Relationship Between DNA Polymorphism and Fixation Time

Fumio Tajima

National Institute of Genetics, Mishima, 411 Japan

Manuscript received November 13, 1989
Accepted for publication February 26, 1990

ABSTRACT

When there is no recombination among nucleotide sites in DNA sequences, DNA polymorphism and fixation of mutants at nucleotide sites are mutually related. Using the method of gene genealogy, the relationship between the DNA polymorphism and the fixation of mutant nucleotide was quantitatively investigated under the assumption that mutants are selectively neutral, that there is no recombination among nucleotide sites, and that the population is a random mating population with N diploid individuals. The results obtained indicate that the expected number of nucleotide differences between two DNA sequences randomly sampled from the population is 42% less when a mutant at a particular nucleotide site reaches fixation than at a random time, and that heterozygosity is also expected to be less when fixation takes place than at a random time, but the amount of reduction depends on the value of $4Nv$ in this case, where v is the mutation rate per DNA sequence per generation. The formula for obtaining the expected number of nucleotide differences between the two DNA sequences for a given fixation time is also derived, and indicates that, even when it takes a large number of generations for a mutant to reach fixation, this number is 33% less than at a random time. The computer simulation conducted suggests that the expected number of nucleotide differences between the two DNA sequences at the time when an advantageous mutant becomes fixed is essentially the same as that of neutral mutant if the fixation time is the same. The effect of recombination on the amount of DNA polymorphism was also investigated by using computer simulation.

DNA polymorphism and fixation of mutants at nucleotide sites are not independent phenomena, but they are mutually related. For example, the fixation of selectively advantageous allele at one locus can reduce the amount of polymorphism at linked locus (Kojima and Schaeffer 1967; Maynard Smith and Haigh 1974; Ohta and Kimura 1975). Recently Kaplan, Hudson and Langley (1989) have examined this hitchhiking effect at the DNA level, and concluded that in the region of low crossing over the fixation of selectively advantageous mutant at one nucleotide site can substantially reduce the number of segregating or polymorphic nucleotide sites in a sample of DNA sequences from that expected under the neutral mutation model.

The effect of the fixation of a mutant on the amount of polymorphism may occur even when the mutant is selectively neutral. Watterson (1982a,b) has shown under the neutral mutation model that fixations tend to occur in clusters rather than behave as a Poisson process. This suggests that there might be some effect of fixation on DNA polymorphism even if all the mutants are selectively neutral.

The purpose of this paper is to examine quantitatively the relationship between the DNA polymorphism and the fixation of a mutant nucleotide.

The amount of DNA polymorphism can be measured by the average number of (pairwise) nucleotide differences among a sample of DNA sequences or by the number of segregating (or polymorphic) sites in a sample of DNA sequences [for their statistical properties under the neutral mutation model, see Watterson (1975) and Tajima (1983)]. In this paper we use the expected number of nucleotide differences between two DNA sequences randomly sampled from a population as a measure of the amount of DNA polymorphism. This number equals not only the expectation of the average number of nucleotide differences among a sample of DNA sequences but also the expected number of segregating sites in a sample of two DNA sequences.

The fixation time is one of the most important quantities that characterize the fixation. In this paper we study the relationship between the amount of DNA polymorphism at the time when a mutant at a particular nucleotide site has fixed and the fixation time for this mutant.

THEORY

Assumptions: Consider a random mating population with N diploid individuals, so that there are $2N$ homologous DNA sequences in the population. Assume that each DNA sequence has infinitely many nucleotide sites (Kimura 1969), and that there is no recombination between them. Also assume that newly arisen mutations are selectively neutral (Kimura

1968, 1983), and that they occur at the rate of v per DNA sequence per generation.

Fixation time: Using the diffusion model, KIMURA (1970) has obtained the probability distribution, $y(t)$, of the number of generations until a newly arisen mutant becomes fixed in the population, excluding the cases where the mutant is lost from it, which is given by

$$y(t) = \sum_{i=1}^{\infty}(2i + 1)(-1)^{i+1}\lambda_i\exp(-\lambda_i t), \tag{1}$$

where $t > 0$ and $\lambda_i = i(i + 1)/(4N)$. Let us obtain this probability, using the method of gene genealogy or coalescent process, since this method directly gives the relationship between fixation and polymorphism as will be shown later.

In order to study branch length and branching pattern, we use the Wright-Fisher model with nonoverlapping generations and Moran's model, respectively. This is because branching pattern can be more easily studied by using Moran's model than the Wright-Fisher model. Needless to say, the mixed use of different models may not be desirable. However, some quantities obtained from the Wright-Fisher model are known to be approximately the same as those of Moran's model although certain changes of definitions, *e.g.*, time scaling and the effective population size, are necessary (KINGMAN 1982a,b; WATTERSON 1984). Furthermore, as will be shown later, computer simulations conducted indicate that this treatment does not cause any serious error.

