# Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism

**Fumio Tajima**

*Department of Biology, Kyushu University, Fukuoka 812, Japan*

Manuscript received February 13, 1989
Accepted for publication July 14, 1989

ABSTRACT

The relationship between the two estimates of genetic variation at the DNA level, namely the number of segregating sites and the average number of nucleotide differences estimated from pairwise comparison, is investigated. It is found that the correlation between these two estimates is large when the sample size is small, and decreases slowly as the sample size increases. Using the relationship obtained, a statistical method for testing the neutral mutation hypothesis is developed. This method needs only the data of DNA polymorphism, namely the genetic variation within population at the DNA level. A simple method of computer simulation, that was used in order to obtain the distribution of a new statistic developed, is also presented. Applying this statistical method to the five regions of DNA sequences in *Drosophila melanogaster*, it is found that large insertion/deletion (>100 bp) is deleterious. It is suggested that the natural selection against large insertion/deletion is so weak that a large amount of variation is maintained in a population.

A large amount of genetic variation is maintained in natural populations. Information about this variation at the DNA level can be obtained from DNA sequencing or restriction enzyme technique. WATTERSON (1975) has shown under the neutral mutation model (KIMURA 1968, 1983) that the expectation and variance of the number ($S$) of segregating (or polymorphic) sites in the sample are given by

$$E(S) = a_1 M, \tag{1}$$

and

$$V(S) = a_1 M + a_2 M^2, \tag{2}$$

respectively, where $M = 4Nu$, $N$ is effective population size, $u$ is the mutation rate per generation per DNA sequence under investigation,

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}, \tag{3}$$

$$a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}, \tag{4}$$

and $n$ is the sample size (the number of DNA sequences studied), so that $M$ can be estimated from

$$\hat{M} = \frac{S}{a_1}. \tag{5}$$

It should be noted that $S$ itself is not a good statistic for estimating the DNA polymorphism, since $S$ depends on the sample size. On the other hand, TAJIMA (1983) has shown under the neutral mutation model that the expectation and variance of the average number ($\hat{k}$) of (pairwise) nucleotide differences between the DNA sequences examined are given by

$$E(\hat{k}) = M, \tag{6}$$

and

$$V(\hat{k}) = b_1 M + b_2 M^2, \tag{7}$$

respectively, where

$$b_1 = \frac{n+1}{3(n-1)}, \tag{8}$$

and

$$b_2 = \frac{2(n^2+n+3)}{9n(n-1)}. \tag{9}$$

This number ($\hat{k}$) not only has clear biological meanings, but also gives the estimate of $M$ directly.

The remarkable and important difference between the number of segregating sites and the average number of nucleotide differences is the effect of selection. Deleterious mutants are maintained in a population with low frequency. Since the number of segregating sites ignores the frequency of mutants, this number might be strongly affected by the existence of deleterious mutants. On the other hand, the existence of deleterious mutants with low frequency does not affect the average number of nucleotide differences very much, since in this case the frequency of mutants is considered. In other words, if some of the mutants observed have selective effects, then the estimate of $M$ obtained from (5) by using the number of segre-

gating sites may not be the same as the average number of nucleotide differences which also is the estimate of $M$.

In this paper I shall investigate the relationship between the number of segregating sites and the average number of nucleotide differences under the neutral mutation model. Using this relationship obtained, I shall also present a statistical method for testing the neutral mutation hypothesis.

## RELATIONSHIP BETWEEN THE NUMBER OF SEGREGATING SITES AND THE AVERAGE NUMBER OF NUCLEOTIDE DIFFERENCES

**Assumption:** In this paper we consider a random mating population of $N$ diploid individuals and assume that there is no selection and no recombination between DNA sequences. We also assume that the number of sites on a DNA sequence is so large that a newly arisen mutation takes place at a site different from the sites where the previous mutations have occurred [infinite site model (KIMURA 1969)]. Under these assumptions the expectation and variance of the number of segregating sites are given by (1) and (2), and the expectation and variance of the average number of nucleotide differences are given by (6) and (7).

**Covariance between the number of segregating sites and the average number of nucleotide differences:** If we denote the number of nucleotide differences between the $i$th and $j$th DNA sequences by $k_{ij}$, the average number of (pairwise) nucleotide differences between the DNA sequences sampled is given by

$$\hat{k} = \frac{\sum\sum_{i<j} k_{ij}}{\binom{n}{2}}, \tag{10}$$

where $n$ is the number of DNA sequences sampled. Incidentally $\hat{k}$ can also be estimated from

$$\hat{k} = \sum_{i=1}^{S} h_i, \tag{11}$$

where $S$ is the number of segregating sites, and $h_i$ is the unbiased estimate of nucleotide diversity (or heterozygosity) for the $i$th segregating site, which is given by

$$h_i = \frac{n\left(1 - \sum_j x_{ji}^2\right)}{n-1}, \tag{12}$$

where $x_{ji}$ is the sample frequency of the $j$th allelic nucleotide in the $i$th segregating site. When the sample size ($n$) is large, (11) is more practical than (10).

If we use (10), the covariance between the number of segregating sites and the average number of nucleotide differences can be given by

$$\mathrm{Cov}(S, \hat{k}) = \mathrm{Cov}(S, k_{ij}). \tag{13}$$

This covariance can be obtained from the genealogical relationship of DNA sequences.

When $n$ is 2, $S$ is equal to $k_{ij}$ (Figure 1a), so that $\mathrm{Cov}(S, k_{ij}) = V(k_{ij}) = V(S)$. From (2) $V(S)$ is equal to $M + M^2$. Therefore, we have

$$\mathrm{Cov}(S, \hat{k}) = M + M^2. \tag{14}$$

The genealogical relationship when $n$ is 3 is shown in Figure 1b. In this case there are two possible common ancestors (namely $A$ and $B$) between the two DNA sequences which are randomly chosen from the three DNA sequences. Since $B$ is the common ancestor when $C$ and $D$ are chosen, and $A$ is the common ancestor when $C$ and $E$, or $D$ and $E$ are chosen, the probability that $B$ is the common ancestor is ⅓, and that of $A$ is ⅔. Therefore, the covariance is given by

$$\mathrm{Cov}(S, \hat{k}) = \tfrac{1}{3}\,\mathrm{Cov}(S, k_{CD}) + \tfrac{2}{3}\,\mathrm{Cov}(S, k_{CE}),$$

where $S = k_{BF} + k_{BC} + k_{BD} + k_{EF}$. If we notice that the distributions of $k_{BC}$, $k_{BD}$, and $k_{EF}$ are the same, we can get

$$\mathrm{Cov}(S, k_{CE}) = V(k_{BF}) + \mathrm{Cov}(S, k_{CD}).$$

TAJIMA (1983) has shown that

$$V(k_{BF}) = M + M^2, \qquad V(k_{BC}) = M/6 + M^2/36,$$

$$\mathrm{Cov}(k_{BC}, k_{BD}) = M^2/36,$$

so that we have

$$\begin{aligned}\mathrm{Cov}(S, k_{CD}) &= V(k_{CD}) + 2\mathrm{Cov}(k_{BC}, k_{BD}) \\ &= 2V(k_{BC}) + 4\mathrm{Cov}(k_{BC}, k_{BD}) = M/3 + M^2/6.\end{aligned}$$

