EVOLUTIONARY RELATIONSHIP OF DNA SEQUENCES IN FINITE POPULATIONS

FUMIO TAJIMA

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

Manuscript received February 11, 1983
Revised copy accepted June 22, 1983

ABSTRACT

With the aim of analyzing and interpreting data on DNA polymorphism obtained by DNA sequencing or restriction enzyme technique, a mathematical theory on the expected evolutionary relationship among DNA sequences (nucleons) sampled is developed under the assumption that the evolutionary change of nucleons is determined solely by mutation and random genetic drift. The statistical property of the number of nucleotide differences between randomly chosen nucleons and that of heterozygosity or nucleon diversity is investigated using this theory. These studies indicate that the estimates of the average number of nucleotide differences and nucleon diversity have a large variance, and a large part of this variance is due to stochastic factors. Therefore, increasing sample size does not help reduce the variance significantly. The distribution of sample allele (nucleomorph) frequencies is also studied, and it is shown that a small number of samples are sufficient in order to know the distribution pattern.

IN some groups of genes, such as mitochondrial DNA, recombination is negligible, and in this case it is possible to construct an evolutionary tree of alleles or nucleomorphs (NEI and TAJIMA 1981). These evolutionary trees indicate that the nucleomorphs sampled from different populations are often more similar in nucleotide sequence than some pairs of the nucleomorphs sampled from the same populations (BROWN 1980; NEI 1982; CANN, BROWN and WILSON 1982). There are three possible explanations for this observation. The first is natural selection, which has conserved the nucleotide sequence of a particular gene or gene set (nucleon) in both populations. The second is recent gene migration between populations. For example, American whites and blacks are expected to share some common nucleomorphs because of recent gene migration. The third is the stochastic error that is generated by random genetic drift. However, there is no theoretical study about the expected phylogenetic trees under these hypotheses.

The purpose of this paper is to present a mathematical theory on the expected genealogy of a group of nucleons sampled and the number of nucleotide differences among them when the evolutionary change of nucleons is determined solely by mutation and random genetic drift. I shall also investigate the expected distribution of sample nucleomorph frequencies.

ASSUMPTION

In this paper we consider a random mating population of N diploid individuals and assume that there is no gene migration from outside populations. We also assume that there is no selection and no recombination between DNA sequences. Following NEI and TAJIMA (1981), we call any segment of DNA under investigation a nucleon and a particular DNA sequence for the segment a nucleomorph. Nucleon and nucleomorph correspond to gene and allele in classical genetics.

EXPECTED EVOLUTIONARY RELATIONSHIP OF A SAMPLE OF NUCLEONS

Single populations

Topological relationship among nucleons: First, we consider the case in which nucleons are randomly sampled from a single population. When two nucleons are sampled, we have one common ancestral nucleon. Figure 1a shows this relationship. When three nucleons are sampled, we have two possible relationships. One is that a common ancestral nucleon bifurcates, and one of the branches again bifurcates. This relationship is shown in Figure 1b. Another relationship can be obtained when a common ancestral nucleon trifurcates. As will be shown later, the probability of the latter event is negligibly small unless population size is very small. Therefore, we assume that all of the branches are created by bifurcation. When four nucleons are sampled, there are two possible relationships as shown in Figure 2. The probabilities of getting relationships a and b in this figure can be obtained by using Figure 1b. If bifurcation takes place at point C or D, relationship a is obtained, whereas if bifurcation occurs at point E, relationship b is obtained. Since these three bifurcation events take place with the same probability, the probabilities of getting relationships a and b are $2/3$ and $1/3$, respectively. The probability of getting a particular relationship among five nucleons sampled, which is shown in Figure 3, can be obtained by the same method using the relationships among four nucleons.

In general, the probability of getting a particular type of relationship for n nucleons is given by

$$P = 2^{n-1-s}/(n-1)!, \tag{1}$$

where s is the number of branching points that lead to exactly two nucleon descendants in the sample. For example, in the relationship of Figure 4 branching points E, F, G and H lead to two nucleon descendants. Therefore, $s = 4$. Since $n = 9$, the probability of getting the relationship in this figure is $P = 2^{9-1-4}/(9-1)! = 1/2520$. Proof of (1) is given in APPENDIX I.

We are often interested only in topological relationship. For example, relationships c, d and e in Figure 3 can be regarded as the same topology. In this case the probability of getting this topology can be obtained by summing the probabilities of getting relationships c, d and e, *i.e.*, $1/6 + 1/6 + 1/6 = 1/2$. In general, the probability of getting a particular topology for nucleons sampled can be obtained by using the following probability. The probability that a

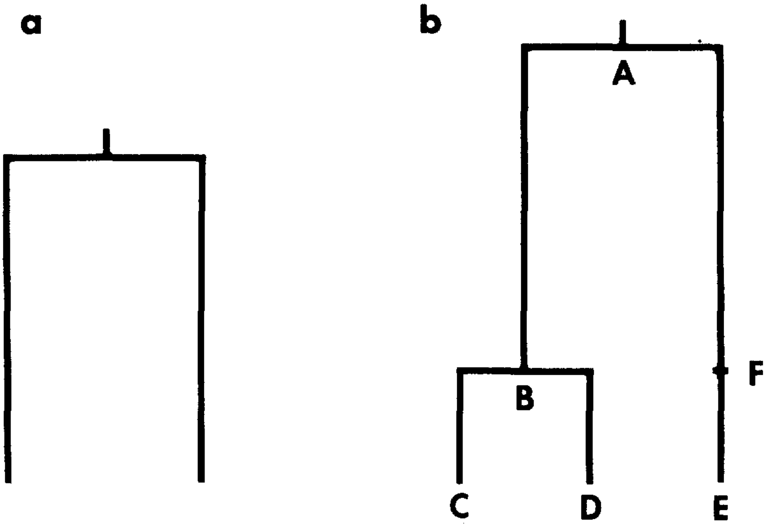

FIGURE 1.—Expected evolutionary relationships, (a) when two nucleons are sampled and (b) when three nucleons are sampled from a population.

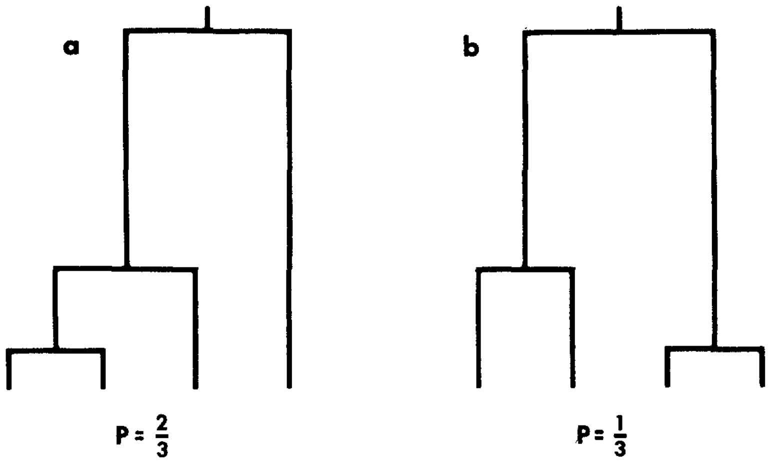

FIGURE 2.—Expected evolutionary relationships among four nucleons sampled from a population.

certain branching point divides n nucleons into n_1 and n_2 nucleons (order not important) is

$$\begin{aligned} P(n_1, n_2) &= 2/(n-1) \quad \text{if} \quad n_1 \neq n_2, \\ P(n_1, n_2) &= 1/(n-1) \quad \text{if} \quad n_1 = n_2, \end{aligned} \tag{2}$$

where $n = n_1 + n_2$. Proof of (2) is given in APPENDIX I. To show how to use (2), let us again use Figure 4. Point A in this figure divides nine nucleons into five and four nucleons. From (2) this probability is $2/(9-1) = 1/4$. Point B divides five nucleons into two and three nucleons. Therefore, this probability

