DNA POLYMORPHISM DETECTABLE BY RESTRICTION ENDONUCLEASES[1]

MASATOSHI NEI AND FUMIO TAJIMA

Center for Demographic and Population Genetics, University of Texas at Houston, Houston, Texas 77025

Manuscript received July 21, 1980
Revised copy received December 12, 1980

ABSTRACT

Data on DNA polymorphisms detected by restriction endonucleases are rapidly accumulating. With the aim of analyzing these data, several different measures of nucleon (DNA segment) diversity within and between populations are proposed, and statistical methods for estimating these quantities are developed. These statistical methods are applicable to both nuclear and nonnuclear DNAs. When evolutionary change of nucleons occurs mainly by mutation and genetic drift, all the measures can be expressed in terms of the product of mutation rate per nucleon and effective population size. A method for estimating *nucleotide* diversity from *nucleon* diversity is also presented under certain assumptions. It is shown that DNA divergence between two populations can be studied either by the average number of restriction site differences or by the average number of nucleotide differences. In either case, a large number of different restriction enzymes should be used for studying phylogenetic relationships among related organisms, since the effect of stochastic factors on these quantities is very large. The statistical methods developed have been applied to data of SHAH and LANGLEY on mitochondrial (mt)DNA from *Drosophila melanogaster*, *simulans* and *virilis*. This application has suggested that the evolutionary change of mtDNA in higher animals occurs mainly by nucleotide substitution rather than by deletion and insertion. The evolutionary distances among the three species have also been estimated.

IN recent years, an increasing number of authors have examined genetic variation of DNA by means of restriction endonucleases. Although this technique does not necessarily detect all genetic variation of DNA, it is much simpler than complete DNA sequencing and thus has a great utility for population genetics and evolutionary studies. Application of this technique to mitochondrial DNA has already generated important information on the rate of nucleotide substitution in evolution and genetic structure of populations (UPHOLT and DAWID 1977; LEVINGS and PRING 1977; AVISE, LANSMAN and SHADE 1979; BROWN, GEORGE and WILSON 1979; SHAH and LANGLEY 1979). The study of nuclear DNA by means of restriction enzymes requires additional techniques such as DNA cloning, so that the progress has been slower. Yet, KAN and DOZY (1978) have identified the restriction-site polymorphism of the β-globin gene region in man and successfully applied the polymorphism to genetic counseling with respect to sickle-cell anemia. Furthermore, BOTSTEIN *et al.* (1980) have initiated an ambitious project

[1] Dedicated to Professor SEWALL WRIGHT on the occasion of his 90th birthday.

for detecting DNA polymorphisms with the aim of constructing a detailed human linkage map.

For the analysis of data generated by these works, new statistical methods are required. Such methods have already been studied by UPHOLT (1977), NEI and LI (1979), KAPLAN and LANGLEY (1979) and GOTOH *et al.* (1979) with respect to the measurement of DNA divergence among different species or populations. However, the theory pertaining to DNA polymorphism within populations is still insufficient. Particularly, the relationship between restriction-site polymorphism and nucleotide polymorphism is not well established. The relationship between DNA divergence and polymorphism is also not clear. In this paper, we shall study these relationships in detail and present some statistical methods for estimating important genetic parameters. We shall present some data analyses to illustrate the methods developed.

MUTATION RATE PER NUCLEON

Nomenclature: Before going into the detail, we would like to introduce two new technical terms to avoid cumbersome descriptive phrases such as "restriction fragment length polymorphism" used by BOTSTEIN *et al.* (1980). In the study of DNA polymorphism, a given segment of DNA is often identified, and the polymorphism of this segment is investigated. Such a segment is sometimes quite long (often about 20 kilobases) and includes many structural genes; whereas, in other cases, it may include only a part of one gene. Furthermore, on some occasions, the DNA segment may include only a noncoding region. In this paper, any segment of DNA will be called a *nucleon.* In physics, this word has been used to indicate a particle of the atomic nucleus, such as a proton or neutron. To our knowledge, however, this word has not been used in biology. Each nucleon is characterized by the restriction-site map or the number and lengths of DNA fragments. Different restriction-site patterns for a given nucleon are called *nucleomorphs.* Nucleomorphs correspond to allelomorphs or alleles at a locus in classical genetics; whereas, nucleons correspond to genes. Two nucleomorphs that differ at the nucleotide level may be regarded as the same when only a small number of restriction enzymes are used. However, as long as the same set of restriction enzymes are used, the nucleomorphs are the unit of polymorphism.

Mitochondrial DNA: Mitochondrial DNA (mtDNA) is circular, and the entire DNA molecule is usually used as a nucleon in studies of mtDNA polymorphisms (AVISE, LANSMAN and SHADE 1979; BROWN 1980). In some cases, however, the nucleon studied represents a part of the DNA (SHAH and LANGLEY 1979). Let m_T be the total number of nucleotides in a nucleon and g_1, g_2, g_3 and g_4 be the frequencies of guanines (G), cytosines (C), adenines (A) and thymines (T) in the 5′–3′ strand of the nucleon. We assume that mutational changes among the four nucleotides are balanced, so that g_i remains constant (see APPENDIX). We also assume that all nucleotides are randomly distributed in the nucleon, so that the expected frequency of restriction sites with r nucleotide pairs is given by

$$a = g_1^{r_1} g_2^{r_2} g_3^{r_3} g_4^{r_4} \ , \qquad (1)$$

where r_1, r_2, r_3 and r_4 are the numbers of nucleotides G, C, A and T in the restriction site. Since a is generally small, the number of restriction sites per nucleon (m) will be distributed approximately as a Poisson variate with mean $E(m) = m_T a$.

The number and locations of restriction sites in a nucleon change when one or more of the existing sites are lost or new sites are created by mutation. We assume that all mutational changes of DNA occur by nucleotide substitution. The effect of insertion or deletion will be considered later. Let μ_1, μ_2, μ_3 and μ_4 be the rates of mutation of nucleotides G, C, A and T to other nucleotides, respectively, per host generation. The probability of loss of a restriction site by mutation is therefore

$$1-(1-\mu_1)^{r_1}(1-\mu_2)^{r_2}(1-\mu_3)^{r_3}(1-\mu_4)^{r_4} \\ \approx r_1\mu_1+r_2\mu_2+r_3\mu_3+r_4\mu_4=r\mu \ , \tag{2}$$

where μ is the weighted mean of mutation rates per nucleotide site. Therefore, unless the mutation rate is the same for all nucleotides, μ depends on the type of restriction enzyme used. At any rate, when there are m restriction sites in a nucleon, the total probability of loss of at least one restriction site is $1-(1-r\mu)^m \approx mr\mu$. Since the expected value of m is $m_T a$, this probability may be written as $m_T ar\mu$. At steady state, the probability of appearance of new restriction sites must be equal to the probability of loss. Indeed, this can be proved mathematically, as shown in the APPENDIX. Therefore, the probability of mutational change of restriction-site sequence or nucleomorph is

$$v \approx 2m_T ar\mu \ . \tag{3}$$

In the case of linear nucleons, the total number of sequences of r nucleotides is $m_T - r + 1$, but since m_T is generally very large, formula (3) still applies approximately. KAPLAN and LANGLEY (1979) derived a formula equivalent to (3), but their formula is much more complicated because of the different assumptions they made (see APPENDIX).