Using these equations, we obtain

$$\begin{aligned}\mathrm{Cov}(S, \hat{k}) &= \tfrac{2}{3}\,V(k_{BF}) + \mathrm{Cov}(S, k_{CD}) \\ &= M + \tfrac{5}{6}\,M^2.\end{aligned} \tag{15}$$

Next, we consider the case where the number of DNA sequences sampled is more than 3. $n$ DNA sequences take place when one of $n-1$ DNA sequences bifurcates. Suppose that such a bifurcation occurred at point $A$ in Figure 1c, and that its descendants are $B$ and $C$. Then the covariance between $S$ and $\hat{k}$ is given by

$$\begin{aligned}\mathrm{Cov}(S, \hat{k}) = &\frac{1}{\binom{n}{2}}\,\mathrm{Cov}(S, k_{BC}) \\ &+\left(1 - \frac{1}{\binom{n}{2}}\right)\mathrm{Cov}(S, k_{ij}),\end{aligned} \tag{16}$$

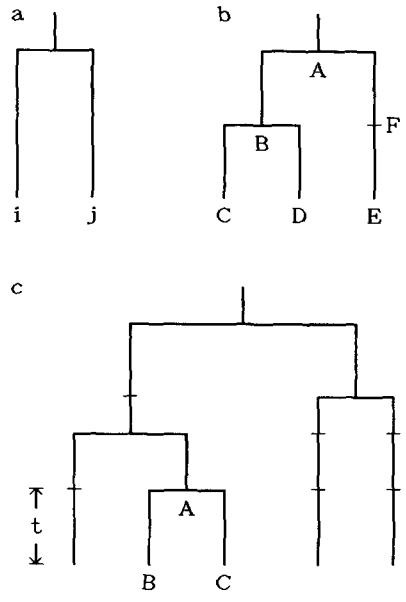


FIGURE 1.—(a) Expected genealogical relationship when two DNA sequences are sampled from a population. (b) Expected genealogical relationship when three DNA sequences are sampled from a population. (c) One example of the genealogical relationship among five DNA sequences sampled from a population.

where $k_{ij}$ is not equal to $k_{BC}$. Using the same method as the above, we can have

$$\mathrm{Cov}(S, k_{BC}) = V(k_{BC}) + 2(n-2)\mathrm{Cov}(k_{AB}, k_{AC}), \quad (17)$$

and

$$\mathrm{Cov}(S, k_{ij}) = \mathrm{Cov}(S^*, \hat{k}^*) + V(k_{BC}) + 2(n-2)\mathrm{Cov}(k_{AB}, k_{AC}), \quad (18)$$

where $S^*$ and $\hat{k}^*$ are the number of segregating sites and the average number of nucleotide differences for $n-1$ DNA sequences, respectively. Substituting (17) and (18) into (16), we have

$$\mathrm{Cov}(S, \hat{k}) = \frac{(n+1)(n-2)}{n(n-1)}\mathrm{Cov}(S^*, \hat{k}^*) + V(k_{BC}) + 2(n-2)\mathrm{Cov}(k_{AB}, k_{AC}). \quad (19)$$

Following TAJIMA (1983), we can obtain $V(k_{BC})$ and $\mathrm{Cov}(k_{AB}, k_{AC})$. HUDSON (1983) and TAJIMA (1983) have shown that the probability that $n$ DNA sequences randomly sampled from a population are derived from $n-1$ DNA sequences $t$ generations ago and the divergence took place $t-1$ generations ago (see $t$ in Figure 1c) is given by

$$f_{n-1}(t) = \frac{\binom{n}{2}}{2N}\left(1 - \frac{\binom{n}{2}}{2N}\right)^{t-1} \approx \frac{\binom{n}{2}}{2N}\exp\left(-\frac{\binom{n}{2}}{2N}t\right). \quad (20)$$

Following TAJIMA (1983), we have

$$E(k_{AB}) = \sum_{t=1}^{\infty}\sum_{k=0}^{\infty} k f_{n-1}(t)Q(k\,|\,t) = \frac{M}{n(n-1)},$$

$$V(k_{AB}) = \sum_{t=1}^{\infty}\sum_{k=0}^{\infty} k^2 f_{n-1}(t)Q(k\,|\,t) - \{E(k_{AB})\}^2 = \frac{M}{n(n-1)} + \left\{\frac{M}{n(n-1)}\right\}^2,$$

$$\mathrm{Cov}(k_{AB}, k_{AC}) = \sum_{t=1}^{\infty}\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty} k_1 k_2 f_{n-1}(t)Q(k_1\,|\,t)Q(k_2\,|\,t) - \{E(k_{AB})\}^2 = \left\{\frac{M}{n(n-1)}\right\}^2, \quad (21)$$

$$V(k_{BC}) = 2V(k_{AB}) + 2\mathrm{Cov}(k_{AB}, k_{AC}) = \frac{2M}{n(n-1)} + \left\{\frac{2M}{n(n-1)}\right\}^2, \quad (22)$$

where

$$Q(k\,|\,t) = \frac{\exp(-ut)(ut)^k}{k!}.$$

Substituting (21) and (22) into (19), we have

$$\mathrm{Cov}(S, \hat{k}) = \frac{(n+1)(n-2)}{n(n-1)}\mathrm{Cov}(S^*, \hat{k}^*) + \frac{2M}{n(n-1)} + \frac{2M^2}{n(n-1)^2}. \quad (23)$$

Since $\mathrm{Cov}(S, \hat{k})$ is $M + M^2$ when $n$ is 2, we finally have

$$\mathrm{Cov}(S, \hat{k}) = M + \left(\frac{1}{2} + \frac{1}{n}\right)M^2. \quad (24)$$

As $n$ increases, (24) approaches

$$\mathrm{Cov}_{st}(S, \hat{k}) = M + \tfrac{1}{2}M^2. \quad (25)$$

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

$$\mathrm{Cov}_s(S, \hat{k}) = \mathrm{Cov}(S, \hat{k}) - \mathrm{Cov}_{st}(S, \hat{k}) = \frac{1}{n}M^2. \quad (26)$$

The correlation coefficient ($r$) between $S$ and $\hat{k}$ is defined as

$$r = \frac{\mathrm{Cov}(S, \hat{k})}{\sqrt{V(S)V(\hat{k})}}. \quad (27)$$

Numerical computations show that this correlation coefficient is large when the sample size ($n$) is small, and decreases slowly as the sample size increases.

**Difference between the two estimates of $4Nu$:** As mentioned earlier, $M$ ($= 4Nu$) can be estimated from $S$ by using (5), or from $\hat{k}$. In this section we consider the difference between these estimates of $M$.

Let us define $d$ as

$$d = \hat{k} - \frac{S}{a_1}, \tag{28}$$

where $a_1$ is given by (3). Then, the expectation of $d$ is 0 and the variance of $d$ is given by

$$V(d) = V(\hat{k}) - \frac{2}{a_1}\mathrm{Cov}(S, \hat{k}) + \frac{1}{a_1^2}V(S), \tag{29}$$

where $V(\hat{k})$, $V(S)$, and $\mathrm{Cov}(S, \hat{k})$ are given by (7), (2), and (24), respectively. Substituting these quantities into (29), we have

$$V(d) = c_1 M + c_2 M^2, \tag{30}$$

where

$$c_1 = b_1 - \frac{1}{a_1}, \tag{31}$$

and

$$c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}. \tag{32}$$

These equations indicate that, unlike the other variances such as the variances of $S$ and $\hat{k}$, the variance of $d$ increases as $n$ increases and reaches to the asymptotic value which is identical with the variance of $\hat{k}$.