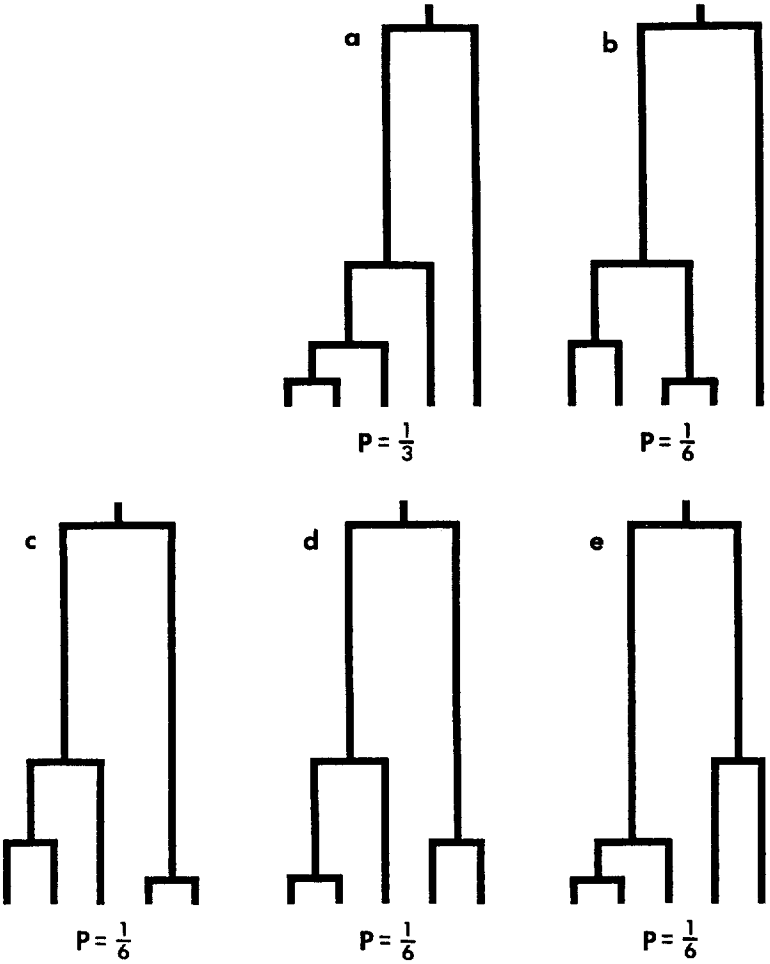

FIGURE 3.—Expected evolutionary relationships among five nucleons sampled from a population.

is $2/(5 - 1) = \frac{1}{2}$ from (2). The probability that point C divides four nucleons into the two groups of two nucleons each is $1/(4 - 1) = \frac{1}{3}$ from (2). Then the probability of getting the topology in Figure 4 is $\frac{1}{4} \times \frac{1}{2} \times \frac{1}{3} = \frac{1}{24}$.

One interesting point emerging from this study is that the probability of one nucleon being quite different from the others is not very low. For example, when we sample 20 nucleons, the probability that 20 nucleons are divided into one and 19 nucleons is $2/(20 - 1) = \frac{2}{19}$ from (2), which is not very low.

Branch length: Next, we consider the branch length of a nucleon genealogy. It is convenient to measure the branch length in terms of the number of generations. Let $f_n(t)$ be the probability that $n + 1$ nucleons randomly sampled from a population are derived from n nucleons t generations ago and the divergence took place $t - 1$ generations ago. Here, t is a random variable, and n is a fixed number.

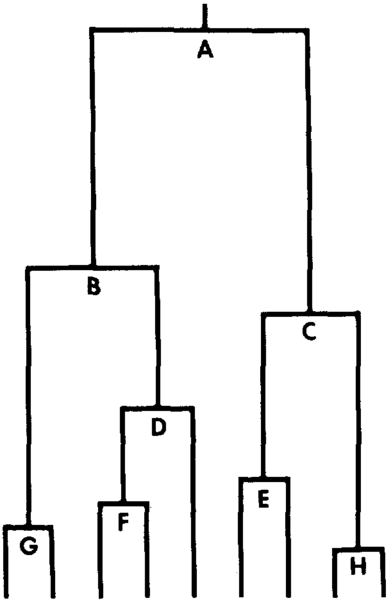

FIGURE 4.—One example of the topological relationship among nine nucleons sampled from a population.

We first consider the case in which two nucleons are sampled. The topological relationship is shown in Figure 1a. The probability that two nucleons are derived from a common ancestral nucleon in the immediately previous generation is

$$f_1(1) = 1/(2N),$$

where N is the number of diploid individuals in a population. Therefore, we have

$$\begin{aligned} f_1(t) &= f_1(1)\{1 - f_1(1)\}^{t-1} = \{1/(2N)\}\{1 - 1/(2N)\}^{t-1} \\ &\approx \{1/(2N)\}\exp\{-t/(2N)\}. \end{aligned} \tag{3}$$

This formula gives the probability distribution of branch length in terms of the number of generations (t). The mean $[E(t)]$ and variance $[V(t)]$ of the distribution are given by

$$E(t) = 2N, \qquad V(t) = 4N^2.$$

When three nucleons are sampled, we have shown that there is only one topological relationship as shown in Figure 1b. The branch length between A and B in this figure is the same as $f_1(t)$. Let us obtain the branch length between B and C, which is $f_2(t)$ by definition. The probability that three nucleons are derived from a common ancestral nucleon in the immediately previous generation is

$$1/(2N)^2 \approx 0.$$

This indicates that trifurcation is a very rare event. Therefore, we disregard it from consideration. The probability that three nucleons were derived from two nucleons in the immediately previous generation is

$$f_2(1) = \binom{3}{2}\{1/(2N)\}\{1 - 1/(2N)\} \approx 3/(2N),$$

whereas the probability that three nucleons are derived from three different nucleons in the immediately previous generation is

$$\{1 - 1/(2N)\}\{1 - 2/(2N)\} \approx 1 - 3/(2N).$$

From these probabilities we obtain

$$f_2(t) = \{3/(2N)\}\{1 - 3/(2N)\}^{t-1} \approx \{3/(2N)\}\exp\{-3t/(2N)\}. \tag{4}$$

This gives the probability distribution of branch length between B and C in Figure 1b. The mean and variance of branch length (t) are given by

$$E(t) = 2N/3, \qquad V(t) = 4N^2/9.$$

This indicates that the mean branch length between B and C is three times shorter than that between A and B.

Similarly, we can obtain $f_n(t)$, which becomes

$$f_n(t) = \{\tbinom{n+1}{2}/(2N)\}\{1 - \tbinom{n+1}{2}/(2N)\}^{t-1} \approx \{\tbinom{n+1}{2}/(2N)\}\exp\{-\tbinom{n+1}{2}t/(2N)\}. \tag{5}$$

Recently HUDSON (1983) has also obtained (5) and used it for simulating the evolution of proteins. The mean and variance of t are given by

$$E(t) = 2N/\tbinom{n+1}{2}, \tag{6}$$

$$V(t) = 4N^2/\tbinom{n+1}{2}^2. \tag{7}$$

The probability that n nucleons randomly sampled from a population are derived from m nucleons t generations ago ($m < n$) can be obtained from the convolution of $f_{n-1}(t), f_{n-2}(t), \ldots, f_m(t)$. The mean and variance of t are then given by

$$E(t) = 2N \sum_{i=m+1}^{n} \{1/\tbinom{i}{2}\} = 4N(1/m - 1/n), \tag{8}$$

$$V(t) = 4N^2 \sum_{i=m+1}^{n} \{1/\tbinom{i}{2}^2\}. \tag{9}$$

In the case of $m = 1$, namely, when n nucleons are derived from a common ancestral nucleon, the mean and variance of t are given by

$$E(t) = 4N(1 - 1/n), \tag{10}$$

$$V(t) = 4N^2 \sum_{i=2}^{n} \{1/\tbinom{i}{2}^2\}. \tag{11}$$

As n increases, $E(t)$ and $V(t)$ quickly approach $4N$ and $16N^2(\pi^2/3 - 3)$, respectively. Note that when $n = 2N$, these values are essentially the same as those of fixation time of a newly arisen neutral mutant (KIMURA and OHTA 1969; BURROWS and COCKERHAM 1974). In fact, we can obtain the mean fixation time when the initial frequency of mutant is $k/(2N)$ by using (8), which becomes

$$E\left(t\,|\,\frac{k}{2N}\right) = 4N\left\{1 - \frac{1}{2N} - \sum_{i=1}^{k-1}\frac{1}{i(i+1)}\prod_{j=1}^{i}\frac{k-j}{2N-j}\right\}. \tag{12}$$

As N increases, (12) approaches $E(t\,|\,p) = -4N(1/p - 1)\log_e(1 - p)$, where $p = k/(2N)$. This is identical with the formula obtained by KIMURA and OHTA (1969) using diffusion approximations.

