Proc. Natl. Acad. Sci. USA
Vol. 76, No. 10, pp. 5269–5273, October 1979
Genetics

Mathematical model for studying genetic variation in terms of restriction endonucleases

(molecular evolution/mitochondrial DNA/nucleotide diversity)

MASATOSHI NEI AND WEN-HSIUNG LI

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

Communicated by Motoo Kimura, August 1, 1979

ABSTRACT **A mathematical model for the evolutionary change of restriction sites in mitochondrial DNA is developed. Formulas based on this model are presented for estimating the number of nucleotide substitutions between two populations or species. To express the degree of polymorphism in a population at the nucleotide level, a measure called "nucleotide diversity" is proposed.**

In recent years a number of authors have studied the genetic variation in mitochondrial DNA (mtDNA) within and between species by using restriction endonucleases (1–6). An important finding from these studies is that mtDNA has a high rate of nucleotide substitution compared with nuclear DNA, and thus it is suited for studying the genetic divergence of closely related species (5–7). However, the mathematical theory for analyzing data from restriction enzyme studies is not well developed. To our knowledge, the only study is that of Upholt (8).

A restriction endonuclease recognizes a specific sequence of nucleotide pairs, generally four or six pairs in length, and cleaves it. Therefore, if a circular DNA such as mtDNA has m such recognition (restriction) sites, it is fragmented into m segments after digestion by this enzyme. The number and locations of restriction sites vary with nucleotide sequence. The higher the similarity of the two DNA sequences compared, the closer the cleavage patterns. Therefore, it is possible to estimate the number of nucleotide substitutions between two homologous DNAs by comparing the locations of restriction sites. Similarly, the number of nucleotide substitutions may be estimated from the proportion of DNA fragments that are common to two organisms. Upholt (8) studied these two problems, but his formulation is not general and seems to involve some errors. Furthermore, Upholt paid no attention to the apparently high degree of heterogeneity of DNA sequences within populations (5). When the genetic divergence between closely related species is to be studied, it is necessary to eliminate the effect of this heterogeneity.

The purpose of this paper is to develop a more rigorous mathematical model of genetic divergence of DNA and present a statistical method for analyzing data from restriction enzyme studies. In the first four sections we shall either assume that there is no polymorphism within populations or consider the genetic divergence between a pair of organisms (individuals) only. The assumption of no polymorphism will be removed in the fifth section.

Evolutionary change of restriction sites

Under certain circumstances it is possible to map restriction sites in DNA. Once these restriction sites are determined for two different organisms, the proportion of sites shared by them can be computed. This proportion is expected to decline as the organisms' DNA sequences diverge. Before studying this problem, however, we consider the evolutionary change of restriction sites in a single population.

Consider a mtDNA of m_T nucleotide pairs with a G+C content of g. We note that in many vertebrate species m_T is about 16,000. If all nucleotides are randomly distributed in the DNA sequence, the expected frequency of restriction sites with r nucleotide pairs is

$$a = (g/2)^{r_1}[(1-g)/2]^{r_2}, \qquad [1]$$

in which r_1 and r_2 are the number of guanines (G) plus cytosines (C) and the number of adenines (A) plus thymines (T) in the restriction site, respectively, and $r_1 + r_2 = r$. (We consider only those restriction enzymes that recognize a unique sequence.) For example, if $g = 0.44$ and $m_T = 16{,}000$, the expected frequency of restriction site G-A-A-T-T-C (*Eco*RI) is 0.0003 and the expected total number (n) of restriction sites is $m_T a = 4.8$. Because a is generally small and m_T is large, n follows the Poisson distribution with mean $m_T a$.