## STATISTICAL METHOD FOR TESTING THE NEUTRAL MUTATION HYPOTHESIS

**Estimating $d$ and $V(d)$:** In the previous section we have obtained the variance of $d$. Formula (30), however, cannot be used directly for estimating the variance of $d$, since we do not know $M$. $M$ can be estimated from $S/a_1$ or $\hat{k}$. We notice from (2) and (7) that the variance of $S/a_1$ is smaller than that of $\hat{k}$ when $n$ is larger than 3. Therefore, $S/a_1$ should be used for estimating $M$ when the neutral mutation hypothesis is correct. Since we assume the neutral mutation hypothesis as a null hypothesis, $M$ is estimated by $S/a_1$ [see (5)]. $(S/a_1)^2$, however, cannot be used for estimating $M^2$, since the expectation of $S^2$ is given by

$$\begin{aligned} E(S^2) &= V(S) + \{E(S)\}^2 \\ &= a_1 M + (a_1^2 + a_2)M^2, \end{aligned} \tag{33}$$

which is not equal to $a_1^2 M^2$. As $E(S^2) - E(S) = (a_1^2 + a_2)M^2$, $M^2$ can be estimated by

$$\frac{S(S-1)}{a_1^2 + a_2}. \tag{34}$$

Therefore, we can estimate $V(d)$ by

$$\hat{V}(d) = e_1 S + e_2 S(S-1), \tag{35}$$

where

$$e_1 = \frac{c_1}{a_1}, \tag{36}$$

and

$$e_2 = \frac{c_2}{a_1^2 + a_2}. \tag{37}$$

**New statistic ($D$):** In order to conduct the statistical test, the following statistic is proposed:

$$D = \frac{d}{\sqrt{\hat{V}(d)}} = \frac{\hat{k} - \frac{S}{a_1}}{\sqrt{e_1 S + e_2 S(S-1)}}, \tag{38}$$

where $a_1$, $e_1$, and $e_2$ are given by (3), (36), and (37).

Then, the mean and variance of $D$ are approximately 0 and 1, respectively. If we know the distribution of $D$, then we can use $D$ in testing the neutral mutation hypothesis. For this purpose the following computer simulation was conducted.

**Computer simulation:** First, genealogical relationships of DNA sequences are generated as follows. When there are $n$ DNA sequences, we randomly choose two DNA sequences among $n$ DNA sequences, combine these two DNA sequences, and obtain new $n-1$ DNA sequences. Figure 2 shows one example of this process. In the case of 5 DNA sequences ($A$, $B$, $C$, $D$, and $E$), if $B$ and $C$ are chosen, we obtain new four DNA sequences ($A$, $BC$, $D$, and $E$). Next, three DNA sequences ($A$, $BC$, and $DE$) are obtained if $D$ and $E$ are chosen. Furthermore, if $A$ and $BC$ are chosen, then we obtain the genealogical relationship of five DNA sequences shown in Figure 2. In this way we can obtain many genealogical relationships of $n$ DNA sequences.

Next, we generate the number of mutations in each branch. Let $S_{in}$ be the number of mutations in the $i$th branch among $n$ branches between $n$ DNA sequences and $n-1$ DNA sequences (Figure 2), and $S_n$ be the total number of mutations in $n$ branches, namely

$$S_n = \sum_{i=1}^{n} S_{in}. \tag{39}$$

Then, $S_n$ follows the geometric distribution,

$$P(S_n) = p_n(1 - p_n)^{S_n}, \tag{40}$$

where

$$p_n = \frac{1}{1 + \frac{M}{n-1}} \tag{41}$$

(see WATTERSON, 1975). The joint probability of $S_{1n}$, $S_{2n}$, . . . , and $S_{nn}$ for a given value of $S_n$ is given by

$$P(S_{1n}, S_{2n}, \ldots, S_{nn} | S_n) = \frac{S_n!}{\prod_{i=1}^{n} S_{in}!}\left(\frac{1}{n}\right)^{S_n}, \tag{42}$$

namely a multinomial distribution. First, we generate $S_n$ according to (40). Then, $S_{in}$'s ($i = 1 \sim n$) are

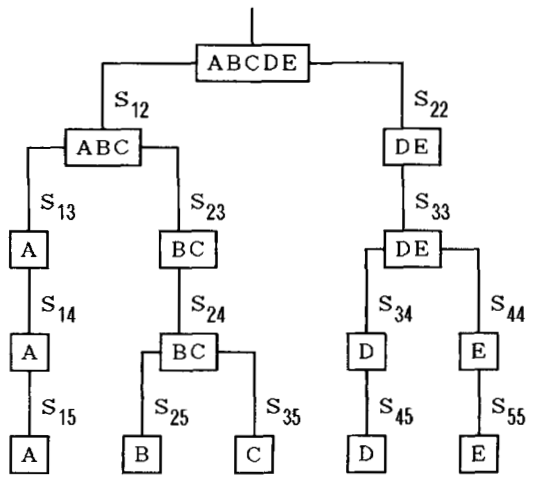


FIGURE 2.—One example of the genealogical relationship among five DNA sequences used for explaining the process of computer simulation.

obtained according to (42). In this way we can get the numbers of mutations in all branches for each genealogical relationship.

Once we have a set of data, we can easily compute the number of segregating sites ($S = S_2 + S_3 + \ldots + S_n$) and the average number of nucleotide differences ($\hat{k}$). Then, we compute $D$ by (38).

In this simulation we used three values of $M$ (1, 10, and 100), and four values of $n$ (5, 10, 20, and 30). In each case we repeated 1000 times. The mean and variance of $D$ in each case are shown in Table 1, and the distribution of $D$ is given in Figure 3. As expected, we can see that the mean of $D$ is nearly zero, although it is negative. The variance of $D$, however, is smaller than 1, especially when $M$ is large. When we conduct a statistical test, this property is not necessarily harmful since it reduces the possibility of rejection. From Figure 3, we can see that the distribution of $D$ is not symmetrical, so that it does not follow the unit normal distribution. For the significant test of neutral mutation hypothesis, however, we can use the unit normal distribution as seen in Table 1. For example, the probability that $D$ is larger than 2 is 0.023 if the unit normal distribution is used. The result obtained from this simulation shows that only in the case of $M = 1$ and $n = 30$ the proportion of $D > 2$ (0.029) is larger than 0.023.