Nucleon polymorphisms are generally studied by using several different restriction enzymes. When s such enzymes are used, the mutation rate per nucleon is given by

$$v = 2m_T \sum_{i=1}^{s} a_i r_i \mu_i = 2m_T \sum_{i=1}^{s} a_i r_i \bar{\mu} \ , \tag{4}$$

where a_i, r_i and μ_i are the values of a, r and μ for the ith enzyme, respectively, and $\bar{\mu}$ is the weighted mean of μ_i, *i.e.*, $\bar{\mu} = \sum_{i=1}^{s} a_i r_i \mu_i / \sum_{i=1}^{s} a_i r_i$.

At this point, let us consider the values of a and r for those restriction enzymes that recognize several different nucleotide sequences. There are a number of different types of enzymes that do not recognize a unique sequence as given in Table 1, but we shall consider only one example, since a and r for other enzymes can be determined in the same way. We consider *Hae*I, which recognizes four

TABLE 1

Examples of various types of restriction enzymes, expected frequency of restriction sites (a), *the mutation rate per site* (rμ) *and the mutation rate per nucleon* (v)

	a		Mutation rate (v)		
			Per site	Per nucleon	
Enzyme	(1)*	(2)†	($r\mu$)	(1)*	(2)†
4-base					
*Hae*III [GGCC]	3.9×10^{-3}	1.0×10^{-4}	4μ	500μ	12.8μ
*Mbo*I [GATC]	3.9×10^{-3}	1.6×10^{-3}	4μ	500μ	204.8μ
5-base					
*Eco*RII [CC$\binom{\mathrm{A}}{\mathrm{T}}$GG]	2.0×10^{-3}	8.0×10^{-5}	$(14/3)\mu$	166.7μ	11.9μ
*Hinf*I [GANTC]	3.9×10^{-3}	1.6×10^{-3}	4μ	500μ	204.8μ
6-base					
*Bam*HI [GGATCC]	2.4×10^{-4}	1.6×10^{-5}	6μ	46.9μ	3.1μ
*Eco*RI [GAATTC]	2.4×10^{-4}	2.6×10^{-4}	6μ	46.9μ	49.2μ
*Hae*I [$\binom{\mathrm{A}}{\mathrm{T}}$GGCC$\binom{\mathrm{T}}{\mathrm{A}}$]	9.8×10^{-4}	6.4×10^{-5}	$(16/3)\mu$	166.7μ	10.9μ
*Hind*II [GTPyPuAC]	9.8×10^{-4}	4.0×10^{-4}	$(16/3)\mu$	166.7μ	68.3μ
*Sma*I [CCCGGG]	2.4×10^{-4}	1.0×10^{-6}	6μ	46.9μ	0.2μ

A nucleon is assumed to consist of 16000 nucleotides.

* (1) $g_G = g_C = g_A = g_T = 0.25$.

† (2) $g_G = g_C = 0.1, g_A = g_T = 0.4$.

different sequences, *i.e.*, AGGCCT, AGGCCA, TGGCCT and TGGCCA. Under the assumption of random nucleotide distribution, the a value for this enzyme is

$$a = (g_3 + g_4)^2 g_1^2 g_2^2 \ .$$

Any mutation in the second to fifth nucleotide sites results in the loss of the restriction site, but the mutation at the first and sixth positions does so only with a certain probability. If we assume that, once mutation occurs, it is equally likely to be to any of the other three nucleotides, this probability is 2/3. Therefore, the probability of loss of existing restriction sites is approximately $2\mu_1 + 2\mu_2 + \frac{2}{3}\mu_3 + \frac{2}{3}\mu_4$, which is equal to $(16/3)\mu$, where μ is the weighted mean of the mutation rates. Thus, if we define $r = 16/3$, formula (4) can be used. The values of a and r for other types of restriction enzymes are given in Table 1.

In the above formulation, we assumed that the frequencies of nucleotides G, C, A and T are known. Unfortunately, this information is not always available. However, the value of $m_T a$ can be estimated by the observed number of restriction sites for a given enzyme. Therefore, the mutation rate per nucleon may be related to the average mutation rate per nucleotide ($\bar{\mu}$) by

$$v = 2 \sum_{i=1}^{s} m_i r_i \bar{\mu} \ , \tag{5}$$

where m_i is the number of restriction sites detected by the ith enzyme. However, since m_i is subject to a large stochastic error, a large number of enzymes should be used to relate v to μ. When r_i is the same for all enzymes used, (5) reduces to $v = 2s\bar{m}r\bar{\mu}$, where $\bar{m}$ is the mean of m_i.

Nuclear DNA: The polymorphism of nuclear DNA is studied in two different ways. One way is to extract and clone a nucleon from the genome by recombinant DNA techniques and digest the cloned nucleon by restriction enzymes. (In this case, the cloned nucleon is generally linear, but the above theory directly applies unless the nucleon is very small.) The other is first to digest the entire DNA of an organism and then separate the different DNA fragment lengths by electrophoresis. The fragment that is homologous to a particular nucleon is then identified by using a DNA probe (BOTSTEIN *et al.* 1980). The DNA probe is the same piece of DNA as the nucleon under investigation, but labeled with a radioactive isotope. In this case, any fragment that is partially or wholly homologous to the nucleon can be detected, as seen from Figure 1. It is clear that in this case nucleomorphs that have one or more restriction sites in the nucleon region will be picked up. The nucleomorphs that have no restriction sites in this region, but cover the entire region, will also be picked up, since they are homologous to the nucleon. Therefore, all nucleomorphs detected by this method refer to the genetic variation of the nucleon under investigation, even though some fragments may include DNA from outside the nucleon region. Thus, the above theory again applies.

Effects of deletion and insertion: The nucleotide sequence of DNA often changes due to deletion and insertion, particularly in noncoding regions of DNA. In higher animals, most changes of mitochondrial DNA seem to result from nucleotide substitution, but, in nuclear DNA, deletion and insertion play an important role. These deletions and insertions are considered generally to result from unequal crossing over, but it is not easy to make a precise formulation of the effects of deletion and insertion on DNA polymorphism, since the pattern of DNA change due to these events is not well understood. However, recent data on nucleotide sequences of the globin family genes suggest that the number of nucleotides involved in each event of deletion and insertion is relatively small and that these events occur recurrently (*e.g.*, EFSTRATIADIS *et al.* 1980). If this observation is general, the change in the restriction site pattern or nucleomorph due to deletion or insertion would occur roughly at a constant rate per generation. In this case, the nucleon polymorphism can be studied in the same way as that due to nucleotide substitution. In this case, v simply represents this mutation rate per nucleon per generation.

In addition to nucleotide substitution, deletion and insertion, usual recombination may also change the nucleotide sequence of DNA. However, the effect of recombination seems to be small, as long as a nucleon of about 20 kilobases or less is studied.