We now study the evolutionary change of the number of restriction sites in mtDNA. Let $n(t)$ be the number of restriction sites at time t and $n(0) = n_0$. We make two assumptions: (*i*) The expected G+C content stays constant and (*ii*) nucleotide substitution occurs randomly and follows the Poisson process with a rate of substitution of λ per unit time (year or generation). We note that as time goes on the original sites will gradually disappear while new sites will be formed. Thus, $n(t)$ can be written as $n_1(t) + n_2(t)$, in which $n_1(t)$ denotes the number of original sites that remain unchanged and $n_2(t)$ that of new sites. Occasionally new sites may be formed at a position where the restriction site sequence once existed but disappeared by mutation. These new sites are included in $n_2(t)$ rather than in $n_1(t)$. Under our assumptions the probability that an original restriction site remains unchanged by time t is $P = e^{-r\lambda t}$. Therefore, the expectation of $n_1(t)$ is $n_0 e^{-r\lambda t}$. The expectation of $n_2(t)$ can be obtained in the following way. Consider a randomly chosen sequence of r nucleotide pairs. The probability that this sequence has undergone one or more nucleotide substitutions by time t is $1 - P$. We assume that nucleotide substitution produces a new random sequence of nucleotides. Then, the probability that a new restriction site is formed at this position is $a(1-P)$. Because there are m_T possible sequences in the entire DNA, the expected value of $n_2(t)$ is $m_T a(1-P)$. This formula can also be derived by a more rigorous but tedious method. At any rate, the expectation $[E(n)]$ of $n(t)$ becomes

$$E(n) = n_0 P + m_T a(1-P). \qquad [2]$$

As expected, $E(n)$ stays constant if $n_0 = m_T a$.

The variance $[V(n)]$ of $n(t)$ is obtained by noting that n_1 is binomially distributed, whereas n_2 follows the Poisson distribution. Because n_1 and n_2 are independent, we have

$$V(n) = n_0 P(1-P) + m_T a(1-P). \qquad [3]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

In the above formulation we have regarded the original restriction sites restored by backward mutations as new sites. For our purpose, however, it is better to regard them as identical with the original sites. In this case we need a slightly different formulation. We first consider the probability (p_t) that the nucleotide at a particular site at time t is the same as that of $t = 0$. If we assume that the mutation rate is the same for all directions among the four nucleotides, the recurrence formula for p_t is given by

$$p_{t+1} = (1 - \lambda)p_t + \tfrac{1}{3}\lambda(1 - p_t). \quad [4]$$

The continuous time solution of this equation with the initial condition $p_0 = 1$ gives

$$p_t = (1 + 3e^{-4\lambda t/3})/4. \quad [5]$$

For a restriction site to exist at the original position, all of the r nucleotides must be identical with the original ones. Thus, the probability that a restriction site exists at the original position at time t is $P = p_t^r$. The mean and variance of n_1 are then given by n_0P and $n_0P(1 - P)$, respectively, with the newly defined P. In practice, however, $P = p_t^r$ is close to $e^{-r\lambda t}$ unless λt is larger than about 0.15. On the other hand, n_2 again follows the Poisson distribution with the mean and variance of $(m_T - n_0)a(1 - P)/(1 - a)$.

DNA divergence between two populations

Let us now consider DNA divergence between two evolutionary lineages or populations X and Y. We assume that the mtDNAs in the two populations were derived from a common ancestral DNA sequence at time 0. Let n_{X1} and n_{X2} be the number of ancestral restriction sites and the number of new sites in population X, respectively, with $n_X = n_{X1} + n_{X2}$, and let n_{Y1}, n_{Y2}, and n_Y be the corresponding values in population Y. We denote the number of identical sites shared by the two populations by n_{XY}. We assume that all identical sites are those that remain unchanged from the common ancestor. Theoretically, new mutations may produce identical sites, but the contribution of new mutations is not so important unless λt is large, as will be discussed elsewhere. At any rate, under the present assumption n_{XY} follows a binomial distribution, and the mean and variance of n_{XY} are given by n_0P^2 and $n_0P^2(1 - P^2)$, respectively, in which P is either $e^{-r\lambda t}$ or the rth power of p_t in Eq. 5.