One of the problems in using the unit normal distribution is that the actual values of $D$ can take only limited values. The minimum value of $d$ is obtained when the frequencies of two allelic nucleotides are $1/n$ and $1 - 1/n$ in every segregating site. In this case we obtain

$$\hat{k}_{\min} = \frac{2}{n} S, \tag{43}$$

using (11), so that we have

$$d_{\min} = \hat{k}_{\min} - \frac{S}{a_1} = \left(\frac{2}{n} - \frac{1}{a_1}\right) S. \tag{44}$$

The minimum value of $D$ is obtained when $S$ is infinitely large. From (38) we can obtain

$$D_{\min} = \lim_{S \to \infty} \frac{d_{\min}}{\sqrt{\hat{V}(d)}} = \frac{\frac{2}{n} - \frac{1}{a_1}}{\sqrt{e_2}}. \tag{45}$$

The maximum value can be obtained in the same way as the above, which is given by

$$D_{\max} = \frac{\frac{n}{2(n-1)} - \frac{1}{a_1}}{\sqrt{e_2}}, \tag{46a}$$

when $n$ is an even number, or

$$D_{\max} = \frac{\frac{n+1}{2n} - \frac{1}{a_1}}{\sqrt{e_2}}, \tag{46b}$$

when $n$ is an odd number. One of the basic distributions which often appear in biological study is a beta distribution. Let us consider the beta distribution an approximate distribution of $D$. Since the mean and variance of $D$ are assumed to be 0 and 1, the beta distribution can be written as probability density function:

$$\phi(D) = \frac{\Gamma(\alpha + \beta)(b - D)^{\alpha-1}(D - a)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)(b - a)^{\alpha+\beta-1}}, \tag{47}$$

where

$$\alpha = -\frac{(1 + ab)b}{b - a},$$

$$\beta = \frac{(1 + ab)a}{b - a},$$

$$a = D_{\min}, \quad \text{and} \quad b = D_{\max}.$$

Figure 4 shows the beta distribution, which well agrees with the actual distribution of $D$ obtained from the computer simulation. Table 1 also shows the beta distribution. From this table we can see that the beta distribution fits the actual distribution better than the normal distribution. Because of the above reason, the beta distribution is recommended for testing the neutral mutation hypothesis.

**Test of the neutral mutation hypothesis:** First, we compute $S$ and $\hat{k}$ from the actual data, then obtain $D$ by using (38). Once we have the value of $D$, we can find the confidence limit from Table 2, which is obtained under the assumption that the distribution of $D$ follows the beta distribution given by (47).

TABLE 1

**Comparisons of the distribution of $D$ obtained by computer simulation with the normal and beta distributions**

| $n$ | $M$ | $D$ | | | | | | | Mean | Variance |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $-3 \sim -2$ | $-2 \sim -1$ | $-1 \sim 0$ | $0 \sim 1$ | $1 \sim 2$ | $2 \sim 3$ | $3 \sim 4$ | | |
| 5 | 1 | | 0.154 | 0.393 | 0.181 | 0.272 | | | −0.007 | 0.949 |
| | 10 | | 0.162 | 0.389 | 0.260 | 0.189 | | | −0.016 | 0.813 |
| | 100 | | 0.133 | 0.416 | 0.279 | 0.179 | | | −0.025 | 0.755 |
| | Beta | | 0.222 | 0.324 | 0.236 | 0.218 | | | 0 | 1 |
| 10 | 1 | 0.000 | 0.226 | 0.287 | 0.313 | 0.164 | 0.011 | | −0.036 | 0.941 |
| | 10 | 0.007 | 0.179 | 0.340 | 0.338 | 0.128 | 0.008 | | −0.072 | 0.851 |
| | 100 | 0.002 | 0.165 | 0.386 | 0.325 | 0.117 | 0.005 | | −0.104 | 0.755 |
| | Beta | 0.003 | 0.177 | 0.336 | 0.304 | 0.156 | 0.023 | | 0 | 1 |
| 20 | 1 | 0.004 | 0.212 | 0.321 | 0.296 | 0.153 | 0.014 | 0.000 | −0.050 | 0.959 |
| | 10 | 0.005 | 0.150 | 0.395 | 0.316 | 0.117 | 0.017 | 0.000 | −0.074 | 0.839 |
| | 100 | 0.004 | 0.149 | 0.403 | 0.327 | 0.114 | 0.003 | 0.000 | −0.080 | 0.724 |
| | Beta | 0.011 | 0.161 | 0.342 | 0.315 | 0.146 | 0.025 | 0.000 | 0 | 1 |
| 30 | 1 | 0.002 | 0.171 | 0.354 | 0.298 | 0.145 | 0.028 | 0.001 | −0.002 | 0.977 |
| | 10 | 0.011 | 0.161 | 0.410 | 0.313 | 0.096 | 0.009 | 0.000 | −0.154 | 0.801 |
| | 100 | 0.007 | 0.140 | 0.423 | 0.321 | 0.100 | 0.009 | 0.000 | −0.110 | 0.751 |
| | Beta | 0.012 | 0.157 | 0.345 | 0.317 | 0.142 | 0.026 | 0.001 | 0 | 1 |
| | Normal | 0.021 | 0.136 | 0.341 | 0.341 | 0.136 | 0.021 | 0.001 | 0 | 1 |

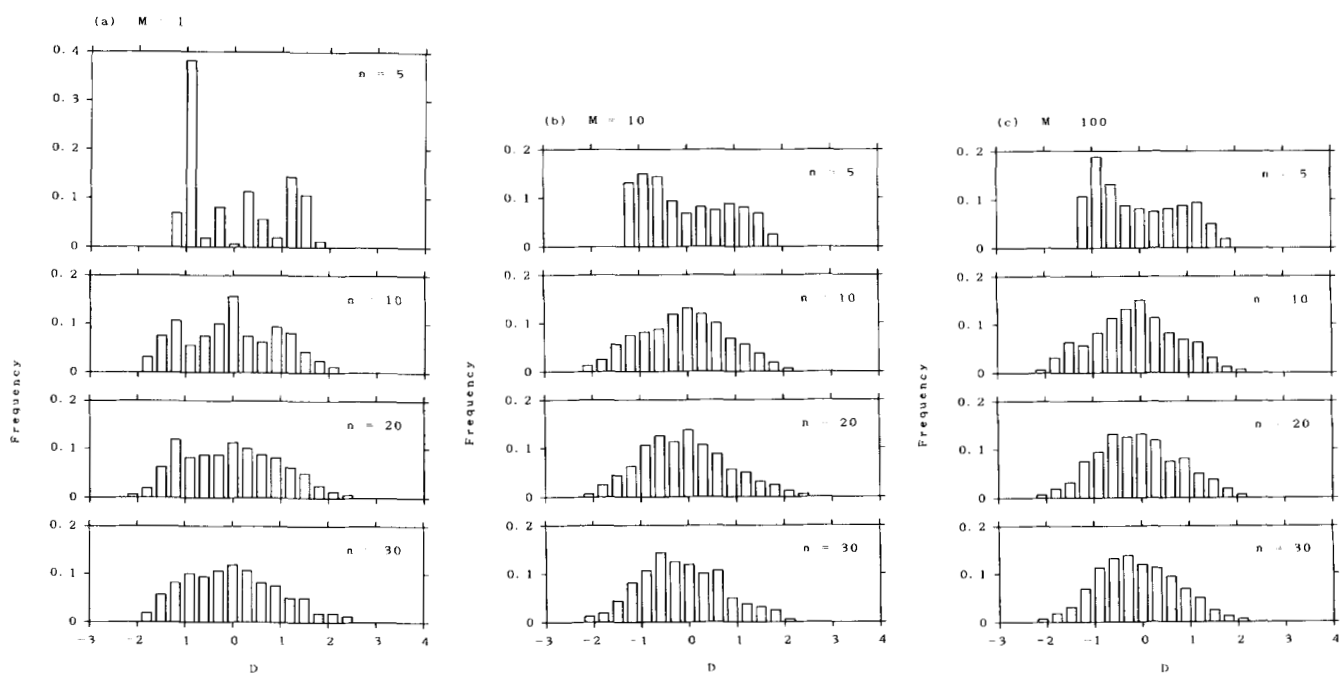


FIGURE 3.—Distributions of $D$ obtained from computer simulation.