Let us now consider the genealogical relationship of DNA sequences when the fixation takes place. We assume that at each unit of time one of the $2N$ DNA sequences is randomly chosen to die, and it is replaced by a replicate of a randomly chosen sequence from the remaining $2N - 1$ DNA sequences. This model is called MORAN's (1958) model, though it is slightly different from Moran's original model [see WATTERSON (1982a)]. One example of the birth and death process in Moran's model is shown in Figure 1 where $2N = 4$ is assumed. In this example three events of fixation are possible, as indicated by arrows. This can be explained in terms of gene genealogy. First, $2N$ DNA sequences from the present population, which are assumed to be mutants, came from $2N - 1$ DNA sequences at the immediately previous time. At this time the remaining one DNA sequence is not the mutant. Then, these $2N - 1$ mutant DNA sequences have the original mutant DNA sequence as common ancestor at some time in the past, and the remaining one nonmutant DNA sequence has an ancestor at this time which is a nonmutant DNA sequence. Figure 1 shows the three genealogical relationships where the fixations can take place in this example. The distribution of the genealogical relationship under the condition of fixation is different from that without any condition, and this difference is caused by branching pattern, but not by branch length, since there is no restriction on branch length. Because of this, the probability distribution of fixation time can be easily obtained.

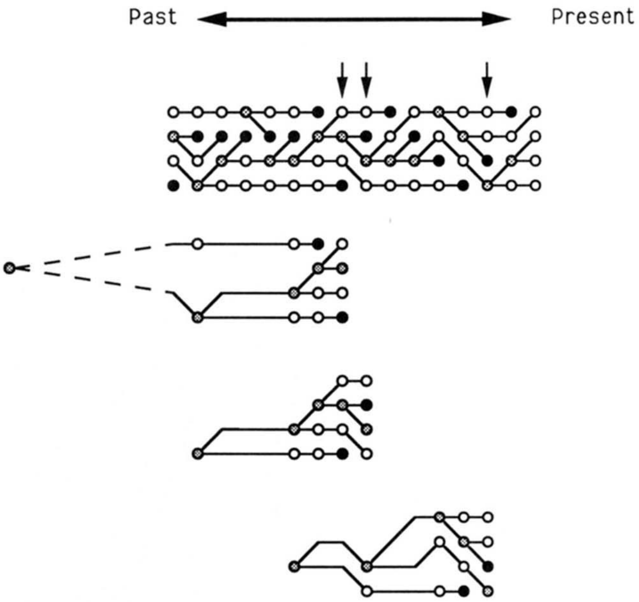

FIGURE 1.—One example of the birth-death process in Moran's model. The arrows in the top figure indicate the time when fixation can take place. The bottom three figures show the genealogical relationships for the three possible events of fixation.

Let $f_n(t)$ be the probability that $n + 1$ DNA sequences randomly chosen from the population are derived from n DNA sequences for the first time t generations ago. Using the Wright-Fisher model with nonoverlapping generations, KINGMAN (1982a), HUDSON (1983) and TAJIMA (1983) have shown that $f_n(t)$ is approximately given by

$$f_n(t) = \lambda_n\exp(-\lambda_n t), \tag{2}$$

where the mean and variance are $1/\lambda_n$ and $1/\lambda_n^2$, respectively. Then, the probability distribution, $y(t)$, of the number of generations until fixation can be obtained from the convolution of $f_{2N-1}(t)$, $f_{2N-2}(t)$, $\ldots, f_1(t)$, namely

$$y(t) = \sum_{i=1}^{2N-1} (2i + 1)(-1)^{i+1}\prod_{j=1}^{i} \frac{2N - j}{2N + j}\lambda_i\exp(-\lambda_i t). \tag{3}$$

This equation can be approximately given by

$$y(t) = \sum_{i=1}^{2N-1} (2i + 1)(-1)^{i+1}\lambda_i\exp[-\lambda_i(t + 2)], \tag{4}$$

which is essentially the same as (1).

DNA polymorphism: WATTERSON (1975) showed that the expected number, $E(k)$, of nucleotide differences between two DNA sequences randomly sampled from the population is given by

$$E(k) = M, \tag{5}$$

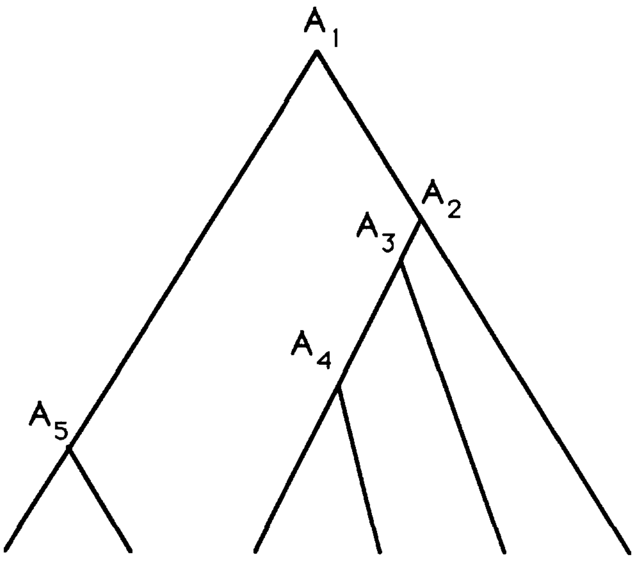

FIGURE 2.—One example of genealogical relationship among six DNA sequences. A_i is the ith oldest ancestor to these six sequences.