On the other hand, the proportion of ancestral restriction sites that remain unchanged in both lines is $S = n_{XY}/n_0$. The mean and variance of S are given by

$$\bar{S} = P^2, \quad [6]$$

$$V(S) = P^2(1 - P^2)/n_0. \quad [7]$$

Therefore, if we use $P = e^{-r\lambda t}$, the mean number of nucleotide substitutions per nucleotide site ($\delta = 2\lambda t$) is given by

$$\delta = -(\ln\bar{S})/r. \quad [8]$$

This relationship is identical with Upholt's (8). On the other hand, if we use the p_t given by Eq. 5, we have

$$\delta = -(3/2)\ln[(4\bar{S}^{1/2r} - 1)/3]. \quad [9]$$

To apply Eq. 8 or Eq. 9 to real data, $\bar{S}$ must be estimated. Brown *et al.* (6) used $n_{XY}/(n_X + n_Y - n_{XY})$ as an estimate of $\bar{S}$, but this gives an underestimate of $\bar{S}$. If n_0 is known, $\bar{S}$ may be estimated by n_{XY}/n_0. In practice, of course, it is not known. However, if we note $E(n_0) = E(n_X) = E(n_Y) \equiv E(n)$, in which $E(n_0)$ refers to the mean of replicate values of n_0, $(n_X + n_Y)/2$ may be used as an estimator of n_0. Therefore, $\bar{S}$ may be estimated by

$$\hat{S} = 2n_{XY}/(x_X + n_Y). \quad [10]$$

Although it is not clear from their description, Upholt and Dawid (2) seem to have used this formula.

We now investigate the statistical properties of this estimator. Using the Taylor expansion and neglecting the third- and higher-order terms, we obtain

$$E(\hat{S}) = \frac{2E(n_{XY})}{E(n_X) + E(n_Y)} - \frac{2\mathrm{Cov}(n_{XY},n_X + n_Y)}{[E(n_X) + E(n_Y)]^2} + \frac{2E(n_{XY})V(n_X + n_Y)}{[E(n_X) + E(n_Y)]^3},$$

approximately. Because n_X and n_Y change independently, $V(n_X + n_Y) = 2V(n)$. We also note that $\mathrm{Cov}(n_{XY},n_X + n_Y) = 2\mathrm{Cov}(n_{XY},n_X)$. Furthermore, $V(n) = E(n)(1 - P^2)$ if we note $E(n_0) = m_Ta$ in Eq. 3. It can also be shown that $\mathrm{Cov}(n_{XY},n_X) = n_0P^2(1 - P)$, which is $E(n)P^2(1 - P)$ when $n_0 = E(n)$. Therefore,

$$E(\hat{S}) = P^2 - P^2(1 - P)^2/[2E(n)]. \quad [11]$$

This indicates that $\hat{S}$ is an underestimate of P^2 but the bias is generally small when $E(n)$ is fairly large.

The approximate variance of $\hat{S}$ can be obtained in the same way. If we replace P^2 by $\hat{S}$ and $E(n)$ by $\bar{n} = (n_X + n_Y)/2$ in the variance obtained, it becomes

$$V(\hat{S}) = [\hat{S}(1 - \hat{S}) - \hat{S}^2(1 - \hat{S}^{1/2})^2]/\bar{n}. \quad [12]$$

This formula may be used for estimating the variance of $\hat{S}$ from data. In practice, the second term in the brackets of Eq. 12 is generally small compared with the first term.

Because $\bar{S}$ can be estimated by $\hat{S}$, the estimate ($\hat{\delta}$) of δ may be obtained by replacing $\bar{S}$ in Eq. 8 or Eq. 9 by $\hat{S}$. The large-sample variance of $\hat{\delta}$ obtained by Eq. 8 is given by

$$V(\hat{\delta}) = [d\hat{\delta}/d\hat{S}]^2V(\hat{S}) = V(\hat{S})/(r\hat{S})^2, \quad [13]$$

approximately, in which $V(\hat{S})$ is given by Eq. 12. On the other hand, the variance of $\hat{\delta}$ obtained by Eq. 9 is

$$V(\hat{\delta}) = [81\hat{S}^{1/r}V(\hat{S})]/[(4\hat{S}^{1/2r} - 1)r\hat{S}]^2. \quad [14]$$

The above two formulas indicate that the variance of $\hat{S}$ is large when $\bar{n}$ is small. Therefore, it is important to increase the reliability of $\hat{S}$ by using many different restriction enzymes. When enzymes with the same r value are used, we can add each of n_X, n_Y, and n_{XY} for all enzymes and then compute $\hat{\delta}$ and $V(\hat{\delta})$. However, when enzymes with different r values are used, $\hat{\delta}$ should be estimated for each r group and then the average weighted with the reciprocals of variances should be computed.