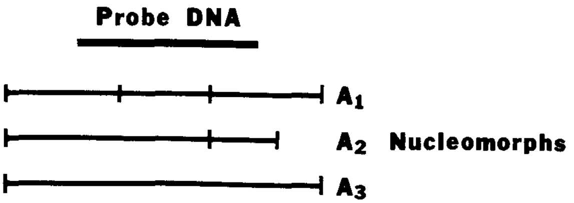

FIGURE 1.—DNA fragments that can be identified by a probe DNA.

GENETIC VARIATION WITHIN POPULATIONS

Nucleon diversity: Suppose that there are l nucleomorphs at a given nucleon site in DNA in a population and let p_i be the frequency of the ith nucleomorph. By analogy to heterozygosity or gene diversity (NEI 1975), we can define the nucleon diversity by

$$h = 1 - \sum_{i=1}^{l} p_i^2 \ . \tag{6}$$

In randomly mating diploid organisms, h for nuclear DNA is equal to heterozygosity. Mitochondrial DNAs are maternally inherited, and in mammals there is little genetic heterogeneity among mtDNAs in one host individual (UPHOLT and DAWID 1977), but it is still a good measure of genetic variation among host individuals. Suppose that, at this nucleon site, a sample of n nucleons is drawn from this population, and the sample frequency of the ith nucleomorph is x_i. Then the nucleon diversity may be estimated by

$$\hat{h} = n(1 - \sum_{i=1}^{l} x_i^2)/(n-1) \ . \tag{7}$$

The sampling variance of $\hat{h}$ can be computed by the same formula as that for gene diversity (NEI 1978).

In addition to the sampling errors at the time of nucleon sampling, $\hat{h}$ is subject to random variation due to stochastic changes of gene frequencies. If we assume no selection, the mean and variance of h over the stochastic process are $H = M/(1+M)$ and $V(h) = 2M/(1+M)^2\,(2+M)\,(3+M)$, respectively (WATTERSON 1974; LI and NEI 1975; STEWART 1976). Here $M = 4N_e v$, where N_e and v denote the effective population size and the mutation rate per nucleon, respectively. It is therefore clear that nucleon diversity depends on the mutation rate, which in turn depends on the types and numbers of restriction enzymes used. As an example, suppose that $g_1 = g_2 = g_3 = g_4$ and that application of *Eco*RI gives $H = 0.091$, with $4N_e v = 0.1$. Then application of *Hae*III is expected to give $H = 0.516$ with $4N_e v = 1.066$ (see Table 1). In this connection, one might wish to estimate the value of $4N_e v$ under the assumption of neutral mutations by equating the observed diversity to its expectation, *i.e.*, using the relationship $h = M/(1+M)$ or $M = h/(1-h)$. However, it is known that this method overestimates $4N_e v$, unless a large number of nucleon sites are used and H is estimated by the average of $\hat{h}$ over all sites (ZOUROS 1979). According to EWENS (1972), a better estimate of $4N_e v$ may be obtained from the number of nucleomorphs in the sample, as long as all nucleomorphs are selectively neutral.

Number of nucleomorphs: The extent of nucleon polymorphism can also be studied in terms of the number of nucleomorphs. The number of nucleomorphs in a sample is, however, highly dependent on the sample size. If we use the infinite-allele model of neutral mutations, the expected number of nucleomorphs $[E(k)]$ in a sample n is given by

$$E(k) = \frac{M}{M} + \frac{M}{M+1} + \cdots + \frac{M}{M+n-1} \tag{8}$$

(EWENS 1972). Thus, $E(k)$ increases as n increases. For example, in the case of $M = 0.5$, $n = 10$ gives $E(k) = 2.5$, but $n = 50$ gives $E(k) = 3.3$. $E(k)$ also depends on the M value, which in turn depends on the types and numbers of restriction enzymes used. For example, in the case of $n = 50$, $M = 0.1$ gives $E(k) = 1.5$; whereas, $M = 0.5$ gives $E(k) = 3.3$. A detailed numerical relationship among $E(k)$, n and M is given by EWENS (1972). In general, it is more efficient to increase M than to increase n in order to detect more nucleomorphs. As mentioned earlier, it is possible to estimate M from the number of nucleomorphs. EWENS (1972) gives a table for obtaining a maximum likelihood estimate ($\hat{M}$) of M and the stochastic variance of $\hat{M}$.

Mean number of restriction-site differences: Another measure of nucleon variation is the mean number of restriction-site differences between two randomly chosen nucleons. This number is defined as

$$\nu = \sum_{ij} p_i p_j \nu_{ij} \ , \tag{9}$$

where v_{ij} is the number of restriction-site differences between the ith and jth nucleomorphs and summation is taken over all combinations of nucleomorphs. Note $\nu_{ii} = 0$. Since the expectation of $x_i x_j$ $(i \neq j)$ is $p_i p_j$ $(1 - 1/n)$, ν may be estimated from sample nucleomorph frequencies by the following unbiased estimator.

$$\hat{\nu} = \frac{n}{n-1} \sum_{ij} x_i x_j \nu_{ij} \ . \tag{10}$$

The sampling variance of $\hat{\nu}$ for given values of p_i is obtained by considering the multinomial sampling of nucleomorphs. It becomes

$$V(\hat{\nu}) = \frac{4}{n(n-1)} [(6-4n) (\sum_{i<j} p_i p_j \nu_{ij})^2 + (n-2) \Sigma p_i p_j p_k \nu_{ij} \nu_{ik} + \sum_{i<j} p_i p_j \nu_{ij}{}^2] \ . \tag{11}$$

When there are only two types of nucleomorphs, (11) reduces to

$$V(\hat{\nu}) = \frac{4}{n(n-1)} [(6-4n) p_1^2 p_2^2 + (n-1) p_1 p_2] \nu_{12}^2 \ . \tag{12}$$

When the number of possible restriction sites per nucleon is large, the mean and variance of ν over the stochastic process can be studied by using the so-called infinite-site model (WATTERSON 1975; LI 1977). Strictly speaking, this model is not appropriate, since some recurrent mutations, as well as back mutations, may occur at the restriction-site level (see NEI and LI 1979). However, the probability of occurrence of these mutations is generally small compared to that of new mutations, so that the infinite-site model approximately applies (see also

GRIFFITHS 1980). At any rate, if we use this model, the mean $[E(\nu)]$ and variance $[V(\nu)]$ of ν are given by

$$E(\nu) = M, \tag{13}$$

$$V(\nu) = M + M^2 \tag{14}$$

(WATTERSON 1975). It is known that (13) is not affected by recombination within the nucleon (KIMURA 1969). Since (10) is an unbiased estimate of M, it is obvious that (13) is better for estimating M than using the formula $\hat{M} = \hat{h}/(1 - \hat{h})$.