Two populations

So far we have considered the case in which nucleons are sampled from a single population. We now consider the case in which nucleons are sampled from two populations. Let us assume that each population consists of N diploid individuals, and that these two populations have separated t generations ago. When four nucleons (two nucleons from each population) are sampled, there are four possible topological relationships as shown in Figure 5. The probability of getting a particular topological relationship can be obtained by applying the relationships obtained for single populations. For example, topology c can be obtained only when the following four conditions are satisfied: (1) Two nucleons sampled from one population do not have a common ancestral nucleon more recently than t generations ago. (2) Two nucleons sampled from another population also do not have a common ancestral nucleon more recently than t generations ago. Each of these probabilities is given by

$$1 - \sum_{i=1}^{t} f_1(i) \approx \exp\{-t/(2N)\}.$$

(3) We have topological relationship b in Figure 2. The probability of having this relationship is $1/3$. (4) Each of two groups consists of two nucleons that are sampled from different populations. This probability is $2/3$. Therefore, the probability of getting topological relationship c in Figure 5 is

$$[\exp\{-t/(2N)\}]^2 \times 1/3 \times 2/3 = (2/9)\exp(-t/N).$$

The other probabilities can be obtained in the same way, and they are given in Figure 5.

Figure 6 shows the relative probabilities of having the four different topological relationships in Figure 5. When the time of divergence between two populations is short, topological relationship d occurs with the highest probability, but as the divergence time becomes longer, the probability of getting relationship a becomes higher. Note that the probability of getting relationship a is not very high unless two populations separated a long time ago. For example, when two populations diverged $4N$ generations ago, this probability is still 0.83.

NUMBER OF NUCLEOTIDE DIFFERENCES BETWEEN RANDOMLY CHOSEN NUCLEONS

In this and the following sections we consider mutation and assume that the mutation rate is the same for all nucleotides.

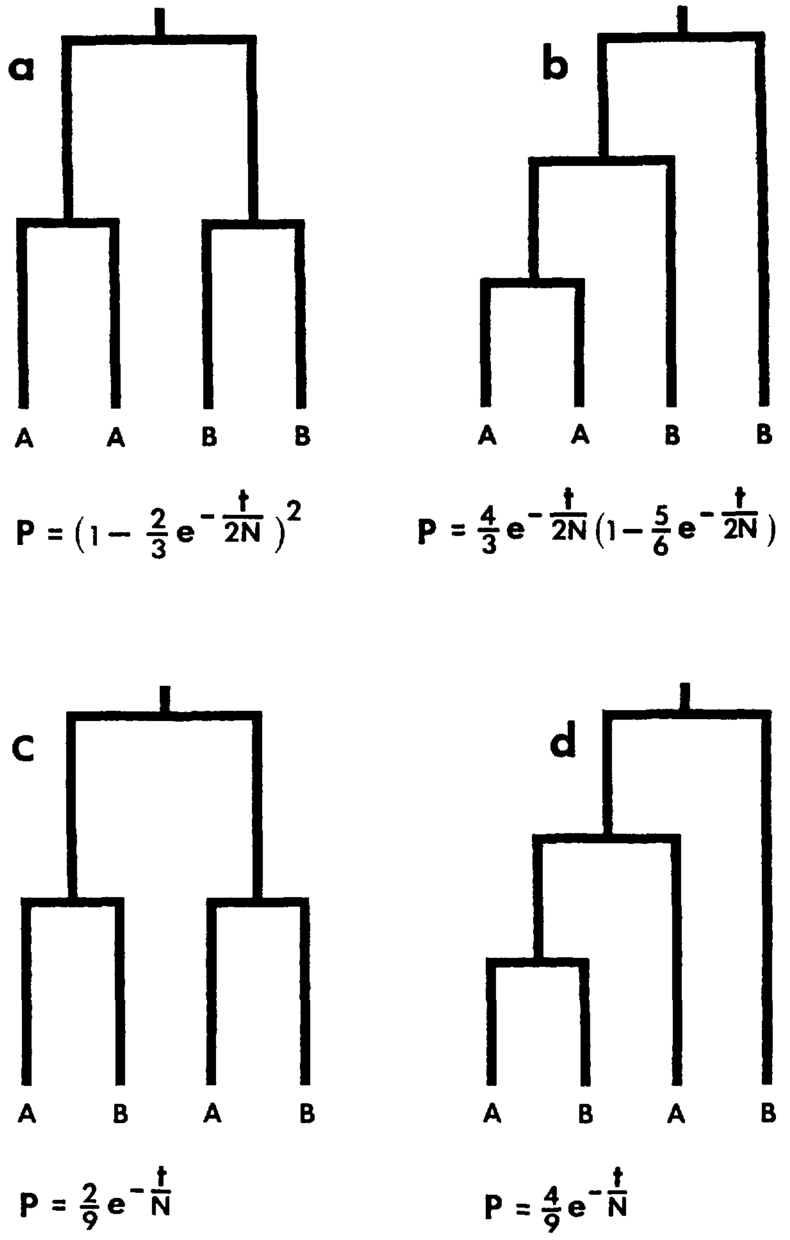
a
A A B B
$P = \left(1 - \frac{2}{3}e^{-\frac{t}{2N}}\right)^2$
b
A A B B
$P = \frac{4}{3}e^{-\frac{t}{2N}}\left(1 - \frac{5}{6}e^{-\frac{t}{2N}}\right)$
c
A B A B
$P = \frac{2}{9}e^{-\frac{t}{N}}$
d
A B A B
$P = \frac{4}{9}e^{-\frac{t}{N}}$

FIGURE 5.—Expected evolutionary relationships among four nucleons sampled from two populations, *e.g.*, populations A and B, which diverged t generations ago. It is assumed that two nucleons are sampled from population A and the other two nucleons are sampled from population B. N is the effective population size in each population.

Probability distribution

Let us consider two randomly chosen nucleons from a population. If a nucleon consists of m sites and each site takes one of K states ($K = 4$ in the case

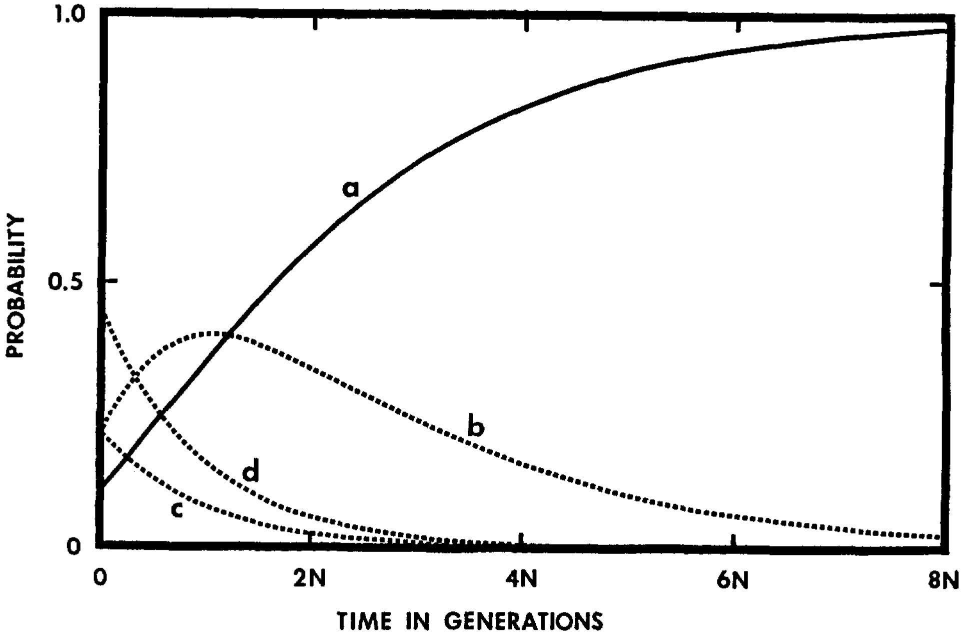

FIGURE 6.—Relationship between the probability of obtaining a certain type of evolutionary relationship in Figure 5 and the divergence time of two populations.