In the derivation of Eqs. 8 and 9 we have assumed that the rate of nucleotide substitution (λ) is constant over time. However, these formulas hold regardless of this assumption, provided nucleotide substitution occurs at random. However, if the rate is constant, δ is linearly related with the time (t) after divergence between the populations, i.e., $\delta = 2\lambda t$, and thus can be used for estimating t when λ is known.

The above formulation depends on the assumption that the probability of nucleotide substitution is the same for all nucleotide sites. In the case of mtDNA this assumption does not seem to be satisfied. Indeed, data from DNA hybridization experiments suggest that the rate of nucleotide substitution greatly varies among sites (6, 7). Uzzell and Corbin (9) have shown that in the cytochrome *c* gene the number of nucleotide substitutions per nucleotide site follows the negative binomial distribution when synonymous codons are disregarded. This suggests that the rate of nucleotide substitution per site follows the gamma distribution. If we assume that the same distribution applies to mtDNA, we can evaluate the effect of variation of

substitution rate (λ) on the estimate of nucleotide substitutions. In the following we assume that λ is constant over evolutionary time but varies among restriction sites following the gamma distribution

$$f(\lambda) = [\beta^\alpha/\Gamma(\alpha)]e^{-\beta\lambda}\lambda^{\alpha-1},$$

in which $\alpha = \bar{\lambda}^2/V_\lambda$, $\beta = \bar{\lambda}/V_\lambda$, in which $\bar{\lambda}$ and V_λ are the mean and variance of λ, respectively. If we use $P = e^{-r\lambda t}$, the mean of $\bar{S}$ in Eq. **6** becomes

$$E(\bar{S}) = \int_0^\infty e^{-2r\lambda t}f(\lambda)d\lambda = \left[\frac{\alpha}{\alpha + 2r\bar{\lambda}t}\right]^\alpha. \quad [15]$$

At the present time the value of α is not known, but probably $\alpha \geq 1$ in most cases. In the cytochrome *c* gene α has been estimated to be about 2. It is noted that when $\alpha \geq 1$ the difference between Eqs. **8** and **15** is small as long as S is larger than 0.7 but increases as S declines further (10). If α is known, the average number of nucleotide substitutions ($\delta = 2\bar{\lambda}t$) should be estimated by using Eq. **15**. For example, if $\alpha = 2$,

$$\delta = (2/r)(1/\sqrt{S} - 1). \quad [16]$$

Evolutionary change of DNA fragments

The current experimental method of comparing restriction-site maps is laborious and may not be suited for a large-scale population survey. A simpler method is to compare the electrophoretic patterns of DNA digested by a restriction endonuclease between the two species or populations in question. The degree of genetic divergence of DNA between the two populations is expected to be correlated with the proportion of DNA fragments shared by them. Let us now study the relationship between these two quantities.

For a given DNA fragment to be conserved in the evolutionary process, two conditions must be met, as noted by Upholt (8). (*i*) Two external restriction sites remain unchanged, and (*ii*) no new restriction sites occur within the fragment. The probability of the first event is obviously P^2. The probability of the second event can be obtained in the following way. Let m be the number of nucleotides in this fragment. Then there are $m - r + 1$ possible sequences of r nucleotides between the two external restriction sites. As shown before, the probability for a randomly chosen r-base sequence to become a new restriction site by time t is $b = a[1 - P]$. Thus, the probability that no new sites are formed in this fragment by time t is $(1 - b)^{m-r+1}$, and the probability that this fragment remains unchanged in both populations is $P^4(1 - b)^{2(m-r+1)}$. Because there are n_0 fragments originally, the proportion of fragments shared by the two populations is

$$F = (1/n_0)\sum_{i=1}^{n_0} P^4(1 - b)^{2(m_i-r+1)}, \quad [17]$$

in which m_i is the number of nucleotide sites in the ith fragment.