Number of segregating restriction sites: A quantity closely related to ν is the number of segregating restriction sites, *i.e.*, the number of restriction sites that are polymorphic in a sample of n nucleons. If we use WATTERSON's (1975) infinite-site model, the mean $[E(\kappa)]$ and variance $[V(\kappa)]$ of this number (κ) over the stochastic process become

$$E(\kappa) = M[1 + 1/2 + 1/3 + \cdots (n - 1)^{-1}] , \tag{15}$$

$$V(\kappa) = E(\kappa) + M^2 \sum_{i=1}^{n-1} 1/i^2 . \tag{16}$$

Therefore, it is possible to estimate M from this number. However, it is interesting to see that $V(\kappa)/E(\kappa)$ is smaller than $V(\nu)/E(\nu)$.

Nucleotide diversity: In the above statistical methods, the genetic variability of nucleons was related to the mutation rate per nucleon or $4N_e v$. Therefore, whether the evolutionary change of restriction sites occurs by nucleotide substitution or by insertion and deletion, the above theories are applicable, as long as the mutation rate per nucleon remains the same. However, when evolutionary change of restriction sites occurs mainly by nucleotide substitution, $4N_e v$ is not a fundamental parameter, since the mutation rate (v) depends on the kind and number of restriction enzymes used. Only when the same kind and same number of restriction enzymes are used can estimates of these parameters be compared between different nucleon sites or different organisms.

A more fundamental measure of DNA polymorphism is the proportion of different nucleotides between two randomly chosen nucleomorphs (the number of nucleotide differences per nucleotide site). NEI and LI (1979) have called this nucleotide diversity. Since the genetic variability is measured per nucleotide site in this case, it can be used for comparing any pair of nucleon sites or organisms.

NEI and LI showed how to estimate the number of nucleotide differences between a pair of nucleomorphs from data on the proportion of shared restriction sites $(\hat{S})$ or of shared restriction fragments $(\hat{F})$. For example, if restriction-site maps are available, the number of nucleotide differences per site (π_{ij}) between the ith and jth nucleomorphs may be estimated either by $(-\ln\hat{S})/r$ or by $-(3/2)\ln[(4\hat{S}^{1/2r} - 1)/3]$ (formulae (8) and (9) in NEI and LI 1979), where r is the value given in Table 1. They defined the nucleotide diversity (π) by

$$\pi = \sum_{i \neq j} p_i p_j \pi_{ij} , \tag{17}$$

where p_i is the population frequency of the ith nucleomorph. In practice, however, the number of nucleons sampled (n) is often very small. In this case the expectation of $\sum_{i \neq j} x_i x_j \pi_{ij}$ becomes $\sum_{i \neq j} p_i p_j \pi_{ij}\ (1 - 1/n)$. Therefore, π in (17) should be estimated by

$$\hat{\pi} = \frac{n}{n-1} \sum_{i \neq j} x_i x_j \hat{\pi}_{ij} \ , \tag{18}$$

where $\hat{\pi}_{ij}$ is the estimate of π_{ij}. If we assume that π_{ij} is a constant, the sampling variance of $\hat{\pi}$ is given by (11), replacing v_{ij} by $\hat{\pi}_{ij}$.

In practice, the estimate $\hat{\pi}_{ij}$ is subject to stochastic errors in the evolutionary process, and the variance of $\hat{\pi}_{ij}$ can be computed by the method of NEI and LI (1979). Unfortunately, there were computational errors in their formula. The variance of the estimate ($\hat{S}$) of S should be

$$V(\hat{S}) = \frac{1}{\bar{m}} [\hat{S}(1-\hat{S}) - \hat{S}^2 (1-\hat{S}^{1/2})\{1 + \frac{1}{2}(1-\hat{S}^{1/2})\}] \ , \tag{19}$$

where $\bar{m} = (m_i + m_j)/2$. Using this value, the variance of $\hat{\pi}_{ij} = -\ (\ln\hat{S})/r$ ($\hat{\delta}$ in NEI and LI's paper) is given by

$$V(\hat{\pi}_{ij}) = V(\hat{S})/(r\hat{S})^2 \ . \tag{20}$$

When π_{ij} is estimated by $-(3/2)\ln[(4\hat{S}^{1/2r} - 1)/3]$, it becomes

$$V(\hat{\pi}_{ij}) = [9\hat{S}^{1/r} V(\hat{S})]/[4\hat{S}^{1/2r} - 1)r\hat{S}]^2 \tag{21}$$

instead of (14) in NEI and LI (1979).

If we assume that the population is in equilibrium and the infinite-site model of neutral mutation applies, the nucleotide diversity can be related to the population size and mutation rate. Namely, $E(\pi) = 4N_e\bar{\mu}$, where $\bar{\mu}$ is the average mutation rate per nucleotide site. Therefore, if there is no effect of insertion and deletion, π can be estimated by $1/(2 \sum_{i=1}^{s} m_i r_i)$ of $\hat{M}$ from (5). This relationship may be used for testing whether or not the underlying assumptions are satisfied.

GENETIC VARIATION BETWEEN POPULATIONS

As in the case of genetic variation within populations, the mathematical theory of interpopulational genetic variation based on the infinite-allele model can be directly applied to data on restriction-site variation, as long as the variation is identified at the nucleomorph level. Thus, NEI'S (1972, 1973) theories of genetic distance and gene-diversity analysis can be used for determining the extent of gene differentiation among populations. However, the number of restriction-site differences between nucleomorphs gives additional information about the interpopulational differentiation of DNA among populations.

Restriction-site differences between populations: Let p_i and q_i be the frequencies of the ith nucleomorph in populations X and Y, respectively. Then, the mean number of restriction-site differences between two randomly chosen nucleomorphs, one each from population X and Y, is

$$\nu_{XY} = \sum_{ij} p_i q_j \nu_{ij} \ , \tag{22}$$

where $\nu_{ii} = 0$. If the sample frequencies of the ith nucleomorph are x_i and y_i in populations X and Y, respectively, this number may be estimated by

$$\hat{\nu}_{XY} = \Sigma x_i y_j \nu_{ij} \ . \tag{23}$$

This number, however, includes the restriction-site differences within populations when there is polymorphism. The number of restriction-site differences within populations can be estimated by (10). Therefore, the number of *net* restriction-site differences between the two populations is estimated by

$$\hat{d} = \hat{\nu}_{XY} - (\hat{\nu}_X + \hat{\nu}_Y)/2 \ , \tag{24}$$

where $\hat{\nu}_X$ and $\hat{\nu}_Y$ are the value of $\hat{\nu}$ given by (10) in populations X and Y, respectively. When the two populations are closely related and n is small, $\hat{d}$ may become negative.

