{\mathrm{AB}}) = (rs)^{-2}[\frac{1}{n}\sum_{k} W_k^2 d_k + \sum_{ij} \mathrm{V}(e_{ij})] \tag{A6}$$

and

$$\mathrm{V}_{\max}(D_{\mathrm{AB}}) = (rs)^{-2}[\frac{1}{n}\sum_{k} W_k^2 d_k + rs\sum_{ij} \mathrm{V}(e_{ij})] \tag{A7}$$

(also see Nei et al. 1985). For π, the same procedure can be used, although the number of sequence comparisons is somewhat different. In this case, equations (A6) and (A7) lead to equations (8a) and (8b), respectively.

We may rewrite equations (A6) and (A7) in more convenient forms in terms of intracluster distances. Let the kth branch in A (B) separate r (s) sequences into n_k and r (s) $- n_k$ sequences. The expectation of $\sum_{i<i'} D_{ii'}$ is

$$s^2 \sum_{i<i'} d_{ii'} = s^2 \sum_{k} \sum_{i<i'} w_{ii'k} d_k = s^2 \sum_{k} n_k(r - n_k)d_k = \sum_{k} (rsW_k - W_k^2)d_k \tag{A8}$$

if A contains the kth branch, and it is

$$r^2 \sum_{j<j'} d_{jj'} = \sum_{k} (rsW_k - W_k^2)d_k \tag{A9}$$

if B contains the kth branch. In the above, $W_k = n_k s$ or $W_k = (r - n_k)s$ for equation (A8), and $W_k = n_k r$ or $W_k = (s - n_k)r$ for equation (A9). Using the expression $W_k^2 d_k$ in equations (A8) or (A9), we can rewrite $\mathrm{V}_{\min}(D_{\mathrm{AB}})$ and $\mathrm{V}_{\max}(D_{\mathrm{AB}})$ as in equations (9a) and (9b).

Acknowledgment

This communication is a product of discussion with Drs. S. Horai and Y. Satta, who wanted to estimate by hand the standard error of branch length in a tree.

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.

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

———. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

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 L. JIN. 1989. Variances of the average numbers of nucleotide substitutions within and between populations. Mol. Biol. Evol. **6**:290–300.

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.

SATTA, Y., and N. TAKAHATA. 1990. Evolution of mitochondrial DNA and the history of the *Drosophila melanogaster* subgroup. Proc. Natl. Acad. Sci. USA **87**:9558–9562.

SERFLING, R. J. 1981. Approximation theorems of mathematical statistics. J. Wiley, New York.

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

TAKAHATA, N. 1991. Overdispersed molecular clock at the major histocompatibility complex loci. Phil. Trans. R. Soc. Lond. **[B] 243**:13–18.

TAVARÉ, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Pp. 57–86 *in* R. M. MIURA, ed. Lectures on mathematics in the life sciences. Vol. **17.** American Mathematical Society, Providence, R.I.

MASATOSHI NEI, reviewing editor

Received August 23, 1990; revision received January 18, 1991

Accepted January 21, 1991