In practice, the above formula is not applicable, because n_0 and m_i are not known. However, it is possible to compute the probability of formation of a fragment of m nucleotides under the assumption of random nucleotide distribution. It is given by $a(1 - a)^{m-r}/T$, in which T is the normalizing factor and given by

$$T = \sum_{m=r}^{m_T} a(1 - a)^{m-r} = 1 - (1 - a)^{m_T-r+1}.$$

The expected proportion of fragments that remain unchanged in both populations at time t is then given by

$$F = \sum_{m=r}^{m_T} P^4(1 - b)^{2(m-r+1)}a(1 - a)^{m-r}/T. \quad [18]$$

Assuming that $(m_T - r + 1)a$ is so large that T is close to 1, we obtain

$$F \approx a(1 - b)P^4/[a(1 - b)^2 + b(2 - b)]. \quad [19]$$

This formula is different from Upholt's. Because a is usually much smaller than 1 and $b = a[1 - P]$, the above formula can be approximated by

$$F \approx P^4/(3 - 2P). \quad [20]$$

Using $P = e^{-r\lambda t}$ and $\delta = 2\lambda t$, F can be related to δ. The relationship between δ and F is shown in Fig. 1 for $r = 4$ and 6. This relationship may be used for estimating δ from F.

To estimate F, we propose the following estimator.

$$\hat{F} = 2n_{XY}/(n_X + n_Y), \quad [21]$$

in which n_X and n_Y are the numbers of fragments in populations X and Y, respectively, whereas n_{XY} is the number of fragments shared by the two populations.

In the above formulation, we have not considered back mutation. This is justified because the "fragment" method can be used only when δ is relatively small.

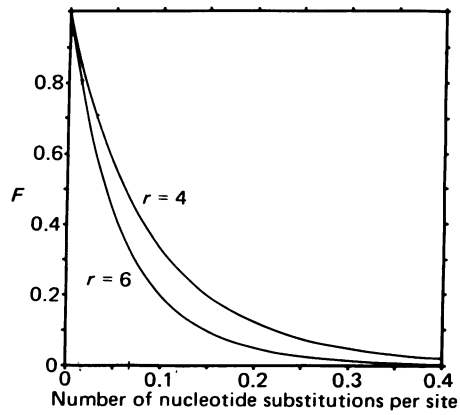

FIG. 1. Relationship between the proportion of shared DNA fragments (F) and the number of nucleotide substitutions per site (δ).

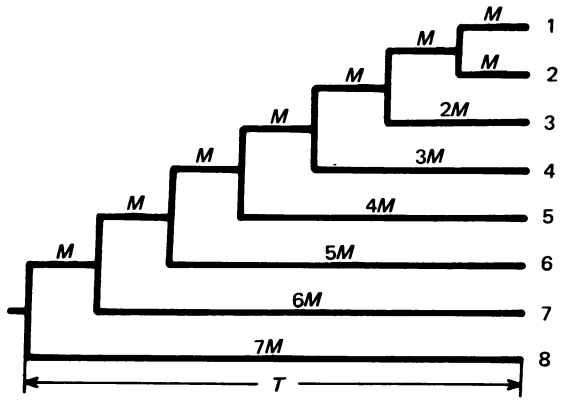

FIG. 2. Evolutionary tree used in the computer simulation. Numbers 1, 2, . . . , 8 represent descendant DNA sequences. M is the expected number of nucleotide substitutions for the shortest branch. In the present simulation M was 8 per 300 nucleotide sites or 100 codons.

Table 1. Number of shared restriction sites and shared DNA fragments between DNA sequences in a computer simulation

Sequence	1	2	3	4	5	6	7	8
1	(105)	86	66	46	46	39	30	29
2	68	(101)	70	51	45	40	28	27
3	34	33	(107)	51	47	39	23	30
4	18	18	20	(100)	44	41	30	28
5	12	14	11	8	(107)	38	30	31
6	13	13	10	10	8	(114)	33	28
7	4	3	3	5	3	7	(115)	24
8	3	2	2	2	4	3	2	(108)

The eight DNA sequences represent those given in Fig. 2. Figures above the diagonal are the numbers of shared restriction sites, whereas those below the diagonal are the number of shared DNA fragments. Figures on the diagonal refer to numbers of restriction sites for each descendant sequence.