of four nucleotides), the probability that the number of nucleotide differences between two nucleons is k, given that these two nucleons are derived from a common ancestral nucleon t generations ago and the divergence took place $t - 1$ generations ago, is given by

$$P(k\,|\,t) = \binom{m}{k}[g(K, m, t)]^{k}[1 - g(K, m, t)]^{m-k}, \qquad (13)$$

where

$$g(K, m, t) = \frac{K-1}{K}\left\{1 - \exp\left[-\frac{2Kv}{(K-1)m}\,t\right]\right\},$$

and v is the mutation rate per nucleon per generation. $g(K, m, t)$ is the probability that a particular site is polymorphic and can be obtained from equation (30) in TAKAHATA (1982). (Note that in his paper v is the mutation rate per site, not per nucleon.) Since the number of nucleotide differences for a given value of t follows the binomial distribution with probability $g(K, m, t)$, we obtain (13). By using (3), we can obtain the probability that the number of nucleotide differences between two randomly chosen nucleons from a population is k. It is given by

$$P(k) = \sum_{t=1}^{\infty} P(k\,|\,t)f_1(t). \qquad (14)$$

The mean and variance of k are given by

$$E(k) = M\Big/\left[1 + \frac{K}{(K-1)m}M\right], \tag{15}$$

$$V(k) = M\Big/\left[1 + \frac{K}{(K-1)m}M\right] + \left\{1 - \frac{2}{m}\left[1 + \frac{K}{(K-1)m}M\right]\right\} M^2\Big/\left\{\left[1 + \frac{K}{(K-1)m}M\right]^2 \cdot \left[1 + \frac{2K}{(K-1)m}M\right]\right\}, \tag{16}$$

where $M = 4Nv$.

Since K is equal to 4 and m is generally very large, we use the infinite-site model in the following. In this model we assume that the number of sites on a nucleon is so large that a newly arisen mutation takes place at a site different from the sites where the previous mutations have occurred. Under this model $P(k\,|\,t)$ is given by

$$P(k\,|\,t) = \exp(-2vt)(2vt)^k/k!. \tag{17}$$

Therefore, from (14) we have

$$P(k) = [1/(1+M)][M/(1+M)]^k, \tag{18}$$

which is identical with the formula obtained by WATTERSON (1975) using a different method. Application of BAYES' theorem gives the probability that two randomly chosen nucleons were derived from a common ancestral nucleon t generations ago and the divergence took place $t - 1$ generations ago, given that the number of nucleotide differences between two nucleons is k. It becomes

$$f_1(t\,|\,k) = P(k\,|\,t)f_1(t)/P(k) = [(1+M)/(2N)]^{k+1}t^k\exp[-(1+M)t/(2N)]/k!. \tag{19}$$

The mean, variance and mode of t for a given value of k are

$$E(t\,|\,k) = 2N(1+k)/(1+M), \tag{20}$$

$$V(t\,|\,k) = 4N^2(1+k)/(1+M)^2, \tag{21}$$

$$\mathrm{Mode}(t\,|\,k) = 2Nk/(1+M). \tag{22}$$

Some of the distributions of t for a given number of nucleotide differences are given in Figure 7, where $N = 10^5$ and $v = 10^{-6}$ are assumed. It is clear that the distribution is very flat for all values of k. Therefore, k does not give a very reliable estimate of t. When k is large, however, it gives a reasonably reliable estimate of t in terms of the coefficient of variation, as will be discussed later.

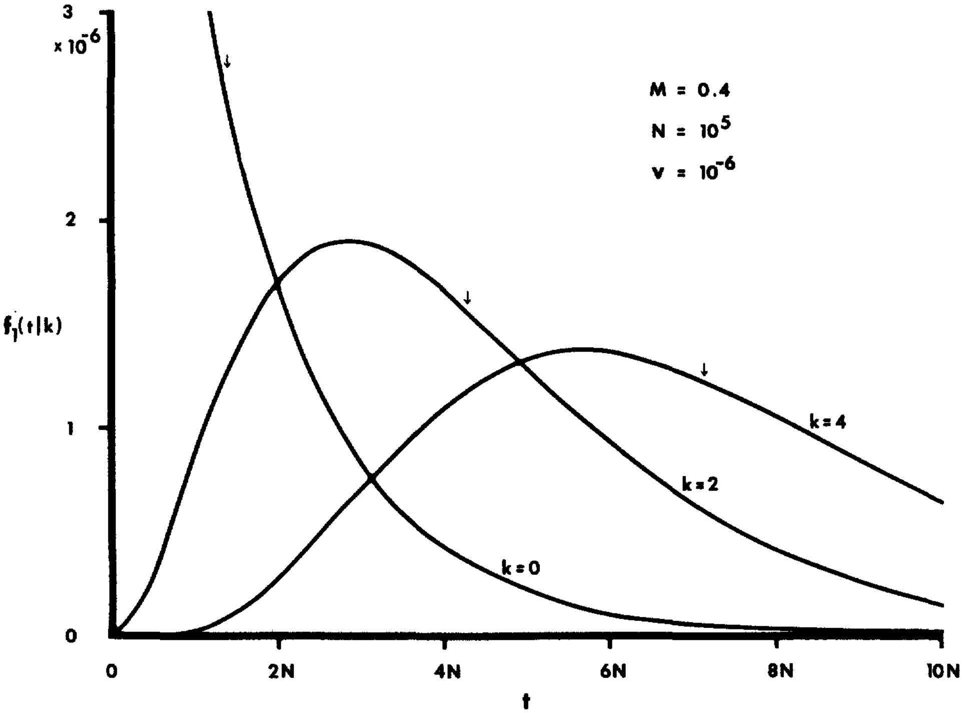

FIGURE 7.—Probability distribution $[f_1(t|k)]$ of the number of generations (t) after divergence of two nucleons, given that k nucleotides are different between two nucleons. Arrow indicates the mean of t.

Mean nucleotide differences for a sample of nucleons

In this section we study the average number of nucleotide differences estimated from pairwise comparison by using the infinite-site model.

When two nucleons are sampled from the population, the mean and variance of k are obtained by substituting $m = \infty$ into (15) and (16). They become

$$E(k) = M, \tag{23}$$

$$V(k) = M + M^2, \tag{24}$$

which agree with WATTERSON'S (1975) result obtained by a different method. When three nucleons are sampled, their evolutionary relationship is given by Figure 1b. In this case we have three estimates, that is, the numbers of nucleotide differences between C and D, between C and E and between D and E (Figure 1b). If we denote the number of nucleotide differences between nucleons i and j by k_{ij}, the following relationships are obtained since under the infinite-site model every mutation can be recognized.

$$k_{CD} = k_{BC} + k_{BD},$$

$$k_{CE} = k_{BF} + k_{BC} + k_{EF},$$

$$k_{DE} = k_{BF} + k_{BD} + k_{EF}.$$

From (23) and (24) we have

$$E(k_{BF}) = M, \qquad V(k_{BF}) = M + M^2.$$

The means, variances and covariances of k_{BC}, k_{BD} and k_{EF} are given by

$$E(k_{BC}) = E(k_{BD}) = E(k_{EF}) = \sum_{t=1}^{\infty} \sum_{k=0}^{\infty} kf_2(t)Q(k\,|\,t) = M/6,$$

$$V(k_{BC}) = V(k_{BD}) = V(k_{EF}) = \sum_{t=1}^{\infty} \sum_{k=0}^{\infty} k^2f_2(t)Q(k\,|\,t) - (M/6)^2 = M/6 + (M/6)^2,$$

$$\begin{aligned}\mathrm{Cov}(k_{BC}, k_{BD}) &= \mathrm{Cov}(k_{BC}, k_{EF}) = \mathrm{Cov}(k_{BD}, k_{EF})\\ &= \sum_{t=1}^{\infty} \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} k_1k_2f_2(t)Q(k_1\,|\,t)Q(k_2\,|\,t) - (M/6)^2 = (M/6)^2,\end{aligned}$$

where $Q(K\,|\,t) = \exp(-vt)(vt)^k/k!$. Therefore, we have

$$E(k_{CD}) = E(k_{BC}) + E(k_{BD}) = M/3,$$

$$V(k_{CD}) = V(k_{BC}) + V(k_{BD}) + 2\ \mathrm{Cov}(k_{BC}, k_{BD}) = M/3 + (M/3)^2,$$