where $M = 4Nv$. This number can easily be obtained by considering gene genealogies. The probability that two randomly chosen DNA sequences are derived from their common ancestral sequence for the first time t generations ago is $f_1(t)$, which gives the probability distribution of branch length between one of the chosen DNA sequences and the common ancestral sequence in terms of the number of generations, so that the mean branch length is $1/\lambda_1$ or $2N$. Since there are two branches between the two sequences randomly chosen from the population, we have

$$E(k) = 2N \times 2v = M.$$

This equation can also be obtained in a different way. Consider the genealogical relationship among $2N$ DNA sequences. When two DNA sequences are randomly chosen from these $2N$ sequences, there are $2N - 1$ possible ancestral sequences. Denote the ith oldest possible ancestral sequence by A_i. One example is shown in Figure 2, where $2N = 6$ is assumed. Also denote by a_i the probability that the common ancestral sequence to two sequences chosen at random is A_i. As shown in the APPENDIX, this probability is given by

$$a_i = \frac{2(2N+1)}{(i+1)(i+2)(2N-1)}, \qquad (6)$$

where $1 \leq i \leq 2N - 1$. For example, when $2N = 6$, we have $a_1 = 7/15$, $a_2 = 7/30$, $a_3 = 7/50$, $a_4 = 7/75$, and $a_5 = 1/15$. When A_i is the common ancestor, the expected number, $E(k_i)$, of nucleotide differences between the two sequences is

$$E(k_i) = \sum_{j=i}^{2N-1} (2v/\lambda_j) = 2M\left(\frac{1}{i} - \frac{1}{2N}\right). \qquad (7)$$

This equation can be obtained as follows: First, we notice from (2) that v/λ_j mutations are expected to take place on each sequence while $j + 1$ sequences are derived from j sequences. Considering two sequences, we obtain (7) since we can detect all the mutations in the infinite site model. Then, the expected number of nucleotide differences between two DNA sequences randomly chosen from the population is given by

$$E(k) = \sum_{i=1}^{2N-1} a_i E(k_i) = M.$$

Thus, we obtain (5).

Using the infinite allele model, KIMURA and CROW (1964) have shown that the expected homozygosity, $E(F)$, or the probability that the randomly chosen two DNA sequences is identical is given by

$$E(F) = \frac{1}{1+M}. \qquad (8)$$

This equation can be obtained in the same way as the above, namely, we have

$$E(F) = \sum_{i=1}^{2N-1} a_i E(F_i), \qquad (9)$$

where $E(F_i)$ is the expected homozygosity when A_i is the common ancestor to the randomly chosen two sequences. Since the probability distribution of the number of generations between A_i and A_{i+1} is given by (2), we have

$$E(F_i) = \prod_{j=i}^{2N-1} b_j, \qquad (10)$$

where b_j is given by

$$b_j = \int_0^\infty f_j(t)\exp(-2vt)dt = \frac{\lambda_j}{\lambda_j + 2v}.$$

In these equations b_j is the probability that there is no mutation on the two sequences while $j + 1$ sequences are derived from j sequences. Since the expected homozygosity is given by the product of b_j's, we obtain (10). Substituting (10) into (9), we obtain (8).

DNA polymorphism at the time of fixation: Let us now study DNA polymorphism at the time when a mutant at a particular site has fixed. In this case the genealogy of $2N$ DNA sequences shows unique topologies as mentioned earlier. Figure 3 gives one example where $2N = 6$ is assumed.

The expected number, $E(k \mid \text{fix})$, of nucleotide differences between two DNA sequences randomly chosen from the population at the time when a mutant at a particular nucleotide site has fixed can be obtained, using the same method as the above. Let A_i be the ith oldest possible common ancestral sequence to the two DNA sequences randomly chosen from $2N$ DNA sequences with the fixed mutant. Then, the probability,