Computer simulation

In order to see the accuracy of the theory developed we have done a computer simulation. In practice, we used artificial nucleotide sequences generated in the work of Y. Tateno and M. Nei (unpublished) on molecular taxonomy. In this study a hypothetical sequence of 6000 nucleotide pairs in a circular form was used. An ancestral sequence of random nucleotides was generated by using pseudorandom numbers with an expected G+C content of 0.5, and from this sequence eight descendant sequences were produced following the evolutionary tree given in Fig. 2. The number of nucleotide substitutions for each branch in this figure followed the Poisson distribution with the mean given along the branch (per 300 nucleotide sites or 100 codons). After generating the eight descendant sequences, we determined the locations of restriction sites for five different hypothetical endonucleases in all of them. Each restriction enzyme was assumed to recognize a particular sequence of four base pairs.

Identity of Restriction Sites. The total number of restriction sites for the five "enzymes" in each descendant sequence is given in Table 1 together with the number of sites shared by each pair of sequences. Using these data, we estimated S and δ. The results obtained are presented in Table 2. When two or more sequence comparisons have the same δ value (e.g., 1–3 vs. 4), the average of δs for all comparisons are presented. The δ value was estimated by Eqs. 8 and 9; the estimate of δ obtained by Eq. 8 is designated by $\hat{\delta}_1$ and that obtained by Eq. 9 by $\hat{\delta}_2$.

Table 1 shows that n is 100 to 115. These values are somewhat smaller than the expected value of $5 \times 23.4 = 117$, but the differences are not statistically significant because the expected standard deviation is 10.8. The values of $\hat{\delta}_1$ and $\hat{\delta}_2$ are also not far from the expected value of δ if we consider the large stochastic error to which they are subject. Theoretically, $\hat{\delta}_2$ is a better estimate than $\hat{\delta}_1$ as mentioned earlier, but in practice there is not much difference between the two estimates. In the comparison of 1–7 vs. 8, $\hat{\delta}_2$ (and also $\hat{\delta}_1$) is somewhat smaller than the expected value. This smaller value occured largely because the proportion of identical sites was affected appreciably by new mutation in this case. Indeed, when we disregarded the identical sites due to new mutation, the $\hat{\delta}_2$ value was 0.378, which is close to the expected value of 0.373. The effect of mutation was observed also in the case of smaller δ values, but it was not so serious as in the case of $\delta = 0.373$.

One important finding in the present simulation is that the estimate of δ is subject to a large stochastic error when n_X, n_Y, and n_{XY} are small. For example, when only one type of "restriction enzyme" is used, $E(n)$ is 23.4. In this case the $\hat{\delta}_2$ value for 6 vs. 8 took the values of 0.452, 0.348, 0.253, 0.423, and 0.358 for the five different types of "restriction enzymes" used. Therefore, it is important to use a large number of restriction enzymes. Of course, the accuracy of $\hat{\delta}_2$ depends on the number of base pairs in the restriction site. The sampling error of $\hat{\delta}_2$ is expected to be smaller for $r = 6$ than for $r = 4$ when S is the same.

Identity of DNA Fragments. Using data on restriction-site maps in the eight descendant DNA sequences, we computed the number of identical DNA fragments that were shared by each pair of sequences (Table 1). We then estimated F and δ; the results are presented in Table 2. The estimate of δ obtained by this method is designated by $\hat{\delta}_3$. It is clear that $\hat{\delta}_3$ again roughly agrees with the expected value. In this case the effect of mutation on the estimate of δ is not so large as in the case of "identical sites" method, because the probability of formation of identical fragments by mutation is smaller than that of formation of identical restriction sites. However, the sampling error of $\hat{\delta}_3$ is generally larger than that of $\hat{\delta}_1$ or $\hat{\delta}_2$.