## DISTRIBUTION OF NUCLEOTIDE FREQUENCY IN THE SAMPLE

In this section we investigate the distribution of nucleotide frequency in the sample. Consider a given site. If we use the infinite allele model, then the expected number of nucleotides with frequency $(p, p + dp)$ in a population is given by

$$\psi(p)dp = 4N\mu p^{-1}(1 - p)^{4N\mu-1}dp \qquad (48)$$

(KIMURA and CROW 1964), where $\mu$ is the mutation rate per site per generation. Then the expected number of nucleotides with frequency $i/n$ in a sample of $n$ DNA sequences is given by

$$F_n(i) = \int_0^1 \binom{n}{i} p^i(1 - p)^{n-i}\psi(p)dp = \frac{4N\mu\left(\frac{1}{i} + \frac{1}{n-i}\right)}{\prod_{j=n-i}^{n-1}\left(1 + \frac{4N\mu}{j}\right)} \qquad (49)$$

(WATTERSON 1974; TAJIMA 1983), when $1 \leqq i \leqq$

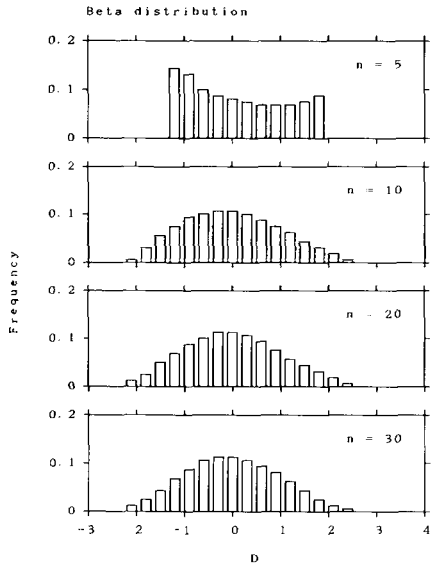


FIGURE 4.—Expected distributions of $D$ obtained by assuming that $D$ follows beta distribution.

$n - 1$. We now assume that there are $m$ sites on the DNA sequence. Then the expected number of nucleotides whose frequency is $i/n$ in a sample of $n$ DNA sequences with $m$ sites can be obtained if we assume $4N\mu m = 4Nu = M$, $\mu \to 0$, and $m \to \infty$, and is given by

$$G_n(i) = M\left(\frac{1}{i} + \frac{1}{n-i}\right), \tag{50}$$

when $1 \leqq i \leqq n - 1$. If we use $S/a_1$ instead of $M$, then we have

$$G_n(i) = \frac{S\left(\frac{1}{i} + \frac{1}{n-i}\right)}{\sum_{j=1}^{n-1} \frac{1}{j}}. \tag{51}$$

Incidentally the sum of $G_n(i)$ for $i = 1$ to $n - 1$ is $2S$, since there are two allelic nucleotides in each segregating site. Using (51), we can compare the observed distribution of nucleotide frequency with the expected one, although we cannot conduct a significant test by using this comparison.

## NUMERICAL EXAMPLE

AQUADRO and GREENBERG (1983) studied a sequence of about 900 nucleotide pairs of the human mitochondrial DNA for seven individuals ($n = 7$). The number of segregating sites ($S$) is 45, and the average number of nucleotide differences ($\hat{k}$) estimated was 15.38. In this case the values of $a_1$, $e_1$, and $e_2$ are 2.4500, 0.01481, and 0.004784, respectively. Using (38), we obtain $D = -0.9382$, which is not significantly different from 0 (Table 2), so that we conclude that the neutral mutation hypothesis can explain the DNA polymorphism of human mitochondrial DNA.

Incidentally, the distribution of nucleotide frequency in the sample is shown in Figure 5, which indicates that the numbers of nucleotides with frequencies $1/7$ and $6/7$ are larger than the expected ones.

Miyashita and Langley (1988) examined a 45-kb region of the *white* locus on the $X$ chromosome in *Drosophila melanogaster*, using 64 $X$ chromosome lines ($n = 64$) with six 6-cutter and ten 4-cutter restriction enzymes. They classified the DNA polymorphisms into three groups, namely restriction site polymorphism, small insertion/deletion (<100 bp) polymorphism, and large insertion/deletion (>100 bp) polymorphism. As long as the infinite site mutation model is applicable, we can use this method. In the cases of restriction site and insertion/deletion, there are many sites where mutations can take place. Therefore, we can apply the present test to these cases. Unlike the mitochondrial DNA, however, some recombination may occur on the nuclear DNA. In this case the actual variance of $d$ is smaller than that of (30), so that the actual value of $D$ takes more extreme value than that estimated by (38). Because of this, the present method might be conservative when the nuclear DNA is analysed. The result of the present test is shown in Table 3. Significant deviation of $D$ from 0 is observed ($P < 0.05$) only in the case of large insertion/deletion polymorphism, so that we can reject the null hypothesis that all the large insertion/deletion polymorphisms are maintained without selection at the 5% level.

One of the possible explanations of this is that the large insertions/deletions are deleterious so that they are maintained with low frequency. This can be seen in Figure 6. In this figure only the numbers of segregating sites whose frequencies are less than or equal to $32/64$ are shown, since the number of segregating sites with frequency $i/n$ is equal to that of $(n - i)/n$. From this figure we can see that the numbers of segregating sites with low frequencies are much larger than the expected ones.

Another possible explanation is that the population does not reach to the equilibrium yet. For example, if a population experienced a bottleneck recently, many sites with low frequency might be observed. Therefore, $D$ is expected to be negative. Since the values of $D$ observed in the cases of restriction site and small insertion/deletion polymorphisms are positive, the bottleneck effect cannot explain the result that the value of $D$ observed in the case of large insertion/deletion is significantly smaller than 0.

Recently, T. TAKANO, S. KUSAKABE and T. MUKAI (in preparation) have obtained the same result. They studied the regions of *Adh, Amy, Pu* (Punch), and *Gpdh* in *Drosophila melanogaster* by using eight 6-cutter restriction enzymes. Eighty-six DNA sequences ($n = 86$) were collected from two Japanese populations (Aomori and Ogasawara populations). Their results, which are shown in Table 4, indicate that the value