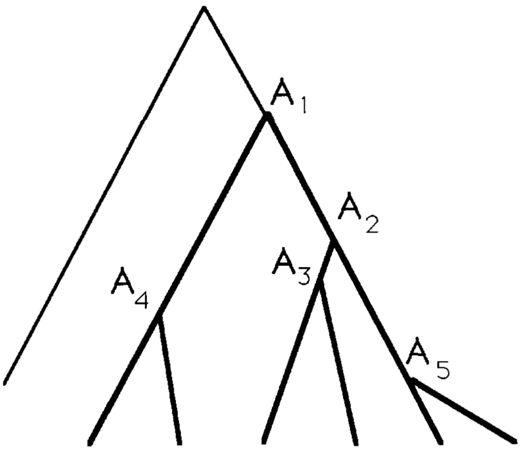

FIGURE 3.—One example of genealogical relationship among six DNA sequences, given that fixation took place. A_i is the ith oldest ancestor to these six sequences. Bold lines show the sequences with the mutant nucleotide fixed.

a_i, that the common ancestor of the two randomly chosen sequences is A_i is given by (6), as shown earlier. The expected number, $E(k_i \mid \text{fix})$, of nucleotide differences between the two sequences when A_i is their common ancestor is given by

$$E(k_i \mid \text{fix}) = \sum_{j=i+1}^{2N-1} (2v/\lambda_j) = 2M\left(\frac{1}{i+1} - \frac{1}{2N}\right). \quad (11)$$

This equation can be obtained by replacing i with $i + 1$ in (7) since there are $i + 1$ sequences, namely i mutant sequences and one nonmutant sequence, at the time when A_i occurred. Then, $E(k \mid \text{fix})$ is given by

$$E(k \mid \text{fix}) = \sum_{i=1}^{2N-2} a_i E(k_i \mid \text{fix}), \quad (12)$$

which approximately becomes

$$E(k \mid \text{fix}) = 2\left(\frac{\pi^2}{3} - 3\right)M = 0.5797\ldots M. \quad (13)$$

This formula indicates that, when fixation takes place, the average number of nucleotide differences is expected to be 42% less than at a random time.

Using the infinite allele model, the expected homozygosity, $E(F \mid \text{fix})$, at the time of fixation can be obtained in exactly the same way as the above. Namely, we have

$$E(F \mid \text{fix}) = \sum_{i=1}^{2N-2} a_i E(F_i \mid \text{fix}), \quad (14)$$

where $E(F_i \mid \text{fix})$ is given by

$$E(F_i \mid \text{fix}) = \prod_{j=i+1}^{2N-1} b_j.$$

This formula can be simplified as

$$E(F \mid \text{fix}) = \frac{2N+1}{(2N-1)M} \cdot \left[1 - \prod_{i=2}^{2N-1} \frac{1}{1 + 2M/\{i(i+1)\}}\right], \quad (15)$$

which is approximately given by

$$E(F \mid \text{fix}) = [1 - \exp(-M + c_1M^2 - c_2M^3)]/M, \quad (16)$$

where $c_1 = (4\pi^2 - 39)/6 = 0.07973\ldots$ and $c_2 = (79 - 8\pi^2)/3 = 0.01438\ldots$. From this formula we can see that, when fixation takes place, the amount of polymorphism in terms of heterozygosity is also expected to be less than at random times, but the amount of reduction depends on the value of M.

DNA polymorphism for a given fixation time: Let us now study the expected number, $E(k \mid T)$, of nucleotide differences between the two DNA sequences randomly chosen from the population at the time of fixation, given that the fixation time is T. As mentioned earlier, the probability distribution of fixation time can be obtained from the convolution of $f_i(t)$'s for all i's, where $f_i(t)$ is given by (2). This can be expressed as

$$y(T) = \int_{t=0}^{T} f_i(t)g_i(T - t)dt,$$

where $g_i(t)$ can be obtained from the convolution of $f_j(t)$'s for all j's except $j = i$. The conditional expectation, $E(t_i \mid T)$, of the number of generations between A_{i-1} and A_i, which has the probability distribution $f_i(t)$, can be given by

$$E(t_i \mid T) = z_i(T)/y(T), \quad (17)$$

where $z_i(T)$ is given by

$$z_i(T) = \int_{t=0}^{T} tf_i(t)g_i(T - t)dt. \quad (18)$$

$y(T)$ can be obtained by using (3) or (4), and $z_i(T)$ can be given by

$$z_i(T) = \sum_{\substack{j=1\\ j\neq i}}^{2N-1} (2j+1)(-1)^{j+1} \prod_{k=1}^{j} \frac{2N-k}{2N+k}\frac{\lambda_j}{\lambda_i - \lambda_j} \cdot[\exp(-\lambda_jT) - \exp(-\lambda_iT)] + (2i+1)(-1)^{i+1} \cdot \prod_{k=1}^{i} \frac{2N-k}{2N+k}\lambda_iT\exp(-\lambda_iT), \quad (19)$$

which is approximately given by

$$z_i(T) = \sum_{\substack{j=1\\ j\neq i}}^{2N-1} (2j+1)(-1)^{j+1} \frac{\lambda_j}{\lambda_i - \lambda_j}[\exp(-\lambda_jT) - \exp(-\lambda_iT)]\exp(-2\lambda_j) + (2i+1)(-1)^{i+1}\lambda_iT\exp[-\lambda_i(T+2)]. \quad (20)$$