Table 2. Estimates ($\hat{\delta}_1$, $\hat{\delta}_2$, $\hat{\delta}_3$) of the number of nucleotide substitutions in comparison with the expected numbers (δ)

Sequence comparison	Restriction sites			DNA fragments		
	$\hat{S}$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{F}$	$\hat{\delta}_3$	δ
1 vs. 2	0.835	0.045	0.045	0.660	0.036	0.053
1–2 vs. 3	0.648	0.109	0.109	0.319	0.103	0.107
1–3 vs. 4	0.483	0.182	0.185	0.183	0.158	0.160
1–4 vs. 5	0.433	0.209	0.213	0.107	0.213	0.213
1–5 vs. 6	0.362	0.254	0.260	0.099	0.222	0.267
1–6 vs. 7	0.263	0.334	0.344	0.038	0.324	0.320
1–7 vs. 8	0.262	0.335	0.345	0.024	0.376	0.373

$\hat{\delta}_1$, $\hat{\delta}_2$, and $\hat{\delta}_3$ were obtained through Eqs. 8, 9, and 20, respectively. When two or more sequence comparisons have the same δ value, the averages of the estimates are presented. Similarly, $\hat{S}$ and $\hat{F}$ are the averages for all comparisons having the same δ value. Therefore, $\hat{\delta}_1$, $\hat{\delta}_2$, and $\hat{\delta}_3$ are not directly obtainable from the $\hat{S}$ and $\hat{F}$ values presented except in the comparison of 1 vs. 2. These results were obtained by computer simulation.

Intrapopulational variation

In population genetics it is customary to measure the genic variation of a population in terms of heterozygosity or gene diversity (11). In the case of mtDNA, however, this measure is not appropriate, because mtDNA contains many genes and thus the gene diversity would be close to 1 in many populations. In this case genic variation may be measured more appropriately by the average number of nucleotide differences per site between two randomly chosen DNA sequences. We call this the *index of nucleotide diversity* or simply *nucleotide diversity*, and denote it by π. It is defined as

$$\pi = \sum_{ij} x_i x_j \pi_{ij}, \quad [22]$$

in which x_i is the frequency of the ith sequence in the popu tion and π_{ij} is the number of nucleotide differences per ... cleotide site between the ith and jth sequences.

The nucleotide diversity may be estimated from restriction enzyme data if we know x_i and π_{ij}. The value of π_{ij} can be estimated either from $\hat{S}$ or from $\hat{F}$ as mentioned above. When data on restriction-site maps are available, it is also possible to compute the average proportion of shared sites between two randomly chosen DNA sequences. It is given by

$$\hat{S} = \sum x_i x_j \hat{S}_{ij}. \quad [23]$$

This will give another estimate of π. That is,

$$\hat{\pi} = (-\ln \hat{S})/r. \quad [24]$$

In the preceding sections we presented formulas for estimating the number of nucleotide substitutions between two populations under the assumption that the effect of intrapopulational variation is negligible. When the populations to be compared are closely related, this assumption will not generally be satisfied. In this case the intrapopulational variation should be subtracted from the total interpopulational difference.

Let x_i and y_i be the frequencies of the ith restriction-site sequence in populations X and Y, respectively. Then, the π values for populations X and Y may be estimated by $\hat{\pi}_X = \sum_{ij} x_i x_j \hat{\pi}_{ij}$ and $\hat{\pi}_Y = \sum_{ij} y_i y_j \hat{\pi}_{ij}$, respectively, whereas the average number of nucleotide differences between two randomly chosen DNA sequences, one from each of X and Y, may be estimated by $\hat{\pi}_{XY} = \sum_{ij} x_i y_j \hat{\pi}_{ij}$. Therefore, the estimate of net nucleotide differences between the two populations is given by

$$\hat{\delta} = \hat{\pi}_{XY} - (\hat{\pi}_X + \hat{\pi}_Y)/2. \quad [25]$$

As mentioned earlier, $\hat{\pi}_{ij}$ may be obtained either from $\hat{S}$ or $\hat{F}$. Another way of estimating $\hat{\delta}$ is to use the normalized proportion of shared sites between X and Y. It is defined as

$$S = S_{XY}/\sqrt{S_X S_Y}, \quad [26]$$

in which $S_X = \Sigma_{ij} x_i x_j S_{ij}$, $S_Y = \Sigma_{ij} y_i y_j S_{ij}$, and $S_{XY} = \Sigma_{ij} x_i y_j S_{ij}$. The δ value is then given by Eq. 8. This method is analogous to that of estimating genetic distance from gene frequency data (11).