**TABLE 2**

**Confidence limit of $D$ obtained by assuming the beta distribution**

| $n$ | Confidence limit of $D$ | | | |
|---|---|---|---|---|
| | 90% | 95% | 99% | 99.9% |
| 4 | −0.876 ~ 2.081 | −0.876 ~ 2.232 | −0.876 ~ 2.324 | −0.876 ~ 2.336 |
| 5 | −1.255 ~ 1.737 | −1.269 ~ 1.834 | −1.275 ~ 1.901 | −1.276 ~ 1.913 |
| 6 | −1.405 ~ 1.786 | −1.478 ~ 1.999 | −1.540 ~ 2.255 | −1.556 ~ 2.373 |
| 7 | −1.498 ~ 1.728 | −1.608 ~ 1.932 | −1.721 ~ 2.185 | −1.761 ~ 2.311 |
| 8 | −1.522 ~ 1.736 | −1.663 ~ 1.975 | −1.830 ~ 2.313 | −1.909 ~ 2.524 |
| 9 | −1.553 ~ 1.715 | −1.713 ~ 1.954 | −1.916 ~ 2.296 | −2.023 ~ 2.519 |
| 10 | −1.559 ~ 1.719 | −1.733 ~ 1.975 | −1.967 ~ 2.362 | −2.105 ~ 2.640 |
| 11 | −1.572 ~ 1.710 | −1.757 ~ 1.966 | −2.014 ~ 2.359 | −2.174 ~ 2.649 |
| 12 | −1.573 ~ 1.713 | −1.765 ~ 1.979 | −2.041 ~ 2.401 | −2.223 ~ 2.729 |
| 13 | −1.580 ~ 1.708 | −1.779 ~ 1.976 | −2.069 ~ 2.403 | −2.267 ~ 2.741 |
| 14 | −1.580 ~ 1.710 | −1.783 ~ 1.985 | −2.085 ~ 2.432 | −2.299 ~ 2.798 |
| 15 | −1.584 ~ 1.708 | −1.791 ~ 1.984 | −2.103 ~ 2.436 | −2.329 ~ 2.811 |
| 16 | −1.583 ~ 1.709 | −1.793 ~ 1.990 | −2.113 ~ 2.457 | −2.350 ~ 2.854 |
| 17 | −1.585 ~ 1.708 | −1.798 ~ 1.990 | −2.126 ~ 2.461 | −2.372 ~ 2.866 |
| 18 | −1.584 ~ 1.709 | −1.799 ~ 1.996 | −2.132 ~ 2.478 | −2.387 ~ 2.900 |
| 19 | −1.585 ~ 1.708 | −1.802 ~ 1.996 | −2.141 ~ 2.483 | −2.403 ~ 2.911 |
| 20 | −1.584 ~ 1.710 | −1.803 ~ 2.001 | −2.146 ~ 2.496 | −2.414 ~ 2.939 |
| 21 | −1.585 ~ 1.709 | −1.805 ~ 2.001 | −2.152 ~ 2.501 | −2.426 ~ 2.950 |
| 22 | −1.584 ~ 1.711 | −1.804 ~ 2.005 | −2.153 ~ 2.512 | −2.434 ~ 2.973 |
| 23 | −1.584 ~ 1.710 | −1.806 ~ 2.006 | −2.160 ~ 2.516 | −2.443 ~ 2.983 |
| 24 | −1.583 ~ 1.712 | −1.806 ~ 2.009 | −2.162 ~ 2.526 | −2.449 ~ 3.002 |
| 25 | −1.583 ~ 1.712 | −1.807 ~ 2.010 | −2.165 ~ 2.530 | −2.457 ~ 3.011 |
| 26 | −1.582 ~ 1.712 | −1.807 ~ 2.013 | −2.167 ~ 2.538 | −2.461 ~ 3.029 |
| 27 | −1.582 ~ 1.712 | −1.807 ~ 2.014 | −2.170 ~ 2.542 | −2.467 ~ 3.037 |
| 28 | −1.581 ~ 1.713 | −1.807 ~ 2.017 | −2.171 ~ 2.549 | −2.471 ~ 3.052 |
| 29 | −1.581 ~ 1.714 | −1.807 ~ 2.018 | −2.173 ~ 2.553 | −2.475 ~ 3.060 |
| 30 | −1.580 ~ 1.714 | −1.807 ~ 2.020 | −2.173 ~ 2.559 | −2.478 ~ 3.073 |
| 31 | −1.580 ~ 1.714 | −1.807 ~ 2.021 | −2.175 ~ 2.563 | −2.482 ~ 3.080 |
| 32 | −1.579 ~ 1.715 | −1.806 − 2.023 | −2.175 ~ 2.569 | −2.484 ~ 3.092 |
| 33 | −1.579 ~ 1.716 | −1.806 ~ 2.024 | −2.177 ~ 2.572 | −2.487 ~ 3.099 |
| 34 | −1.578 ~ 1.716 | −1.806 ~ 2.026 | −2.177 ~ 2.577 | −2.489 ~ 3.110 |
| 35 | −1.578 ~ 1.717 | −1.806 ~ 2.027 | −2.178 ~ 2.580 | −2.492 ~ 3.116 |
| 36 | −1.577 ~ 1.717 | −1.805 ~ 2.029 | −2.178 ~ 2.585 | −2.493 ~ 3.126 |
| 37 | −1.577 ~ 1.717 | −1.805 ~ 2.030 | −2.179 ~ 2.588 | −2.495 ~ 3.132 |
| 38 | −1.576 ~ 1.718 | −1.804 ~ 2.031 | −2.178 ~ 2.592 | −2.496 ~ 3.141 |
| 39 | −1.576 ~ 1.718 | −1.804 ~ 2.032 | −2.179 ~ 2.595 | −2.498 ~ 3.147 |
| 40 | −1.575 ~ 1.719 | −1.804 ~ 2.033 | −2.179 ~ 2.599 | −2.499 ~ 3.155 |
| 41 | −1.575 ~ 1.719 | −1.803 ~ 2.034 | −2.179 ~ 2.601 | −2.500 ~ 3.160 |
| 42 | −1.574 ~ 1.720 | −1.803 ~ 2.036 | −2.179 ~ 2.605 | −2.501 ~ 3.168 |
| 43 | −1.574 ~ 1.720 | −1.803 ~ 2.037 | −2.179 ~ 2.608 | −2.502 ~ 3.173 |
| 44 | −1.573 ~ 1.721 | −1.802 ~ 2.038 | −2.179 ~ 2.611 | −2.502 ~ 3.180 |
| 45 | −1.573 ~ 1.721 | −1.802 ~ 2.039 | −2.179 ~ 2.613 | −2.503 ~ 3.185 |
| 46 | −1.572 ~ 1.721 | −1.801 ~ 2.040 | −2.179 ~ 2.617 | −2.504 ~ 3.191 |
| 47 | −1.572 ~ 1.722 | −1.801 ~ 2.041 | −2.179 ~ 2.619 | −2.504 ~ 3.196 |
| 48 | −1.571 ~ 1.722 | −1.800 ~ 2.042 | −2.178 ~ 2.622 | −2.505 ~ 3.202 |
| 49 | −1.571 ~ 1.722 | −1.800 ~ 2.042 | −2.178 ~ 2.624 | −2.505 ~ 3.207 |
| 50 | −1.570 ~ 1.723 | −1.800 ~ 2.044 | −2.178 ~ 2.627 | −2.505 ~ 3.212 |
| 55 | −1.568 ~ 1.724 | −1.797 ~ 2.048 | −2.177 ~ 2.638 | −2.506 ~ 3.235 |
| 60 | −1.566 ~ 1.726 | −1.795 ~ 2.052 | −2.175 ~ 2.649 | −2.506 ~ 3.256 |
| 65 | −1.565 ~ 1.727 | −1.793 ~ 2.055 | −2.173 ~ 2.658 | −2.506 ~ 3.274 |
| 70 | −1.563 ~ 1.729 | −1.791 ~ 2.058 | −2.171 ~ 2.666 | −2.505 ~ 3.291 |
| 75 | −1.561 ~ 1.730 | −1.790 ~ 2.061 | −2.170 ~ 2.673 | −2.504 ~ 3.306 |
| 80 | −1.560 ~ 1.731 | −1.788 ~ 2.064 | −2.168 ~ 2.681 | −2.502 ~ 3.320 |
| 85 | −1.559 ~ 1.732 | −1.786 ~ 2.066 | −2.166 ~ 2.687 | −2.500 ~ 3.333 |
| 90 | −1.557 ~ 1.733 | −1.784 ~ 2.069 | −2.164 ~ 2.693 | −2.499 ~ 3.345 |
| 95 | −1.556 ~ 1.734 | −1.783 ~ 2.071 | −2.162 ~ 2.699 | −2.497 ~ 3.355 |
| 100 | −1.555 ~ 1.735 | −1.781 ~ 2.073 | −2.160 ~ 2.704 | −2.495 ~ 3.366 |
| 110 | −1.552 ~ 1.737 | −1.779 ~ 2.077 | −2.157 ~ 2.713 | −2.492 ~ 3.385 |
| 120 | −1.550 ~ 1.739 | −1.776 ~ 2.080 | −2.153 ~ 2.722 | −2.488 ~ 3.401 |
| 130 | −1.549 ~ 1.740 | −1.774 ~ 2.084 | −2.150 ~ 2.730 | −2.484 ~ 3.416 |
| 140 | −1.547 ~ 1.741 | −1.771 ~ 2.086 | −2.147 ~ 2.736 | −2.481 ~ 3.430 |