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

$$\begin{aligned} V(\hat{d}) &= V(\hat{\nu}_{XY}) + \frac{1}{4}[V(\hat{\nu}_X) + V(\hat{\nu}_Y)] \\ &\quad - \frac{1}{2}[\mathrm{Cov}(\hat{\nu}_{XY}, \hat{\nu}_X) + \mathrm{Cov}(\hat{\nu}_{XY}, \hat{\nu}_Y)] \ , \end{aligned} \tag{25}$$

where

$$\begin{aligned} V(\hat{\nu}_{XY}) = \frac{1}{n_X n_Y} [& (1 - n_X - n_Y)(\Sigma p_i q_i \nu_{ij})^2 \\ & + (n_X - 1)\, \Sigma q_i p_j p_k \nu_{ij} \nu_{ik} \\ & + (n_Y - 1)\, \Sigma p_i q_j q_k \nu_{ij} \nu_{ik} \\ & + \Sigma p_i q_j \nu_{ij}^2] \ , \end{aligned} \tag{26}$$

$$\begin{aligned} \mathrm{Cov}(\hat{\nu}_{XY}, \hat{\nu}_X) = 2[& \sum_{i \neq j} \sum_k p_i p_j q_k \nu_{ij} \nu_{ik} \\ & - (\Sigma p_i p_j \nu_{ij})(\Sigma p_i q_j \nu_{ij})]/n_X \ , \end{aligned} \tag{27}$$

$$\begin{aligned} \mathrm{Cov}(\hat{\nu}_{XY}, \hat{\nu}_Y) = 2[& \sum_{i \neq j} \sum_k q_i q_j p_k \nu_{ij} \nu_{ik} \\ & - (\Sigma p_i q_j \nu_{ij})(\Sigma q_i q_j \nu_{ij})]/n_Y \ , \end{aligned} \tag{28}$$

and $V(\hat{\nu}_X)$ and $V(\hat{\nu}_Y)$ are given by (11).

If we assume that all the losses or gains of restriction sites occur by nucleotide substitution, $\hat{d}$ can be related to the time (t) after divergence between the two populations. As mentioned earlier, the expectations of v_X and v_Y are approximately M, assuming the same N_e for the two populations. On the other hand, v_{XY} is the mean of $m_X + m_Y - 2m_{XY}$ for all pairs of nucleomorphs, where m_X and m_Y are the numbers of restriction sites of a nucleomorph from populations X and Y, respectively, and m_{XY} is the number of restriction sites that are shared by the pair of nucleomorphs. In equilibrium populations, the expectation of m_X or m_Y is $m_T a$; whereas, the expectation of m_{XY} is $m_0 e^{-2r\lambda t}$, where λ is the rate of nucleotide substitution per generation (or per year depending on the definition) and m_0 is the value of m at $t = 0$ (NEI and LI 1979). In the absence of selection, λ is equal to $\bar{\mu}$. If we note that $E(v_{XY}) = E(v_X) = M$ at time 0 and assume that all nucleomorphs are neutral, we have

$$E(v_{XY}) = M + (2m_T a - M)(1 - e^{-2r\bar{\mu}t}) \ , \quad (29)$$

which becomes

$$E(v_{XY}) = M + 2vt \quad (30)$$

approximately if $2r\bar{\mu}t \ll 1$, since $v = 2m_T a r\bar{\mu}$. Therefore, for a relatively short period of time, $\hat{d}$ is expected to increase linearly with time.

However, $\hat{d}$ has a large stochastic variance in addition to the sampling variance given by (25). Particularly when v_{XY} is close to $2m_T a$, its relationship with t is expected to be poor. In this case, NEI and LI's (1979) method of relating the number of nucleotide substitutions to divergence time should be used with the modification of their variance formula given in (20) of this paper.

When a population is subdivided into many subpopulations, it is often important to partition the total genetic variation into the variation within and between subpopulations. This can be done by applying NEI's (1973) method of gene diversity analysis either to nucleon diversity data or to data on the number of restriction-site differences. In the analysis of the latter data, D_{ij}, H_{ij}, H_i and H_j in NEI's (1973) formula (6) should be replaced by d_{ij}, v_{ij}, v_i and v_j, respectively, where i and j each stand for the ith and jth populations.

Some of the statistics proposed in this paper require an extensive computation, and we have developed a computer program, a copy of which is available by request.

NUMERICAL EXAMPLE

SHAH and LANGLEY (1979) studied the genetic variability of mtDNA in *Drosophila melanogaster, simulans* and *virilis* by using restriction endonucleases *Hae*III, *Hpa*II, *Eco*RI and *Hin*dIII. They identified seven nucleomorphs, and the frequencies of these nucleomorphs are given for each species in Table 2. In *D. simulans*, only five nucleons were sampled, and no polymorphism was found.

Intraspecific variation: In *D. melanogaster*, there were four nucleomorphs, and the estimate of nucleon diversity ($\hat{h}$) is 0.71 ± 0.04 (Table 3). From this value, we can estimate $M \equiv 4N_e v$ by $\hat{M} = \hat{h}/(1-\hat{h})$ under the assumption of no selection; it becomes 2.46. On the other hand, the estimate of M obtained from the number of nucleomorphs by using (8) is 1.21, which is about one-half

TABLE 2

Nucleomorph frequencies in mtDNA samples from three species of Drosophila

Nucleomorph	m	m_a	m_b	m_c	s	v	v_a	n†
D. melanogaster	0.1	0.3	0.5	0.1				10
D. simulans					1.0			5
D. virilis						0.6	0.4	10

† n = sample size.

the estimate obtained from nucleon diversity. This difference apparently occurred because nucleon diversity tends to give an overestimate, as mentioned earlier. In this connection it should be noted that mitochondria are maternally inherited, so that the effective population size is 1/4 that for nuclear genes if the sex ratio is 1:1.

To compute the mean number of restriction-site differences, it is convenient to make a table of the number of restriction-site differences (ν_{ij}) for each pair of nucleomorphs (Table 4). The numbers of restriction-site differences in Table 4 are obtainable from the restriction-site maps for nucleomorphs given in SHAH and LANGLEY's Figure 1. From the values of ν_{ij} for *D. melanogaster* in Table 4, we obtain $\hat{\nu} = 1.22$ by using (10). Under the assumption of neutral nucleomorphs, this is an estimate of $4N_e v$. Interestingly, it is close to the estimate obtained from the number of nucleomorphs.

The nucleon polymorphism in *D. melanogaster* is caused by the polymorphism (segregation) at three restriction sites (SHAH and LANGLEY's Figure 1); namely, $\kappa = 3$ in this species. Therefore, if we equate this number to the expectation in (15), we have $3 = 2.83M$, since $n = 10$. Therefore, we have another estimate of $\hat{M} = 1.06$, which is close to the estimates from k and $\hat{\nu}$. This agreement suggests that the change of mtDNA occurs mainly through nucleotide substitution. (In many Drosophila species, mtDNA contains a large A-T-rich segment, but in

TABLE 3

Nucleon diversity ($\hat{h}$), *number of nucleomorphs* (k), *average number of restriction-site differences* ($\hat{\nu}$), *number of segregating sites* (κ), *and nucleotide diversity* ($\hat{\pi}$) *in* D. melanogaster *and* D. virilis