$$E(k_{CE}) = E(k_{BF}) + E(k_{BC}) + E(k_{EF}) = (4/3)M,$$

$$V(k_{CE}) = V(k_{BF}) + V(k_{BC}) + V(k_{EF}) + 2\ \mathrm{Cov}(k_{BC}, k_{EF}) = (4/3)M + (10/9)M^2,$$

$$E(k_{DE}) = E(k_{CE}), \qquad V(k_{DE}) = V(k_{CE}).$$

Using these formulas, we can obtain the means and variances of the average numbers ($\hat{k}$) of nucleotide differences. They are

$$E(\hat{k}) = E[(k_{CD} + k_{CE} + k_{DE})/3] = M, \tag{25}$$

$$V(\hat{k}) = V[(k_{CD} + k_{CE} + k_{DE})/3] = (2/3)M + (5/9)M^2. \tag{26}$$

The mean and variance of the average number of nucleotide differences for a given sample size can be obtained in the same way. For example, when four nucleons are sampled, there are two types of evolutionary relationships as shown in Figure 2. Using the same method, we obtain

$$E_a(\hat{k}) = (17/18)M, \qquad V_a(\hat{k}) = (53/108)M + (115/324)M^2$$

for type a, and

$$E_b(\hat{k}) = (10/9)M, \qquad V_b(\hat{k}) = (37/54)M + (89/162)M^2$$

for type b. Note that both the mean and variance of $\hat{k}$ for the type a relationship are different from those for the type b relationship and that whether the values are overestimates or underestimates of M depends on the type of relationship. The overall mean and variance of $\hat{k}$ can be obtained by taking into account the probability of getting a certain type of relationship. In this case the probabilities of getting types a and b are 2/3 and 1/3, respectively. Therefore, we have

$$E(\hat{k}) = (2/3)E_a(\hat{k}) + (1/3)E_b(\hat{k}) = M, \tag{27}$$

$$\begin{aligned} V(\hat{k}) &= (2/3)V_a(\hat{k}) + (1/3)V_b(\hat{k}) + (2/3)[E_a(\hat{k}) - E(\hat{k})]^2 + (1/3)[E_b(\hat{k}) - E(\hat{k})]^2 \\ &= (5/9)M + (23/54)M^2. \end{aligned} \tag{28}$$

In general, the mean and variance of the average number of nucleotide differences between two nucleons when n nucleons are sampled from the population are given by

$$E(\hat{k}) = M, \tag{29}$$

$$V(\hat{k}) = \frac{n+1}{3(n-1)}M + \frac{2(n^2+n+3)}{9n(n-1)}M^2. \tag{30}$$

Proof of (30) is given in APPENDIX II. As n increases, (30) approaches

$$V_{st}(\hat{k}) = (1/3)M + (2/9)M^2. \tag{31}$$

We call this variance the stochastic variance. The sampling variance is given by

$$V_s(\hat{k}) = V(\hat{k}) - V_{st}(\hat{k}) = \frac{2}{3(n-1)}M + \frac{2(2n+3)}{9n(n-1)}M^2. \tag{32}$$

Table 1 shows the relationship between the standard deviation ($\sigma_{\hat{k}} = [V(\hat{k})]^{1/2}$) of $\hat{k}$ and sample size n. As expected, $\sigma_{\hat{k}}$ decreases as n increases but quickly reaches the asymptotic value. Thus, if we sample ten nucleons, the estimate of $\hat{k}$ is nearly as reliable as that obtained from a sample of $n = 200$. This indicates that for estimating the number of heterozygous nucleotide sites a sample size of ten (or even five) is sufficient.

NUCLEON DIVERSITY

Nucleon diversity is defined as the probability that two nucleons randomly chosen from a population are different (NEI and TAJIMA 1981). It is essentially the same as heterozygosity used in the study of protein polymorphism. In this section we use the K-allele model, in which a nucleon is assumed to take one of K allelic states, and the rate of mutation is the same for all alleles or nucleomorphs. In the earlier part of the previous section we considered the m-site-K-state model. If a nucleon has only one site, *i.e.*, $m = 1$, and this site takes one of K states, then this model becomes identical with the K-allele model. Let us denote by $P(0)$ the probability that two nucleons randomly chosen from a population are identical and $P(1) = 1 - P(0)$. The probability that two nucleons randomly chosen from a population are identical, given that these two nucleons separated t generations ago, is given by

$$P(0 \mid t) = 1/K + [(K-1)/K]\exp[-2Kvt/(K-1)], \tag{33}$$

whereas the probability that two nucleons randomly chosen from a population are different, given that these two nucleons separated t generations ago, is

$$P(1 \mid t) = [(K-1)/K]\{1 - \exp[-2Kvt/(K-1)]\}. \tag{34}$$

Equation (33) can be obtained by substituting $m = 1$ and $k = 0$ into (13), and

TABLE 1

Effect of sample size (n) *on the standard deviations of the nucleon diversity* (Ĥ) *and the average number of nucleotide differences* (k̂).

n	$E(\hat{H})$	$\sigma_{\hat{H}}$	$E(\hat{k})$	$\sigma_{\hat{k}}$
		$M = 0.1$		
2	0.091	0.287	0.100	0.332
5		0.199		0.232
10		0.178		0.209
50		0.163		0.192
200		0.160		0.190
∞		0.159		0.189
		$M = 1$		
2	0.500	0.500	1.000	1.414
5		0.296		0.931
10		0.247		0.829
50		0.212		0.761
200		0.206		0.749
∞		0.204		0.745
		$M = 10$		
2	0.909	0.287	10.000	10.488
5		0.113		6.455
10		0.070		5.655
50		0.040		5.160
200		0.034		5.081
∞		0.033		5.055

(34) can be obtained from $P(1\,|\,t) = 1 - P(0\,|\,t)$. From (14) we have

$$P(0) = (K - 1 + M)/(K - 1 + KM), \tag{35}$$

$$P(1) = (K - 1)M/(K - 1 + KM). \tag{36}$$

Equation (36) was first derived by KIMURA (1968) using a different method. Application of BAYES' theorem gives the probability that two nucleons diverged t generations ago, given that the two nucleons are identical $[f_1(t\,|\,0)]$ or different $[f_1(t\,|\,1)]$. Namely,

$$\begin{aligned} f_1(t\,|\,0) = P(0\,|\,t)f_1(t)/P(0) = \{&1/K + [(K - 1)/K]\exp[-2Kvt/(K - 1)]\} \\ &\times [1/(2N)]\exp[-t/(2N)](K - 1 + KM)/(K - 1 + M), \end{aligned} \tag{37}$$

$$\begin{aligned} f_1(t\,|\,1) = P(1\,|\,t)f_1(t)/P(1) = \{&1 - \exp[-2Kvt/(K - 1)]\} \\ &\times [1/(2N)]\exp[-t/(2N)](K - 1 + KM)/(KM). \end{aligned} \tag{38}$$

The mean of t for a pair of identical nucleons is given by

$$E(t\,|\,0) = 2N\{1 - (K - 1)^2M/[(K - 1 + M)(K - 1 + KM)]\}, \tag{39}$$

whereas that for a pair of nonidentical nucleons is given by

$$E(t|1) = 2N[1 + (K - 1)/(K - 1 + KM)]. \tag{40}$$

Furthermore, their respective variances are

$$V(t|0) = 4N^2\{1 - (K - 1)^3M[2(K - 1) + (K + 1)M]/[(K - 1 + M)^2(K - 1 + KM)^2]\}, \tag{41}$$

$$V(t|1) = 4N^2[1 + (K - 1)^2/(K - 1 + KM)^2]. \tag{42}$$

As K increases, the mean and variance approach the following formulas.

$$E(t|0) = 2N/(1 + M), \tag{43}$$

$$E(t|1) = 2N[1 + 1/(1 + M)], \tag{44}$$

$$V(t|0) = 4N^2/(1 + M)^2, \tag{45}$$

$$V(t|1) = 4N^2/[1 + 1/(1 + M)^2]. \tag{46}$$

Figure 8 shows the relationship between M and $E(t|k)$, where $K = 4$ and $K = \infty$ are assumed. In the case of $K = \infty$, both the mean of t $[E(t|0)]$ for a pair of identical nucleons and that $[E(t|1)]$ for a pair of nonidentical nucleons decrease as M increases. Interestingly, the difference between them is always $2N$ generations [see (43) and (44)]. On the other hand, in the case in which K is a finite number, both $E(t|0)$ and $E(t|1)$ approach $2N$ generations as M increases [see (39) and (40)]. In either case the variance of t is too large to obtain a reliable estimate of t.