**TABLE 2—Continued**

| $n$ | Confidence limit of $D$ | | | |
|---|---|---|---|---|
| | 90% | 95% | 99% | 99.9% |
| 150 | −1.545 ~ 1.743 | −1.769 ~ 2.089 | −2.144 ~ 2.743 | −2.477 ~ 3.443 |
| 175 | −1.542 ~ 1.746 | −1.765 ~ 2.095 | −2.138 ~ 2.757 | −2.470 ~ 3.470 |
| 200 | −1.539 ~ 1.748 | −1.760 ~ 2.100 | −2.132 ~ 2.768 | −2.462 ~ 3.492 |
| 250 | −1.534 ~ 1.752 | −1.754 ~ 2.107 | −2.122 ~ 2.787 | −2.449 ~ 3.529 |
| 300 | −1.530 ~ 1.755 | −1.748 ~ 2.114 | −2.114 ~ 2.802 | −2.439 ~ 3.558 |
| 350 | −1.526 ~ 1.757 | −1.744 ~ 2.119 | −2.107 ~ 2.814 | −2.430 ~ 3.581 |
| 400 | −1.523 ~ 1.759 | −1.740 ~ 2.123 | −2.101 ~ 2.824 | −2.422 ~ 3.600 |
| 450 | −1.521 ~ 1.761 | −1.737 ~ 2.127 | −2.096 ~ 2.833 | −2.415 ~ 3.617 |
| 500 | −1.519 ~ 1.763 | −1.734 ~ 2.130 | −2.092 ~ 2.840 | −2.409 ~ 3.632 |
| 600 | −1.515 ~ 1.765 | −1.728 ~ 2.135 | −2.084 ~ 2.853 | −2.398 ~ 3.657 |
| 800 | −1.510 ~ 1.769 | −1.721 ~ 2.143 | −2.072 ~ 2.873 | −2.382 ~ 3.694 |
| 1000 | −1.505 ~ 1.772 | −1.715 ~ 2.150 | −2.062 ~ 2.887 | −2.369 ~ 3.722 |

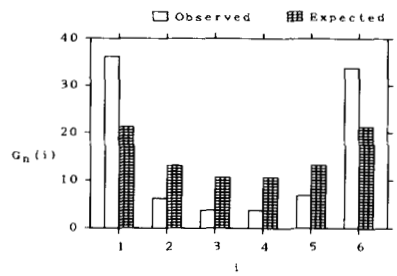


FIGURE 5.—Observed and expected distributions of the number of allelic nucleotides for human mitochondrial DNA. The observed distribution was obtained from AQUADRO and GREENBERG (1983), and the expected distribution was obtained by assuming the neutral mutation model.

**TABLE 3**

**Estimates of $D$ for the three groups of polymorphisms in the *white* locus in *D. melanogaster***

| Type of polymorphism | $S$ | $\hat{k}$ | $D$ |
|---|---|---|---|
| Restriction site | 53 | 11.92 | 0.2128 (NS) |
| Small insertion/deletion | 40 | 10.02 | 0.6075 (NS) |
| Large insertion/deletion | 15 | 0.94 | −2.0709 ($P < 0.05$) |

Data from MIYASHITA and LANGLEY (1988). NS, not significant ($P > 0.1$).

of $D$ in the case of restriction site polymorphism does not show a significant deviation from 0, but that of insertion/deletion (>300 bp) polymorphism shows a significant deviation in the case of *Amy*. If we pool the data of four regions of DNA, the deviation of $D$ from 0 becomes highly significant ($P < 0.01$). In this case the value of $D$ was obtained by the sum of the values of $d$ divided by the square root of the sum of the estimated variances of $d$, since these four regions can be assumed to be unlinked. The distribution of the sum of independent random variables approaches the normal distribution as the number of variables increases, and the computer simulation conducted earlier indicates that the distribution of $D$ is not far from the normal distribution. Because of these reasons, in order to find the confidence limit of $D$, we can use the normal distribution when several regions of DNA are used. At any rate, if we apply the unit normal distribution in this case, the deviation of $D$ from 0 is highly significant ($P < 0.01$), so that the neutral mutation hypothesis is rejected.

## DISCUSSION

In this paper we have obtained a statistical method for testing the neutral mutation hypothesis by using DNA polymorphism. Unlike HUDSON, KREITMAN and AGUADÉ (1987) where not only DNA polymorphism data but also between species divergence data are necessary, only DNA polymorphism data are needed to use this method. In many cases only DNA polymorphism data are available, so that this method might be useful. When we apply this method, however, some caution is necessary. (1) The DNA sequences applied to this method must be a random sample from a population. (2) We must take into consideration whether the population is at equilibrium or not. For example, as shown in the NUMERICAL EXAMPLE section, a negative value of $D$ can also be obtained if the population experienced a bottleneck recently. In this case a comparison between different kinds of DNA polymorphisms such as a comparison between nucleotide and insertion/deletion polymorphisms may help our interpretation, since a bottleneck affects all kinds of DNA polymorphisms. (3) If a selectively neutral site is linked to a site at which natural selection is operating, then the value of $D$ for the neutral site might be affected by the selected site. [For the coalescent process for a neutral site which is linked to a selected site, see KAPLAN, DARDEN and HUDSON (1988) and HUDSON and KAPLAN (1988).]

In the NUMERICAL EXAMPLE section, we analyzed the five regions of DNA sequences, *white, Adh, Amy, Pu*, and *Gpdh*. All the values of $D$ for the restriction site polymorphism were positive (Tables 3 and 4).

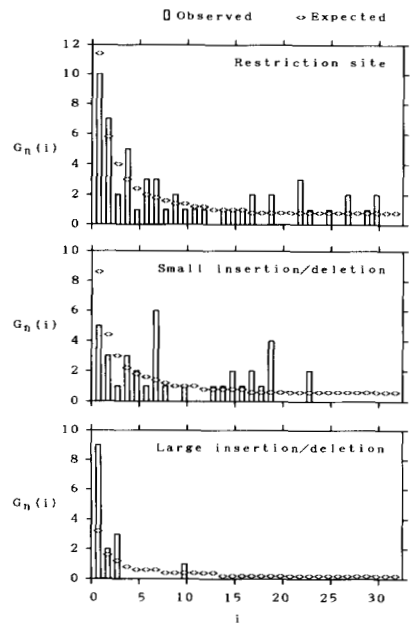


FIGURE 6.—Observed and expected frequency spectrums of polymorphic variation in the *white* locus region of *D. melanogaster*. The observed spectrums were obtained from MIYASHITA and LANGLEY (1988), and the expected spectrums were obtained by assuming the neutral mutation model.