	$\hat{h}$	k	$\hat{\nu}$	κ	$\hat{\pi}$
D. melanogaster					
Estimate	0.71 ± 0.04	4	1.22 ± 0.27	3	0.007 ± 0.002
$4N_e v$	2.46	1.21	1.22	1.06	
$4N_e\bar{\mu}$	0.017	0.008	0.008	0.007	0.007
D. virilis					
Estimate	0.53 ± 0.03	2	0.53 ± 0.09	1	0.004 ± 0.001
$4N_e v$	1.14	0.32	0.53	0.35	
$4N_e\bar{\mu}$	0.008	0.002	0.004	0.002	0.004

TABLE 4

Restriction-site differences and $\hat{S}_{ij}$ values for pairs of nucleomorphs

Nucleomorph	m	m_a	m_b	m_c	s	v	v_d
m		1	1	1	11	13	14
m_a	0.94 1.00		2	2	10	12	13
m_b	1.00 0.94	0.94 0.94		2	12	14	15
m_c	1.00 0.94	0.94 0.94	1.00 0.89		12	14	15
s	0.40 0.84	0.43 0.84	0.40 0.80	0.40 0.80		14	15
v	0.17 0.71	0.18 0.71	0.17 0.67	0.17 0.67	0.22 0.59		1
v_d	0.17 0.67	0.18 0.67	0.17 0.63	0.17 0.63	0.22 0.56	1.00 0.92	

The figures above the diagonal are the restriction-site differences; those below are the $\hat{S}_{ij}$ values. The upper $\hat{S}_{ij}$ value for each pair of nucleomorphs is for *Hae*III and *Hpa*II ($r = 4$); whereas, the lower $\hat{S}_{ij}$ value is for *Eco*RI and *Hind*III ($r = 6$).

D. virilis this segment is deleted. However, since there are no restriction sites for the four enzymes used in this segment in *D. melanogaster* and *D. simulans*, the evolutionary change in this segment can be disregarded. Furthermore, this type of deletion or addition seems to be rare in the mtDNA's of higher animals (BROWN, GEORGE and WILSON 1979).

Let us now relate $4N_ev$ to $4N_e\bar{\mu}$ under the assumption that all evolutionary changes in these species occurred by nucleotide substitution. The average number of restriction sites for *Hae*III, *Hpa*II, *Hind*III and *Eco*RI for the four nucleomorphs of *D. melanogaster* are 3.7, 4, 4.6 and 4, respectively. Therefore, the estimate of v is $\hat{v} = 2\Sigma m_i r_i \bar{\mu} = 164.8\bar{\mu}$. Similarly, we obtain $\hat{v} = 180\bar{\mu}$ for *D. simulans* and $\hat{v} = 100.8\bar{\mu}$ for *D. virilis*. The average of these estimates is $149\bar{\mu}$. In this connection, it should be noted that the G+C content of mtDNA in Drosophila is about 0.22 even if the A-T-rich region is excluded (KAPLAN and LANGLEY 1979), and thus *Hae*III and *Hpa*II do not produce many restriction sites. At any rate, if we use the relationship $v = 149\bar{\mu}$, the estimates of $4N_e\bar{\mu}$ can be obtained from the $4N_ev$ value. They are about 0.008, though the estimate obtained from nucleon diversity is two times larger. The latter value is, however, probably an overestimate for the reason mentioned earlier.

To estimate the nucleotide diversity, π, we must first compute the proportion of shared restriction sites for each pair of nucleomorphs by using the formula $\hat{S}_{ij} = 2m_{ij}/(m_i + m_j)$, where m_i and m_j are the numbers of restriction sites for the *i*th and *j*th nucleomorphs, respectively, and m_{ij} is the number of shared restriction sites (NEI and LI 1979). Then, the estimate of π_{ij} is given by $\hat{\pi}_{ij} = (-\ln\hat{S}_{ij})/r$. When more than one enzyme with the same r value is used, $\hat{S}_{ij}$ should be computed by pooling m_i, m_j and m_{ij} over all enzymes; but, if r is not the same, they should be computed separately. In Table 4, $\hat{S}_{ij}$ values are given separately for *Hae*III and *Hpa*II ($r = 4$) and for *Eco*RI and *Hin*dIII ($r = 6$). From these values, $\hat{\pi}$ can be computed in the same way as $\hat{\nu}$ is computed. In *D. melanogaster* it becomes 0.0061 for enzymes with $r = 4$ and 0.0076 for enzymes with $r = 6$, the average being 0.007. This value is another estimate of $4N_e\bar{\mu}$ and close to the other estimates.

In *D. virilis*, the same computations were done, and the results obtained are given in Table 3. It is clear that all estimates of genetic variability in this species are smaller than those in *D. melanogaster*, and the estimates of $4N_e\bar{\mu}$ are again more-or-less the same, except the estimate from nucleon diversity. This result suggests that the mtDNA in *D. virilis* is less variable than that in *D. melanogaster*. However, since the number of nucleons sampled and the restriction enzymes used are both small, a more extensive study should be done before any definite conclusion is derived. In this connection, it should be noted that the standard errors given in Table 3 are those for nucleon sampling, and the standard errors arising from the stochastic process in the evolutionary process are much larger than these. For example, the $\hat{\nu}$ value for *D. melanogaster* is 1.22. Therefore, the stochastic standard error under the assumption of neutral mutations is $(1.22 + 1.22^2)^{1/2} = 1.6$ from (14). This is larger than the estimate itself. Similarly, the stochastic standard error of $\hat{\nu}$ for *D. virilis* is 0.90, compared with the estimate of 0.53. Therefore, if we use these standard errors, the difference in $\hat{\nu}$ between the two species is not statistically significant. To reduce the standard error relative to the estimate, it is necessary to use a large number of restriction enzymes.

Interspecific variation: The number of net restriction-site differences between species can be computed from the ν_{ij} matrix given in Table 4 by using (24). Noting that $\hat{\nu}$ for *D. simulans* is 0, we obtain $\hat{d}_{ms} = 10.7 \pm 0.3$ as a value of $\hat{d}$ between *D. melanogaster* and *D. simulans*. Similarly, the values for the pairs of *melanogaster-virilis* and *simulans-virilis* are $\hat{d}_{mv} = 12.8 \pm 0.4$ and $\hat{d}_{sv} = 14.1 \pm 0.1$, respectively. Note that the standard error given here includes only that due to nucleon sampling; the stochastic error is much larger than this. At any rate, the sibling species, *melanogaster* and *simulans*, show the smallest value, as expected. However, considering the morphological differences between *virilis* and the *melanogaster-simulans* complex, the $\hat{d}$ values for the other two pairs seem to be too small. This is apparently due to the fact that d is not a good measure of genetic distance when it is large, as mentioned earlier.

A better estimate of interspecific genetic distance is provided by formula (25) of NEI and LI (1979), *i.e.*,

$$\hat{\delta} = \hat{\pi}_{XY} - (\hat{\pi}_X + \hat{\pi}_Y)/2 \ , \tag{31}$$

where $\hat{\pi}_X$ and $\hat{\pi}_Y$ are the estimates of nucleotide diversities in species X and Y, respectively; whereas, $\hat{\pi}_{XY}$ is the estimate of average nucleotide differences between the two species. We must again compute $\hat{\delta}$ for the enzymes with $r=4$ and $r=6$ separately and take the average. It can be obtained from the $\hat{S}_{ij}$ values given in Table 4. In the comparison of *D. melanogaster* with *D. simulans* $\hat{\delta}$ becomes 0.2204 ± 0.0031 for the 4-base enzymes and 0.0300 ± 0.0017 for the 6-base enzymes. The standard error given here is that due to immediate sampling of nucleons and is much smaller than the estimate itself. This error seems to be important only when two closely related populations are compared. On the other hand, the stochastic standard error increases with increasing δ and is much larger than the sampling variance when δ is large.