In the case of DNA sequences there are many possible nucleomorphs, so that K is very large. Therefore, we can use the infinite-allele model, in which the number of possible nucleomorphs is assumed to be so large that any mutation creates a new nucleomorph. When two nucleons are sampled from a population, the mean and variance of nucleon diversity (H) are given by

$$E(H) = M/(1 + M), \tag{47}$$

$$V(H) = M/(1 + M)^2. \tag{48}$$

These formulas can be obtained by substituting $K = \infty$ and $m = 1$ into (15) and (16).

When the sample size is more than two, we must again consider the relationship between nucleons as we did in the case of the infinite-site model. We can then obtain the probability that the average nucleon diversity is $\hat{H}$ for a given sample size. For example, in the case in which three nucleons are sampled from the population, this probability is given by

$$Pr\{\hat{H} = 0\} = 2/[(1 + M)(2 + M)],$$

$$Pr\{\hat{H} = 2/3\} = 3M/[(1 + M)(2 + M)],$$

$$Pr\{\hat{H} = 1\} = M^2/[(1 + M)(2 + M)].$$

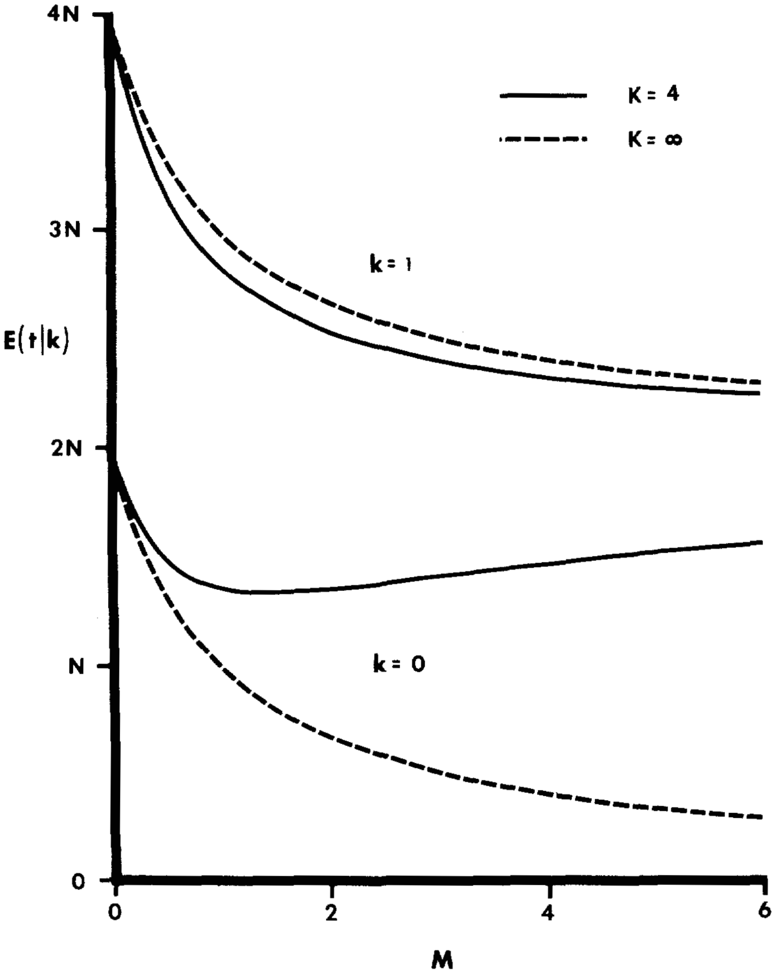

FIGURE 8.—Relationship between the expected number of generations [$E(t|k)$] after divergence of two nucleons, given that two nucleons are identical ($k = 0$) or that two nucleons are different ($k = 1$), and the value of M.

From this we obtain

$$E(\hat{H}) = M/(1 + M),$$
$$V(\hat{H}) = M(4 + M)/[3(1 + M)^2(2 + M)].$$

In general the mean and variance of $\hat{H}$ when n nucleons are sampled from a population are given by

$$E(\hat{H}) = M/(1 + M), \tag{49}$$
$$V(\hat{H}) = 2M(n + M)(n + 1 + M)/[n(n - 1)(1 + M)^2(2 + M)(3 + M)], \tag{50}$$

respectively. As n increases, this variance approaches

$$V_{st}(\hat{H}) = 2M/\{(1 + M)^2(2 + M)(3 + M)\}, \tag{51}$$

which agrees with the earlier result obtained by WATTERSON (1974a), LI and NEI (1975) and STEWART (1976). The sampling variance is given by

$$\begin{aligned} V_s(\hat{H}) &= V(\hat{H}) - V_{st}(\hat{H}) \\ &= 2M(2n + M)/\{n(n - 1)(1 + M)(2 + M)(3 + M)\}, \end{aligned} \tag{52}$$

which again agrees with the earlier result by NEI (1978).

The effect of sample size on the standard deviation ($\sigma_{\hat{H}} = [V(\hat{H})]^{1/2}$) of $\hat{H}$ can be seen from Table 1. When M is relatively small, the effect of increasing n beyond 10 is relatively small as in the case of $\sigma_{\hat{k}}$. However, when M is large, $\sigma_{\hat{H}}$ is significantly reduced by increasing n to more than 10.

DISTRIBUTION OF SAMPLE NUCLEOMORPH FREQUENCIES

WATTERSON (1974b) presented a formula for computing the expected number of nucleomorphs (alleles) with a given frequency in the sample. In this section, we derive this formula by using a different but simple method. We use the infinite-allele model in this section.

In the case of $n = 2$, the probability that the two nucleons sampled are identical is $1/(1 + M)$ and that the probability that the two nucleons are different is $M/(1 + M)$. Therefore, the expected number of nucleomorphs with frequency 1 is $F_2(1) = 1/(1 + M)$, and the expected number of nucleomorphs with frequency ½ is $F_2(½) = 2M/(1 + M)$.

When n is larger than 2, we must again consider the relationship between nucleons studied earlier. The expected number of nucleomorphs with frequency x in a sample of n nucleons is given by

$$\begin{aligned} &F_n(0) = 0, \\ &F_n(x) = Mx^{-1}(1 - x)^{-1}\Big/\left[n \prod_{i=n(1-x)}^{n-1} (1 + M/i)\right] \quad \text{when} \quad 0 < x < 1, \\ &F_n(1) = 1\Big/\prod_{i=1}^{n-1} (1 + M/i). \end{aligned} \tag{53}$$

Proof of (53) is given in APPENDIX III. As n increases, $F_n(x)$ approaches $Mx^{-1}(1 - x)^{M-1}/n$, since

$$\lim_{n\to\infty} \prod_{i=n(1-x)}^{n-1} (1 + M/i) = (1 - x)^{-M}.$$

Some examples of this distribution are given in Table 2. It is clear that a small number of samples are sufficient in order to know the distribution pattern unless M is very large.