In the same way as the above, the expected number, $E(k_i \mid T)$, of nucleotide differences between the two

DNA sequences when their common ancestor is A_i, given that the fixation time is T, is given by

$$E(k_i \mid T) = \sum_{j=i+1}^{2N-1} 2vE(t_j \mid T). \tag{21}$$

Since the probability that A_i is the common ancestor to the randomly chosen two sequences is a_i, the expected number, $E(k \mid T)$, of nucleotide differences between the two DNA sequences randomly chosen from the population at the time of fixation, given that the fixation time is T, is given by

$$E(k \mid T) = \sum_{i=1}^{2N-2} a_i E(k_i \mid T), \tag{22}$$

which can be simplified as

$$E(k \mid T) = \sum_{i=1}^{2N-2} 2vd_i E(t_{i+1} \mid T), \tag{23}$$

where d_i is given by

$$d_i = \frac{i(2N+1)}{(i+2)(2N-1)}.$$

Although numerical examples will be shown later, here we notice that, as T increases, $E(t_i \mid T)$ approaches

$$E(t_i \mid \infty) = \frac{4N}{(i-1)(i+2)}, \tag{24}$$

so that from (23) we have

$$E(k \mid \infty) = \frac{2(2N-2)}{3(2N-1)}M \approx \frac{2}{3}M = 0.6666\ldots M. \tag{25}$$

This formula indicates that, even when it takes a large number of generations for a mutant to reach fixation, the amount of DNA polymorphism at the time of fixation, in terms of the number of nucleotide differences between the two DNA sequences randomly chosen from the population or the average number of (pairwise) nucleotide differences among a sample of sequences, is expected to be 33% less than at a random time.

NUMERICAL EXAMPLES AND COMPUTER SIMULATION

In order to check the accuracy of the formulae obtained, computer simulations were conducted. To save computer time, the telescoping method proposed by KIMURA and TAKAHATA (1983) was used, which is an improved version of the pseudosampling-variable method (KIMURA 1980).

Neutral mutation: First, we consider the case where a newly arisen mutant is selectively neutral. Let x_i be the relative frequency of mutant at generation i. Assume that there is no mutant at generation 0 and a new mutation takes place at generation 1, so that $x_0 = 0$ and $x_1 = 1/(2N)$. Then, we start computer simulation followed by the telescoping method and record the frequency of mutant at every generation until the mutant reaches fixation or extinction. In the case of extinction we discard all the records and repeat the simulation from the beginning. In this simulation the population size (N) was assumed to be 100 and we have obtained 10,000 events of fixation.

From a set of data we can easily obtain the fixation time. Since the probability that the two randomly chosen mutant sequences at generation $i + 1$ has their common ancestor at generation i is $1/(2Nx_i)$, the expected number of nucleotide differences between the two sequences can be obtained, using

$$E(k_{i+1}) = \left(1 - \frac{1}{2Nx_i}\right)[E(k_i) + 2v] + \frac{1}{2Nx_i}2v$$
$$= \left(1 - \frac{1}{2Nx_i}\right)E(k_i) + \frac{M}{2N} \tag{26}$$

repeatedly until fixation, where $E(k_1) = 0$. The result of computer simulation is shown in Table 1. The average fixation time obtained was 399 generations, which is almost the same as $4N$. The average of the expected number of nucleotide differences between the two DNA sequences obtained was $0.573M$. This indicates that (13) is a good approximation. Figure 4, as well as Table 1, shows the relationship between the fixation time and the expected number of nucleotide differences between the two DNA sequences randomly chosen from the population at the time of fixation obtained in this simulation. In this figure the

TABLE 1

Results of computer simulation

Fixation time	No. of cases	Average fixation time	Average no. of nucleotide differences between two DNAs[a]
-49	0		
50-99	29	89.5	0.301
100-149	351	131.3	0.374
150-199	947	177.6	0.450
200-249	1,226	225.4	0.501
250-299	1,316	275.0	0.547
300-349	1,223	324.1	0.576
350-399	1,013	375.0	0.596
400-449	828	423.9	0.622
450-499	662	474.5	0.621
500-549	512	522.9	0.654
550-599	424	572.6	0.651
600-649	327	623.4	0.662
650-699	227	674.2	0.688
700-749	180	723.1	0.680
750-799	149	773.8	0.679
800-849	129	823.1	0.650
850-899	114	874.6	0.690
900-949	73	924.2	0.688
950-999	67	976.8	0.666
1000-	203	1,196.2	0.639
Total	10,000	399.1	0.573

[a] This number is measured with $4Nv$.