Computation of the total variance of $\hat{\delta}$, including both sampling and stochastic variances, is very complicated. However, in interspecific comparisons $\hat{\pi}_{XY}$ in $\hat{\delta}$ are generally much larger than $\hat{\pi}_X$ or $\hat{\pi}_Y$. Therefore, the variance of $\hat{\delta}$ is largely due to the variance of π_{XY}. Furthermore, the variance of $\hat{\pi}_{XY}$ is approximately given by (20) if we replace $\hat{S}$ by the weighted average ($\bar{S}$) of $\hat{S}_{ij}$ for all nucleomorph comparisons between two species. In the present case $\bar{S}$ becomes 0.4086 for the 4-base enzymes and 0.8168 for the 6-base enzymes. The former value has a standard error of 0.0932, the latter 0.0187. Obviously, these are minimum standard errors, but they are much larger than the sampling standard errors. If we use these stochastic standard errors, the normal deviate for the difference between the 4-base and 6-base enzymes is $(0.2204-0.0300)/[0.0932^2+0.0187^2]^{1/2}=2.0$, which barely reaches the 5% significance level. However, since our estimates of standard errors are minimum, it is possible that the difference in $\hat{\delta}$ between the 4-base and 6-base enzymes is due to stochastic errors. Indeed, in the remaining two species comparisons, the differences between the 4-base and 6-base enzymes were not statistically significant. At any rate, if we take the average of $\hat{\delta}$ for the two types of enzymes, it becomes 0.1252 ± 0.0672. This value is close to KAPLAN and LANGLEY's (1979) estimate ($2\hat{\eta}=0.1302$) by a different method. (KAPLAN and LANGLEY eliminated the rRNA region, as well as the A-T-rich region, from their study.)

We have done similar computations for the other two species comparisons. The average $\hat{\delta}$ value for *melanogaster* vs. *virilis* was 0.2490 ± 0.1495; whereas, the value of *simulans* vs. *virilis* was 0.2324 ± 0.1429. Therefore, the genetic distance between *D. melanogaster* and *D. simulans* is about half the distances for the other species comparisons. Considering the morphological differences among the three species, however, the latter distances relative to the former still seem to be too small. To make a more definite conclusion, we must use a larger number of restriction enzymes.

DISCUSSION

Our mathematical formulation depends on a number of assumptions. One of the important assumptions is that nucleotides are randomly distributed in a nucleon. Strictly speaking, this assumption does not hold in mtDNA (BROWN 1976). However, for our purpose, a small deviation from randomness does not matter, and the approximate validity can be tested by comparing the observed and expected numbers of restriction sites per nucleon. Actually, in mtDNA the mean number of restriction sites among independent nucleons seems to agree roughly with the expected number ($m_T a$) under the assumption of random distribution. For example, in Drosophila the total number of nucleotides in mtDNA is about 14,000 (SHAH and LANGLEY 1979), excluding the A-T-rich region, and the G + C content is 0.22 (KAPLAN and LANGLEY 1979). Therefore, if we assume $g_G = g_C$ and $g_A = g_T$, the expected number of restriction sites for *Hae*III (*GGCC*) and *Hpa*II (*CCGG*) is $m_T a = 2.0$. SHAH and LANGLEY list the number of restriction sites for these enzymes for each nucleomorph. If we use one nucleomorph (the most common one) from each species, the mean number is 2.83 ± 0.48. Similarly, the expected number of restriction sites for *Eco*RI (*GAATTC*) or *Hind*III (*AAGCTT*) is 3.9; whereas, the observed number is 4.33 ± 0.61. Therefore, the observed and expected numbers agree with each other fairly well. We have done a similar but more extensive analysis of human data, which indicated that the agreement between the observed and expected numbers is also reasonably good. This result will be published elsewhere.

In our study we have implicitly assumed that all DNA fragments digested by restriction enzymes are identifiable by the experimental method used. In practice, small fragments are often disregarded because of experimental difficulty (BROWN, GEORGE and WILSON 1979). In these cases our theory should be modified to some extent. However, the effect of elimination of small fragments is generally unimportant, as will be discussed in a separate paper.

Another assumption we have made in relating nucleon diversity to nucleotide diversity is that the rate of nucleotide substitution or mutation rate is the same for all nucleotide sites. This assumption apparently does not hold in general. There is evidence that the rate of nucleotide substitution varies considerably from gene to gene in a genome and from region to region in a gene (*e.g.*, NEI 1975). However, when the number of nucleotide differences per nucleotide site is small, say less than 0.3, the assumption of equal rate does not lead to any serious error in the computation of average nucleotide differences (NEI and CHAKRABORTY 1976). In practice, the average number of nucleotide differences per site between two randomly chosen nucleomorphs seems to be generally smaller than 0.1 within a species (Table 3; see also NEI and LI 1979), so that the present method for analyzing polymorphism within populations will not be seriously affected by our assumption. However, when the number of nucleotide differences between a pair of species is large, our formula (31) tends to give an underestimate of $\delta = 2r\lambda t$ (see NEI and LI 1979).

The study of polymorphism of nuclear DNA by means of restriction enzymes has just started, and there are not enough data to test the validity of our theory. In nuclear DNA, however, deletion and insertion seem to be quite important. Therefore, DNA polymorphism should be studied in terms of nucleon diversity rather than in terms of nucleotide diversity, unless the nucleon studied refers only to the coding region.

This study was supported by Public Health Service grant GM 20293 and National Science Foundation grant DEB 76-06069.

APPENDIX

PROBABILITY OF PRODUCTION OF NEW RESTRICTION SITES BY MUTATION

Let g_1, g_2, g_3 and g_4 be the frequencies of nucleotides G, C, A and T in a nucleon consisting of a large number of nucleotides, respectively, and μ_{ij} be the mutation rate of the ith nucleotide to the jth nucleotide per generation. The total mutation rate of the ith nucleotide to the other nucleotides is then given by $\mu_i = \Sigma_{j=1}^{4} \mu_{ij}$ for $j \neq i$. Let $\mathbf{g}$ be the vector of nucleotide frequencies in a generation, *i.e.*, $\mathbf{g} = (g_1, g_2, g_3, g_4)$. If we assume no selection, the transition matrix for nucleotide state (i=1,2,3,4) is given by

$$\mathbf{P} = \begin{bmatrix} 1-\mu_1 & \mu_{12} & \mu_{13} & \mu_{14} \\ \mu_{21} & 1-\mu_2 & \mu_{23} & \mu_{24} \\ \mu_{31} & \mu_{32} & 1-\mu_3 & \mu_{34} \\ \mu_{41} & \mu_{42} & \mu_{43} & 1-\mu_4 \end{bmatrix}. \tag{A1}$$