DISCUSSION

Our study on the topological relationship among nucleons has shown that when the time since divergence between two populations is relatively short a nucleon sampled from one population is often more similar to the one sampled

TABLE 2

Expected number of nucleomorphs with frequency x *for a given sample size* (n)

	n			
x	10	50	200	10,000
$M = 0.1$				
0.00–0.05	0.000	0.154	0.292	0.684
0.05–0.15	0.110	0.119	0.120	0.120
0.15–0.25	0.061	0.062	0.062	0.062
0.25–0.35	0.046	0.046	0.046	0.046
0.35–0.45	0.039	0.040	0.040	0.040
0.45–0.55	0.037	0.037	0.037	0.037
0.55–0.65	0.038	0.038	0.038	0.038
0.65–0.75	0.042	0.042	0.042	0.042
0.75–0.85	0.052	0.054	0.054	0.054
0.85–0.95	0.084	0.094	0.095	0.095
0.95–1.00	0.759	0.747	0.745	0.745
Total	1.269	1.433	1.572	1.963
$M = 1$				
0.00–0.05	0.000	1.500	2.879	6.792
0.05–0.15	1.000	1.093	1.099	1.099
0.15–0.25	0.500	0.508	0.511	0.511
0.25–0.35	0.333	0.336	0.336	0.336
0.35–0.45	0.250	0.251	0.251	0.251
0.45–0.55	0.200	0.201	0.201	0.201
0.55–0.65	0.167	0.167	0.167	0.167
0.65–0.75	0.143	0.143	0.143	0.143
0.75–0.85	0.125	0.125	0.125	0.125
0.85–0.95	0.111	0.111	0.111	0.111
0.95–1.00	0.100	0.061	0.054	0.051
Total	2.929	4.499	5.878	9.788
$M = 10$				
0.00–0.05	0.000	12.054	25.016	63.841
0.05–0.15	5.263	5.127	4.921	4.818
0.15–0.25	1.316	0.899	0.791	0.753
0.25–0.35	0.413	0.204	0.163	0.150
0.35–0.45	0.135	0.046	0.032	0.028
0.45–0.55	0.043	0.009	0.006	0.005
0.55–0.65	0.013	0.002	0.001	0.001
0.65–0.75	0.003			
0.75–0.85	0.001			
0.85–0.95				
0.95–1.00				
Total	7.188	18.342	30.930	69.595

from the other population than to another nucleon sampled from the same population (Figure 5b, c, d). This is counter-intuitive, but the probability of this event is large when $t \leq 2N$. This finding has an important implication for interpretation of data. For example, CANN, BROWN and WILSON (1982) and

NEI (1982) constructed evolutionary trees for mitochondrial DNAs sampled from various human races (whites, American blacks, Orientals, etc.). Although mtDNAs from the same race generally showed a higher similarity with each other than with those from other races, there were many exceptions. This led to the suggestion that some of the exceptions are due to interracial gene admixture. The present finding indicates that such exceptions can occur even without gene admixture.

When genetic variation within populations is measured, the average number of nucleotide differences (nucleotide diversity) is more informative than heterozygosity or nucleon diversity, since the former gives information on the extent of DNA difference between two randomly chosen genes, whereas the latter gives information only on whether a pair of genes (or nucleons) are the same or not. There are two other quantities for measuring genetic variation, *i.e.*, the number of alleles or nucleomorphs by EWENS (1972) and the number of segregating (nucleotide) sites by WATTERSON (1975). When all of the mutants observed are selectively neutral, the number of segregating sites gives the best estimate of M or nucleotide diversity among the four methods because of the smallest variance of the estimate of M. However, when some of the mutants observed are deleterious, this measure is not necessarily good. Deleterious mutants are maintained in a population with low frequency. Since the number of alleles (nucleomorphs) and the number of segregating sites ignore the frequency of mutants, these two quantities might be strongly affected by the existence of deleterious mutants. On the other hand, in the average number of nucleotide differences and heterozygosity the frequency of mutants is considered, so that the existence of deleterious mutants with low frequency does not affect these quantities very much. Furthermore, even if deleterious mutants affect these four quantities, the average number of nucleotide differences and heterozygosity have clear biological meanings. Although the number of alleles (nucleomorphs) and the number of segregating sites are biologically clear, these two quantities cannot be used directly since they are dependent on the sample size. For these reasons, I would like to recommend that the average number of nucleotide differences be used for measuring genetic variation within populations.

Nevertheless, it should be noted that the average number of nucleotide differences is accompanied by a large variance. This can be seen from AQUADRO and GREENBERG'S (1983) data. These authors studied a sequence of about 900 nucleotide pairs of the human mitochondrial DNA for seven individuals (nucleons). The average number of nucleotide differences ($\hat{k}$) estimated was 15.4. From (30), (31) and (32), we estimate the total variance, stochastic variance and sampling variances of $\hat{k}$ to be 80.9, 57.8 and 23.0, respectively. Thus, a majority of the total variance is attributed to stochastic errors.

As mentioned earlier, we can estimate the number of generations after divergence between a pair of nucleons from (20). In AQUADRO and GREENBERG'S (1983) data, the number of nucleotide differences (k) between nucleons 1 and 2 (see Table 2 in their paper) is 5. If $\hat{k} = 15.4$ is an accurate estimate of M [see (29)], the estimate of the number of generations after divergence of these nucleons is $0.7N \pm 0.3N$ from (20) and (21). On the other hand, the number of nucleotide differences between nucleons 3 and 4 is 28, and the estimate of the number of generations is $3.5N \pm 0.7N$. This example shows that the estimate of t has a smaller coefficient of variation when k is large than when k is small.

When M is large and sample size is small, all nucleons sampled are often different. For example, all 21 human mitochondrial DNAs studied by BROWN (1980) were different with respect to restriction-site sequence. It is, therefore, interesting to know the probability that all n nucleons sampled from a population are different. This probability can be obtained in the following way. We first note that for all n nucleons sampled to be different from each other a particular nucleon of the n nucleons sampled must be different from the remaining $n - 1$ nucleons. At the same time, a particular nucleon of the $n - 1$ nucleons must be different from the $n - 2$ remaining nucleons. A similar condition is required for all $n - i$, where i is $0, 1, 2, \ldots, n - 1$. We also note that the probability of a particular nucleon of the i nucleons sampled being different from the remaining $i - 1$ nucleons is $F_i(1/i)/i$ from (53). Thus, the probability that all n nucleons sampled from a population are different from each other is given by

$$\prod_{i=2}^{n} F_i(1/i)/i = 1/\prod_{i=1}^{n-1} (1 + i/M). \tag{54}$$

Incidentally, the probability that all n nucleons sampled from a population are identical is

$$F_n(1) = 1/\prod_{i=1}^{n-1} (1 + M/i). \tag{55}$$

These formulas are identical with the formulas obtained by EWENS (1972) using a different method. For example, in the case of AQUADRO and GREENBERG'S (1983) data, our estimate of M is 15.4 and n is 7. From (54) the probability that all seven sampled nucleons are different is 0.30.

Recently, WATTERSON (1982a, b) showed that the times at which mutations at nucleotide sites become fixed in a population tend to cluster together rather than behave as a Poisson process when the nucleotide sites are completely linked. This implies that, when two loci are linked, the behaviors of two loci are not independent. We can see this from two different points of views.

Let us assume that two loci, say loci A and B, are completely linked and that the mutation rate at each locus is v. When we sample two chromosomes from a population, the probability that both loci A and B are homozygous is $1/(1 + 2M)$, which is not equal to $1/(1 + M)^2$. Thus, two loci are not independent. Namely, given that locus B is homozygous, the probability that locus A is homozygous is $(1 + M)/(1 + 2M)$. Similarly, given that locus B is heterozygous, the probability that locus A is heterozygous is $2M/(1 + 2M)$. These probabilities are larger than those for independent loci. In fact, the heterozygosity for locus A is positively correlated with that for locus B as will be published elsewhere.

Another way to see the correlation between linked loci is to study the average number of nucleotide differences. If we denote $\hat{k}$'s for loci A and B by $\hat{k}_A$ and $\hat{k}_B$, the covariance between $\hat{k}_A$ and $\hat{k}_B$ is given by

$$\mathrm{Cov}(\hat{k}_A, \hat{k}_B) = \frac{2(n^2 + n + 3)}{9n(n - 1)} M^2, \tag{56}$$

and the correlation coefficient is

$$\rho(\hat{k}_A, \hat{k}_B) = 1/[1 + 3n(n + 1)/\{2(n^2 + n + 3)M\}]. \tag{57}$$

These results can be obtained from the following relationship.

$$V(\hat{k}_A + \hat{k}_B) = V(\hat{k}_A) + 2\ \mathrm{Cov}(\hat{k}_A, \hat{k}_B) + V(\hat{k}_B),$$

where $V(\hat{k}_A)$ and $V(\hat{k}_B)$ are given by (30), and $V(\hat{k}_A + \hat{k}_B)$ are obtained from (30) by using $2M$ instead of M. Again, this quantity is positively correlated.

From these studies we can conclude that, when a locus is highly polymorphic, a locus linked to it also tends to be highly polymorphic.

I thank M. NEI, G. A. WATTERSON and R. CHAKRABORTY for valuable suggestions and comments.

LITERATURE CITED

AQUADRO, C. F. and B. D. GREENBERG, 1983 Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics **103:** 287–312.

BROWN, W. M., 1980 Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. Proc. Natl. Acad. Sci. USA **77:** 3605–3609.