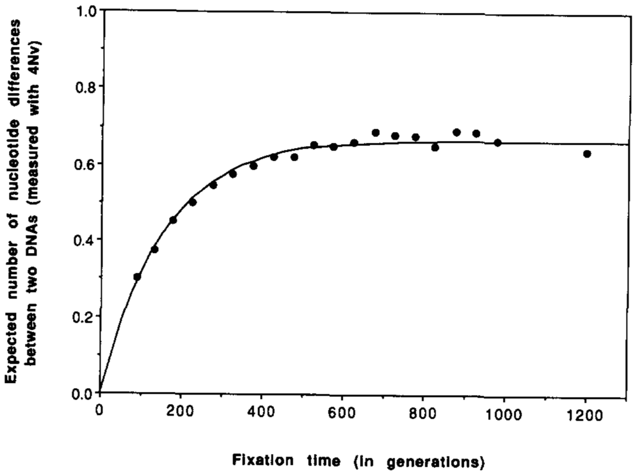

FIGURE 4.—Relationship between the expected number of nucleotide differences between the two DNA sequences randomly sampled from the population and the fixation time, where $N = 100$ was assumed. The line was obtained from (23) with (17), (4) and (20). Closed circles are the results of computer simulation, whose data are shown in Table 1.

line that was obtained by using (23) with (17), (4) and (20) shows a good agreement with the result of simulation.

Advantageous mutation: So far we consider only the case of neutral mutation. Here we examine the effect of an advantageous mutant on the expected number of nucleotide differences between the two sequences, using computer simulation. If we denote by s the selective advantage of mutant over nonmutant, the mean rate of change in x_i per generation is approximately given by

$$\Delta x_i = sx_i(1 - x_i).$$

Then, computer simulation was conducted under this selection model, where $s = 0.005$, 0.01, 0.02, 0.05, 0.1 and 0.2 are used together with $s = 0$. The method of simulation is the same as the above, except the change in frequency of the mutant is affected by selection. For each value of s, 100 events of fixation were collected, where $N = 100$ was also assumed.

The results of simulation are shown in Figure 5, where each point is the sum of ten replicates classified according to the length of fixation time so that there are ten points for each value of s. Interestingly, this figure shows that the expected number of nucleotide differences between the two sequences at the time when an advantageous mutant becomes fixed is essentially the same as that of neutral mutant if the fixation time is the same. It should be noted here that the fixation of an advantageous mutant tends to reduce the amount of DNA polymorphism more strongly than that of a neutral mutant since the fixation of an advantageous mutant tends to take place more rapidly than that of a neutral mutation. This conclusion is consistent with that of KAPLAN, HUDSON and LANGLEY (1989).

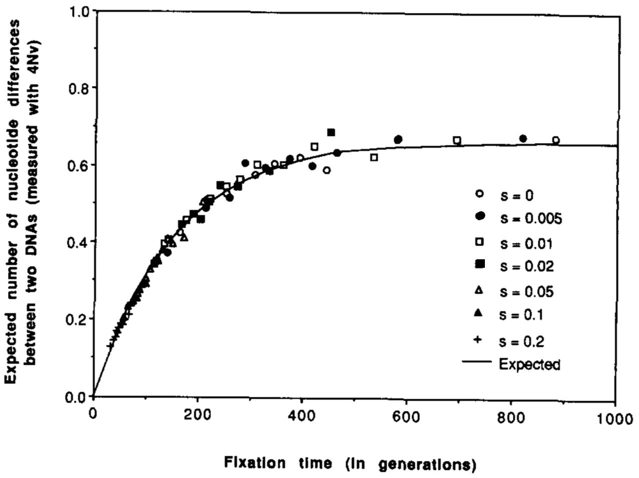

FIGURE 5.—Results of computer simulation conducted under the genic selection model of advantageous mutation. The line was obtained from (23) with (17), (4) and (20) under the neutral mutation model.

Effect of recombination: We have shown that the amount of DNA polymorphism is less when a mutant at a particular nucleotide site becomes fixed, compared to random times. This conclusion was obtained under the assumption that there is no recombination in DNA sequence. When there is some recombination, the degree of reduction might not be so large as that of no recombination. To know it quantitatively, computer simulation was conducted.

Let r be the recombination rate between the site where a mutant fixes and sites where divergence is being considered, and x_i be the relative frequency of the mutant at generation i. Then, RICHARD R. HUDSON (personal communication) has shown that the expected number of nucleotide differences between two sequences can be obtained by using

$$E(k_{i+1}) = \left[1 - \frac{1}{2Nx_i} - \frac{R(1-x_i)}{2N}\right]E(k_i) + \frac{R(1-x_i)}{2N}E(k_i') + \frac{M}{2N}, \tag{27a}$$

$$E(k_{i+1}') = \left(1 - \frac{R}{4N}\right)E(k_i') + \frac{Rx_i}{4N}E(k_i) + \frac{R(1-x_i)}{4N}E(k_i'') + \frac{M}{2N}, \tag{27b}$$

$$E(k_{i+1}'') = \left[1 - \frac{1}{2N(1-x_i)} - \frac{Rx_i}{2N}\right]E(k_i'') + \frac{Rx_i}{2N}E(k_i') + \frac{M}{2N}, \tag{27c}$$

where $M = 4Nv$, $R = 4Nr$, $E(k_i)$ is the expected number of nucleotide differences between two sequences randomly chosen from the sequences which bear the mutation going to fixation, $E(k_i')$ is the expected num-