This matrix is different from a similar matrix of KAPLAN and LANGLEY (1979), who assumed that the mutation rate is the same for all nucleotides, but once mutation occurs, each nucleotide changes to one of the other three nucleotides with a specific probability. Furthermore, they considered only the case where $g_1 = g_2$ and $g_3 = g_4$. Our matrix is more general than KAPLAN and LANGLEY's. At any rate, at steady state we have

$$\mathbf{g} = \mathbf{g}\mathbf{P}. \tag{A2}$$

Namely,

$$g_i = g_i\,(1 - \mu_i) + \sum_{\substack{j=1 \\ j \neq i}}^{4} g_j\,\mu_{ji}\ .$$

Therefore,

$$g_i\,\mu_i = \sum_{\substack{j=1 \\ j \neq i}}^{4} g_j\,\mu_{ji}\ . \tag{A3}$$

As in the text, we consider a restriction-site sequence of r nucleotides and denote the numbers of G, C, A and T in the restriction site by r_1, r_2, r_3 and r_4, respectively. In the computation of the probability of appearance of new restriction sites, we first note that in most cases new sites are formed from those DNA sequences in which one nucleotide is different from the restriction-site sequence, since the probability of double and triple mutations in a sequence of four to six nucleotides is extremely small. We therefore ignore the case where multiple mutations produce a restriction site. The expected frequency of a sequence of r nucleotides that differs from the restriction-site sequence by one nucleotide is given by

$$\begin{aligned} a_1 &= r_1 g_1^{r_1-1}\, g_2^{r_2+1}\, g_3^{r_3}\, g_4^{r_4} + r_1 g_1^{r_1-1}\, g_2^{r_2}\, g_3^{r_3+1}\, g_4^{r_4} \\ &\quad + r_1 g_1^{r_1-1}\, g_2^{r_2}\, g_3^{r_3}\, g_4^{r_4+1} + r_2 g_1^{r_1+1}\, g_2^{r_2-1}\, g_3^{r_3}\, g_4^{r_4} \\ &\quad + r_2 g_1^{r_1}\, g_2^{r_2-1}\, g_3^{r_3+1}\, g_4^{r_4} + \cdots + r_4 g_1^{r_1}\, g_2^{r_2}\, g_3^{r_3+1}\, g_4^{r_4-1} \\ &= a \sum_{i=1}^{4} \frac{r_i}{g_i} \sum_{\substack{j=1 \\ j \neq i}}^{4} g_j\ , \end{aligned}$$

where a is given by (1). Since there are m_T possible sequences of r nucleotides in a circular DNA of m_T nucleotides, the total number of such sequences is $m_T a_1$. Therefore, the probability of appearance of new sites is

$$\begin{aligned} & m_T a \sum_{i=1}^{4} \frac{r_i}{g_i} \sum_{\substack{j=1 \\ j \neq i}}^{4} g_j\ \mu_{ji} \\ &= m_T a \sum_{i=1}^{4} \frac{r_i}{g_i}\, g_i\, \mu_i && \text{(from (A3))} \\ &= m_T a \sum_{i=1}^{4} r_i\, \mu_i \\ &= m_T a r \mu. && \text{(from (2))} \end{aligned}$$

This is identical with the probability of loss of restriction sites.

LITERATURE CITED

AVISE, J. C., R. A. LANSMAN and R. O. SHADE, 1979 The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. I. Population structure and evolution in the genus Peromyscus. Genetics **92**: 279–295.

BOTSTEIN, D., R. L. WHITE, M. SKOLNICK and R. W. DAVIS, 1980 Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Amer. J. Hum. Genet. **32**: 314–331.

BROWN, W. M., 1976 Animal mitochondrial DNA. I. Intraspecific comparisons. II. The fidelity of replication of mitochondrial DNA fragments grown as recombinant plasmids in *Escherichia coli.* Ph.D. thesis. California Institute of Technology, Pasadena, California. ——, 1980 Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. Proc. Natl. Acad. Sci. U.S. **77**: 3605–3609.

BROWN, W. M., M. GEORGE, JR. and A. C. WILSON, 1979 Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. U.S. **76**: 1967–1971.

EFSTRATIADIS, A., J. W. POSAKONY, T. MANIATIS, R. M. LAWN, C. O'CONNELL, R. A. SPRITZ, J. K. DERIEL, B. G. FORGET, S. M. WEISSMAN, J. L. SLIGHTOM, A. E. BLECHL, O. SMITHIES, F. E. BARALLE, C. C. SHOULDERS and N. J. PROUDFOOT, 1980 The structure and evolution of the human β-globin gene family. Cell **21**: 653–668.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theoret. Popul. Biol. **3**: 87–112.

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.

GRIFFITHS, R. C., 1980 Genetic identity between populations when mutation rates vary within and across loci. J. Math. Biol. (in press).

KAN, Y. W. and A. M. DOZY, 1978 Polymorphism of DNA sequence adjacent to human β-globin structural gene: relationship to sickle mutation. Proc. Natl. Acad. Sci. U.S. **75**: 5631–5635.

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

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**: 893–903.

LEVINGS, C. S. and D. R. PRING, 1977 Diversity of mitochondrial genomes among normal cytoplasms of maize. J. Hered. **68**: 350–354.

LI. W.-H., 1977 Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. Genetics **85**: 331–337.

LI, W.-H. and M. NEI, 1975 Drift variances of heterozygosity and genetic distance in transient states. Genet. Res. **25**: 229–248.

NEI, M., 1972 Genetic distance between populations. Am. Naturalist **106**: 283–292. ——, 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. U.S. **70**: 3321–3323. ——, 1975 *Molecular Population Genetics and Evolution.* North Holland, Amsterdam and 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. U.S. **76**: 5269–5273.

SHAH, D. M. and C. H. LANGLEY, 1979 Inter- and intraspecific variation in restriction maps of *Drosophila* mitochondrial DNAs. Nature **281**: 696–699.

STEWART, F. M., 1976 Variability in the amount of heterozygosity maintained by neutral mutations. Theoret. Popul. Biol. **9**: 188–201.

UPHOLT, W. B., 1977 Estimation of DNA sequence divergence from comparison of restriction endonuclease digests. Nucleic Acids Research **4**: 1257–1265.

UPHOLT, W. B. and I. B. DAWID, 1977 Mapping of mitochondrial DNA of individual sheep and goats: rapid evolution in the D loop region. Cell **11**: 571–583.

WATTERSON, G. A., 1974 Models for the logarithmic species abundance distributions. Theoret. Popul. Biol. **6**: 217–250. ——, 1975 On the number of segregating sites in genetical models without recombination. Theoret. Popul. Biol. **7**: 256–276.

ZOUROS, E., 1979 Mutation rates, population sizes and amounts of electrophoretic variation of enzyme loci in natural populations. Genetics **92**: 623–646.

Corresponding editor: B. S. WEIR