Under the neutral mutation hypothesis, the probability that $D$ is positive is less than ½ (see Table 1), so that the probability that all the five $D$ values are positive is less than 1/32. Therefore, there may be a site at which natural selection, which increases the genetic variation, is operating. Among the five values of $D$, the value of $D$ for *Adh* is quite large, although it is not significantly different from 0. The exceptionally high level of variation was observed in the *Adh* coding region by KREITMAN and AGUADÉ (1986), so that the large $D$ value may be explained by natural selection.

Negative values of $D$ were observed in the case of large insertion/deletion for all five regions of DNA. [The length of insertion/deletion in T. TAKANO, S. KUSAKABE and T. MUKAI (in preparation) is longer than 300 bp, so that we call it large insertion/deletion according to MIYASHITA and LANGLEY (1988).] Under the deleterious mutation model, we can estimate the total number of deleterious mutants per DNA sequence (or per genome).

Let $q_i$ be the frequency of deleterious nucleotide in the $i$th deleterious site in a population. If we consider the deleterious site as well as the neutral site, then the expectation of the number of segregating sites for a sample of $n$ DNA sequences is given by

$$E(S) = a_1 M + \sum [1 - q_i^n - (1 - q_i)^n].$$

If we assume that $q_i$ is very small, then $E(S)$ is approximately given by

$$E(S) = a_1 M + n \sum q_i. \tag{52}$$

TABLE 4

**Estimates of $D$ for four regions of DNA in *D. melanogaster***

| | Restriction site | | | Insertion/deletion | | |
|---|---|---|---|---|---|---|
| Region | $S$ | $\hat{k}$ | $D$ | $S$ | $\hat{k}$ | $D$ |
| *Adh* | 4 | 1.39 | 1.520 (NS) | 10 | 0.82 | −1.537 (NS) |
| *Amy* | 7 | 1.74 | 0.599 (NS) | 10 | 0.59 | −1.839 ($P < 0.05$) |
| *Pu* | 6 | 1.25 | 0.108 (NS) | 2 | 0.07 | −1.305 (NS) |
| *Gpdh* | 18 | 4.27 | 0.559 (NS) | 15 | 1.10 | −1.784 ($P < 0.1$) |
| Sum | 35 | 8.64 | 1.111 (NS) | 37 | 2.58 | −3.127 ($P < 0.01$) |

Data from T. TAKANO, S. KUSAKABE and T. MUKAI (in preparation). NS, not significant ($P > 0.1$).

On the other hand, the expectation of the average number of nucleotide differences is given by

$$E(\hat{k}) = M + \sum 2q_i(1 - q_i) \approx M + 2 \sum q_i. \tag{53}$$

If we define $d$ by (28) as before, then the expectation of $d$ becomes

$$E(d) = \left(2 - \frac{n}{a_1}\right) \sum q_i. \tag{54}$$

Therefore, the total number of deleterious mutants per DNA sequence ($\sum q_i$) can be estimated by

$$Q = -\frac{d}{\frac{n}{a_1} - 2}. \tag{55}$$

The estimates ($Q$) of $\sum q_i$ for large insertion/deletion polymorphism in the five regions of DNA are shown in Table 5. In this case $Q$ is the estimate of the total number of deleterious insertions/deletions per DNA sequence under investigation. If we sum up all the five regions, then $Q$ becomes 0.510 per 107 kb. This means that on the average there is one deleterious insertion/deletion every 200 kb. If this estimate is correct for the whole regions of DNA, then it is expected that on the average there are 700 deleterious insertions/deletions per genome, assuming $1.4 \times 10^8$ bp per genome (LEWIN 1975). In order to explain this large value, we need to assume not only high mutation rate but also weak selection. LEIGH BROWN (1983) and AQUADRO *et al.* (1986) suggest that the majority of large insertions are caused by transposable elements. If this is the case, then the mutation rate of deleterious insertion/deletion might be high. If we assume that the mutant is maintained by the mutation-selection balance and that the selection coefficient of heterozygote is $hs$ and the mutation rate of deleterious insertion/deletion is $v$, then the equilibrium frequency is $v/hs$. If the mutation rate per genome is 0.01, the selection coefficient must be $1.5 \times 10^{-5}$, assuming that the selection coefficient is the same for all sites. If the mutation rate is as high as 0.1, then the selection

TABLE 5

**Estimated numbers ($Q$) of deleterious insertions/deletions for the five regions of DNA**

| Region | Length of DNA examined (kb) | $d$ | $Q$ | $Q$/kb |
|---|---|---|---|---|
| *white* | 45 | −2.228 | 0.193 | 0.0043 |
| *Adh* | 11 | −1.170 | 0.077 | 0.0070 |
| *Amy* | 14 | −1.400 | 0.093 | 0.0066 |
| *Pu* | 14 | −0.328 | 0.022 | 0.0016 |
| *Gpdh* | 23 | −1.885 | 0.125 | 0.0054 |
| Sum | 107 | | 0.510 | 0.0048 |

Data for *white* from MIYASHITA and LANGLEY (1988). Data for the others from T. TAKANO, S. KUSAKABE and T. MUKAI (in preparation).

coefficient is 0.00015. OHTA (1973, 1974) has proposed the very slightly deleterious or nearly neutral mutation hypothesis, and this hypothesis has been further developed by KIMURA (1979). The large insertion/deletion seems to support this hypothesis.

I thank T. MUKAI, T. TAKANO and S. KUSAKABE for allowing me to use their unpublished data. I also thank T. OHTA, B. S. WEIR, and two anonymous reviewers for their 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.

AQUADRO, C. F., S. F. DEESE, M. M. BLAND, C. H. LANGLEY and C. C. LAURIE-AHLBERG, 1986 Molecular population genetics of alcohol dehydrogenase gene region of *Drosophila melanogaster*. Genetics **114:** 1165–1190.

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

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

KAPLAN, N. L., T. DARDEN and R. R. HUDSON, 1988 The coalescent process in models with selection. Genetics **120:** 819–829.

KIMURA, M., 1968 Evolutionary rate at the molecular level. Nature **217:** 624–626.

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

KIMURA, M., 1979 Model of effectively neutral mutations in which selection constraint is incorporated. Proc. Natl. Acad. Sci. USA **76:** 3440–3444.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, London.

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

KREITMAN, M., and M. AGUADÉ, 1986 Excess polymorphism at the *Adh* locus in *Drosophila melanogaster*. Genetics **114:** 93–110.

LEIGH BROWN, A. J., 1983 Variation at the 87A heat-shock locus in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **80:** 5350–5354.

LEWIN, B., 1975 Units of transcription and translation: sequence components of heterogeneous nuclear RNA and messenger RNA. Cell **4:** 77–93.

MIYASHITA, N., and C. H. LANGLEY, 1988 Molecular and phenotypic variation of the *white* locus region in *Drosophila melanogaster*. Genetics **120:** 199–212.

OHTA, T., 1973 Slightly deleterious mutant substitutions in evolution. Nature **246:** 96–98.

OHTA, T., 1974 Mutational pressure as the main cause of molecular evolution and polymorphism. Nature **252:** 351–354.

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

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

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

Communicating editor: B. S. WEIR