TABLE 2

Expected number of nucleotide differences between two DNA sequences when a mutant at a linked site is fixed, obtained by computer simulation

	Selection coefficient (s)						
$4Nr$	0	0.005	0.01	0.02	0.05	0.1	0.2
0	0.576	0.573	0.552	0.506	0.380	0.266	0.170
0.01	0.579	0.577	0.555	0.508	0.381	0.268	0.170
0.1	0.607	0.604	0.578	0.527	0.394	0.277	0.177
0.2	0.634	0.631	0.602	0.547	0.408	0.287	0.184
0.5	0.700	0.696	0.662	0.599	0.448	0.316	0.204
1	0.773	0.769	0.734	0.667	0.507	0.362	0.237
2	0.855	0.851	0.821	0.760	0.601	0.442	0.298
5	0.941	0.938	0.923	0.886	0.769	0.616	0.450
10	0.977	0.975	0.969	0.951	0.887	0.778	0.625
20	0.993	0.992	0.990	0.984	0.959	0.907	0.810
Average fixation time	401	389	326	244	138	82	47

$N = 100$ and $4Nv = 1$ were assumed. For each selection coefficient, s, 1000 events of fixation were collected.

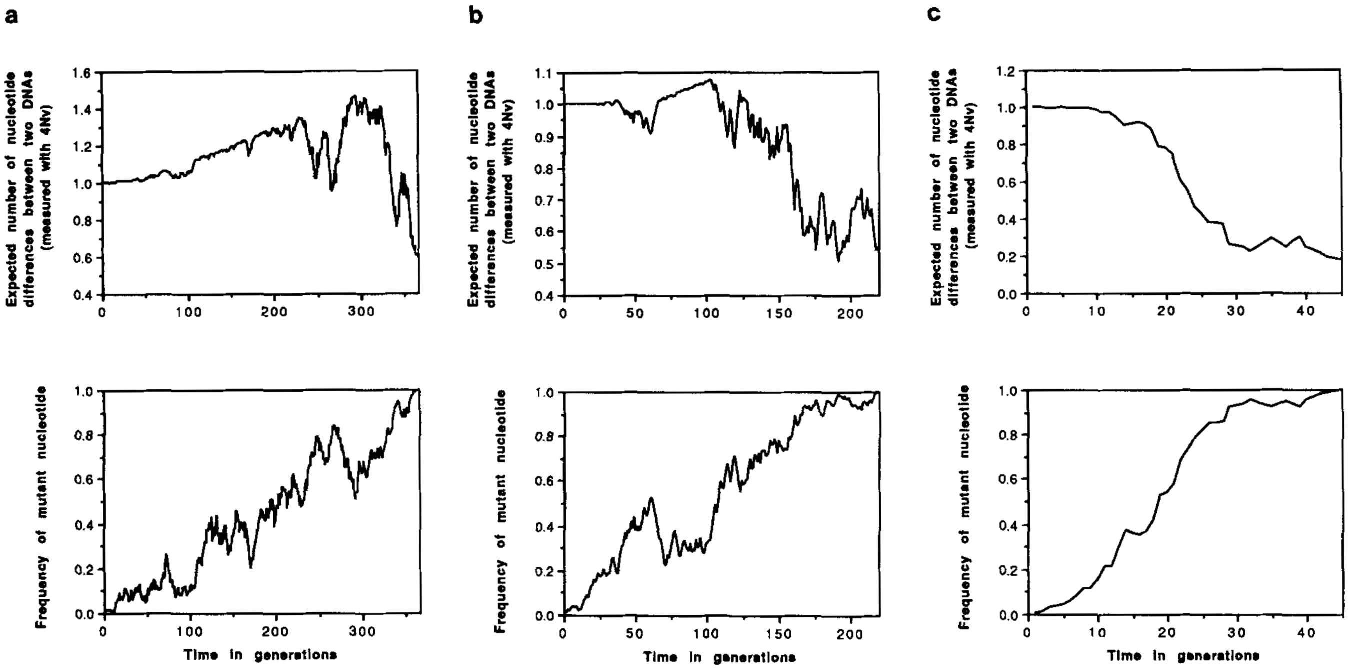

FIGURE 6.—Three examples of the changes in the frequency of DNA sequence with mutant nucleotide and in the expected number of nucleotide differences between the two DNA sequences randomly chosen from the population which were obtained from the computer simulation where $N = 100$ was assumed. (a) $s = 0$ and $T = 367$; (b) $s = 0.02$ and $T = 220$; (c) $s = 0.2$ and $T = 45$.

ber of nucleotide differences between two sequences one bearing the mutant and one not, and $E(k_i'')$ is the expected number of nucleotide differences for two sequences not bearing the mutation. These three equations can be derived from (26) with the same reasoning as equations (15) of KAPLAN, HUDSON and LANGLEY (1989), assuming $r \ll 1$. Starting from $E(k_1) = 0$ and $E(k_1') = E(k_1'') = 4Nv$, we obtain $E(k_i)$.