Discussion

The theory developed in this paper is dependent on the assumption that all nucleotides are distributed at random over the DNA sequences with a given G+C content. Available data suggest that this assumption is not always satisfied. Brown (12) has shown that the contents of thymines and guanines in the heavy strand of mtDNA are considerably different from those of the light strand in man, green monkey, and mouse. However, because we are concerned with the evolutionary change of mtDNA, the nonrandom distribution would not affect our estimate of nucleotide substitutions seriously unless it is extreme.

At the present time the magnitude of nucleotide diversity (π) in natural populations is not well known. The *Peromyscus polinotus* data of Avise *et al.* (5) suggest that it is of the order of 0.01, whereas in man it seems to be of the order of 0.002 (6). This quantity is expected to vary from population to population even in the same species. Therefore, it is important to make correction for this factor in the estimation of the degree of nucleotide divergence between closely related species.

Theoretically, it is possible to express nucleotide diversity π in terms of the mutation rate per nucleotide site per host generation (μ) and the effective population size (13–15). In the case of mitochondria, which are maternally inherited, π is approximately given by $2N_m\mu$, in which N_m is the number of female adult individuals. We note that there is little genetic heterogeneity among mtDNAs of one host individual in mammals. On the other hand, the average heterozygosity for nuclear genes may be expressed as $H = 4N_n v/(4N_n v + 1)$, in which N_n is the effective population size for nuclear genes and equal to the number of both male and female individuals, and v is the mutation rate per gene. In *P. polinotus*, H has been estimated to be 0.08 for isozyme data (16). If we assume that an average structural gene consists of 1000 nucleotide pairs and only 1/10th of nucleotide variation in structural genes is detectable by electrophoresis, the average nucleotide difference per site between two randomly chosen nuclear genes becomes 0.0008. Therefore, it seems that mtDNA is much more variable than structural genes in nuclear DNA. This conclusion is different from Langley and Shah's (17) that they are almost equally variable in *Drosophila*.

We thank W. M. Brown and A. C. Wilson for their valuable comments on the manuscript. This study was supported by research grants from the National Science Foundation and the National Institutes of Health.

1. Potter, S. T., Newbold, J. E., Hutchison, C. A. & Edgell, M. H. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 4496–4500.
2. Upholt, W. B. & Dawid, I. B. (1977) *Cell* **11**, 571–583.
3. Levings, C. S. & Pring, D. R. (1977) *J. Hered.* **68**, 350–354.
4. Parker, R. C. & Watson, R. M. (1977) *Nucleic Acids Res.* **4**, 1291–1300.
5. Avise, J. C., Lansman, R. A. & Shade, R. O. (1979) *Genetics* **92**, 279–295.
6. Brown, W. M., George, M. & Wilson, A. C. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1967–1971.
7. Dawid, I. B. (1972) *Dev. Biol.* **29**, 139–151.
8. Upholt, W. B. (1977) *Nucleic Acids Res.* **4**, 1257–1265.
9. Uzzell, T. & Corbin, K. W. (1971) *Science* **172**, 1089–1096.
10. Nei, M. (1980) *Proceedings of the XIV International Congress of Genetics*, Moscow, U.S.S.R., in press.
11. Nei, M. (1975) *Molecular Population Genetics and Evolution* (North-Holland, Amsterdam).
12. Brown, W. M. (1976) Dissertation (California Inst. Tech., Pasadena, CA).
13. Kimura, M. (1969) *Genetics* **61**, 893–903.
14. Watterson, G. A. (1975) *Theor. Pop. Biol.* **7**, 256–276.
15. Li, W.-H. (1977) *Genetics* **85**, 331–337.
16. Selander, R. K., Smith, M. H., Yang, S. Y., Johnson, W. E. & Gentry, J. B. (1971) *Stud. Genet.* **6**, 49–90.
17. Langley, C. H. & Shah, D. M. (1979) *Nature (London)*, in press.