BURROWS, P. M. and C. C. COCKERHAM, 1974 Distributions of time to fixation of neutral genes. Theor. Pop. Biol. **5:** 192–207.

CANN, R. L., W. M. BROWN and A. C. WILSON, 1982 Evolution of human mitochondrial DNA: molecular, genetic, and anthropological implications. pp. 157–165. In: *Human Genetics, Part A: The Unfolding Genome*, edited by B. Bonné-Tamir. Alan R. Liss, New York.

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

HUDSON, R. R., 1983 Testing the constant-rate neutral allele model with protein sequence data. Evolution **37:** 203–217.

KIMURA, M., 1968 Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. **11:** 247–269.

KIMURA, M. and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725–738.

KIMURA, M. and T. OHTA, 1969 The average number of generations until fixation of a mutant gene in a finite population. Genetics **61:** 763–771.

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

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

NEI, M., 1982 Evolution of human races at the gene level. pp. 167–181. In: *Human Genetics, Part A: The Unfolding Genome*, edited by B. Bonné-Tamir. Alan R. Liss, New York.

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

STEWART, F. M., 1976 Variability in the amount of heterozygosity maintained in neutral populations. Theor. Pop. Biol. **9:** 188–201.

TAKAHATA, N., 1982 Linkage disequilibrium, genetic distance and evolutionary distance under a general model of linked genes or a part of the genome. Genet. Res. **39:** 63–77.

WATTERSON, G. A., 1974a Models for the logarithmic species abundance distributions. Theor. Pop. Biol. **6:** 217–250.

WATTERSON, G. A., 1974b The sampling theory of selectively neutral alleles. Adv. Appl. Prob. **6:** 463–488.

WATTERSON, G. A., 1975 On the number of segregating sites in genetic models without recombination. Theor. Popul. Biol. **7:** 256–276.

WATTERSON, G. A., 1982a Mutant substitutions at linked nucleotide sites. Adv. Appl. Prob. **14:** 206–224.

WATTERSON, G. A., 1982b Substitution times for mutant nucleotides. J. Appl. Prob. **19A:** 59–70.

Corresponding editor: B. S. WEIR

APPENDIX I: TOPOLOGICAL RELATIONSHIP AMONG NUCLEONS RANDOMLY SAMPLED FROM A POPULATION

Let us consider the evolutionary relationship among n nucleons. We assume that all of the branches are created by bifurcation. Let us start from a common ancestral nucleon. The bifurcation of this ancestral nucleon creates two branches (see Figure 1a). Next, one of two branches bifurcates. In this case there are two possible ways of bifurcation, although they create the same relationship (see Figure 1b). After continuation of this process, one of $n - 1$ nucleons finally bifurcates. In this case there are $n - 1$ possible ways of bifurcation. Thus, there are $(n - 1)!$ possible ways of bifurcation to create n nucleons from a common ancestral nucleon. As we have noticed, however, some bifurcations create the same relationship. There are $n - 1$ branching points in a tree with n nucleons. When a particular branching point is asymmetrical, there are two ways of bifurcation that create the same relationship. This is because we regard two trees that are mirror imaged as the same relationship. The number of asymmetrical branching points is $n - 1 - s$, where s is the number of branching points that lead to exactly two nucleon descendants in the sample. Thus, we obtain (1).

We now study the topological relationship among n nucleons. Let us consider the case in which a particular branching point divides n nucleons. We denote by $Q(n_1, n_2|n)$ the probability that the left side of this branching point has n_1 nucleons and the right side of it has n_2 nucleons, where $n_1 + n_2 = n$. When one of these n nucleons bifurcates, bifurcation occurs on the left side with probability n_1/n and on the right side with probability n_2/n. Thus, we have

$$Q(n_1, n_2|n) = Q(n_1 - 1, n_2|n - 1)(n_1 - 1)/(n - 1) + Q(n_1, n_2 - 1|n - 1)(n_2 - 1)/(n - 1). \quad (A1)$$

By using $Q(1, 1|2) = 1$ as the initial condition, we obtain

$$Q(n_1, n_2|n) = 1/(n - 1). \quad (A2)$$

If we denote by $P(n_1, n_2)$ the probability that a certain branching point divides n nucleons into n_1 and n_2 nucleons (order not important), we obtain (2).

APPENDIX II. MEAN AND VARIANCE OF THE AVERAGE NUMBER OF NUCLEOTIDE DIFFERENCES

The average number of nucleotide differences is defined by

$$\hat{k} = \sum_{i<j}\sum k_{ij}/\tbinom{n}{2}, \quad (A3)$$

where k_{ij} is the number of nucleotide differences between the ith and jth nucleons. Since $E(k_{ij}) = M$ from (23), we obtain

$$E(\hat{k}) = \sum_{i<j}\sum E(k_{ij})/\tbinom{n}{2} = M. \tag{A4}$$

The variance of $\hat{k}$ is

$$V(\hat{k}) = E(\hat{k^2}) - \{E(\hat{k})\}^2. \tag{A5}$$

$\hat{k^2}$ can be written as

$$\begin{aligned}\hat{k^2} &= \left(\sum_{i<j}\sum k_{ij}\right)^2/\tbinom{n}{2}^2 \\ &= \left(\sum_{i<j}\sum {k_{ij}}^2 + \sum_{i\neq j\neq r}\sum\sum k_{ij}k_{ir} + \sum_{\substack{i<j \\ i\neq j\neq r\neq s}}\sum \sum_{r<s}\sum k_{ij}k_{rs}\right)/\tbinom{n}{2}^2.\end{aligned} \tag{A6}$$

Let us now define

$$\begin{aligned}U_2 &= E({k_{ij}}^2) - M^2, \\ U_3 &= E(k_{ij}k_{ir}) - M^2, \\ U_4 &= E(k_{ij}k_{rs}) - M^2.\end{aligned} \tag{A7}$$

Since $E(\hat{k}) = M$ from (A4), (A5) becomes

$$\begin{aligned}V(\hat{k}) &= \left(\sum_{i<j}\sum U_2 + \sum_{i\neq j\neq r}\sum\sum U_3 + \sum_{\substack{i<j \\ i\neq j\neq r\neq s}}\sum \sum_{r<s}\sum U_4\right)/\tbinom{n}{2}^2 \\ &= \{U_2 + 2(n-2)U_3 + \tbinom{n-2}{2}U_4\}/\tbinom{n}{2}.\end{aligned} \tag{A8}$$

When $n = 2$, $V(\hat{k}) = M + M^2$ from (24). Therefore, from (A8) we have

$$M + M^2 = U_2. \tag{A9}$$

When $n = 3$, $V(\hat{k}) = (2/3)M + (5/9)M^2$ from (26). Thus, we have

$$(2/3)M + (5/9)M^2 = (U_2 + 2U_3)/3. \tag{A10}$$

When $n = 4$, from (28) we have

$$(5/9)M + (23/54)M^2 = (U_2 + 4U_3 + U_4)/6. \tag{A11}$$

From (A9), (A10) and (A11), we get

$$\begin{aligned}U_2 &= M + M_2, \\ U_3 &= (1/2)M + (1/3)M^2, \\ U_4 &= (1/3)M + (2/9)M^2.\end{aligned} \tag{A12}$$

By substituting (A12) into (A8), we obtain

$$V(\hat{k}) = \frac{n+1}{3(n-1)}M + \frac{2(n^2+n+3)}{9n(n-1)}M^2. \tag{A13}$$

APPENDIX III. EXPECTED NUMBER OF NUCLEOMORPHS WITH A GIVEN FREQUENCY IN THE SAMPLE

The expected number of nucleomorphs with frequency $(p, p + dp)$ in a population is

$$\phi(p)dp = Mp^{-1}(1-p)^{M-1}dp$$

(KIMURA and CROW 1964). The expected number of nucleomorphs with frequency x ($=i/n$) in a sample of n nucleons is given by

$$F_n(x) = \int_0^1 \binom{n}{i} p^i(1-p)^{n-i}\phi(p)dp. \tag{A14}$$

When $0 < x < 1$, (A14) becomes

$$F_n(x) = Mx^{-1}(1-x)^{-1}/\left\{n \prod_{i=n(1-x)}^{n-1} (1+M/i)\right\}. \tag{A15}$$

When $x = 1$, (A14) becomes

$$F_n(1) = 1/\prod_{i=1}^{n-1} (1+M/i). \tag{A16}$$