The method of simulation is the same as the above, except (27) is used instead of (26). In this simulation $N = 100$ was also assumed, and $4Nv = 1$, $s = 0$, 0.005, 0.01, 0.02, 0.05, 0.1 and 0.2, and $4Nr = 0$, 0.01, 0.1, 0.2, 0.5, 1, 2, 5, 10 and 20 were used. For each value of s, 1000 events of fixation were collected. The results are shown in Table 2, which indicates that, as the recombination rate increases, the amount of DNA polymorphism also increases. In the case of a neutral mutant ($s = 0$), if $4Nr$ is larger than 10, the amount of polymorphism at a time of fixation is almost the same as that of a random time. This table also indicates that, when a strongly advantageous ($s > 0.1$) mutant is fixed, the amount of polymorphism is substantially smaller than at a random time even if $4Nv$ is larger than 10. This result is consistent with that of KAPLAN, HUDSON and LANGLEY (1989).

DISCUSSION

We consider the DNA polymorphism only at the time of fixation. In the case of neutral mutation the expected number of nucleotide differences between the two DNA sequences randomly chosen from the population is M (= $4Nv$), while the number becomes $0.58M$ at the time of fixation. This difference might be caused by the large amount of DNA polymorphism on the way to fixation. Figure 6 shows some examples of the relationship between the frequency of mutant and the expected number of nucleotide differences between the two sequences, which were obtained from the computer simulation in the previous section. Figure 6a shows the example in the case of $s = 0$, where the fixation time was 367 generations. In this example the expected number of nucleotide differences is larger than $4Nv$ when the frequency of mutant is intermediate. This does not occur in the case of rapid fixation as shown in Figure 6c. At any rate, the amount of DNA polymorphism changes drastically, depending on the frequency of the mutant. This might be one of the main reasons for a large stochastic variance of the amount of DNA polymorphism.

The results of the computer simulation conducted have shown that the formulas (22), (23) and (25) obtained under the assumption of neutral mutation also hold in the case of advantageous mutation. This conclusion, however, might be correct only in the genic selection model or the semi-dominant mutation model. In fact, in the case where the advantageous mutation is recessive or dominant the results of the computer simulation conducted in the same way as the above show that this is not the case (data not shown). At any rate more extensive studies are needed for various types of selection model, including overdominance selection model.

I thank R. R. Hudson for showing me equations (27) which improved this paper greatly. I also thank R. R. Hudson and two anonymous reviewers for their valuable suggestions and comments. Contribution No. 1823 from the National Institute of Genetics, Mishima, 411 Japan.

LITERATURE CITED

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

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

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., 1970 The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. Genet. Res. **15:** 131–133.

Kimura, M., 1980 Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical, and pseudo-sampling methods. Proc. Natl. Acad. Sci. USA **77:** 522–526.

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

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 N. Takahata, 1983 Selective constraint in protein polymorphism: Study of the effective neutral mutation model by using an improved pseudosampling method. Proc. Natl. Acad. Sci. USA **80:** 1048–1052.

Kingman, J. F. C., 1982a On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

Kingman, J. F. C., 1982b Exchangeability and the evolution of large populations, pp. 97–112 in *Exchangeability in Probability and Statistics*, edited by G. Koch and F. Spizzichino. North-Holland, Amsterdam.

Kojima, K. I., and H. E. Schaeffer, 1967 Survival process of linked genes. Evolution **21:** 518–531.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23–35.

Moran, P. A. P., 1958 Random processes in genetics. Proc. Camb. Philos. Soc. **54:** 60–71.

Ohta, T., and M. Kimura, 1975 The effect of a selected linked locus on heterozygosity of neutral alleles (the hitch-hiking effect). Genet. Res. **25:** 313–325.

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

Watterson, G. A., 1975 On the number of segregating sites in genetic models without recombination. Theor. Popul. Biol. **10:** 256–276.

Watterson, G. A., 1982a Substitution times for mutant nucleotides. J. Appl. Probab. **19A:** 59–70.

Watterson, G. A., 1982b Mutant substitutions at linked nucleotide sites. Adv. Appl. Probab. **14:** 206–224.

Watterson, G. A., 1984 Lines of descent and the coalescent. Theor. Popul. Biol. **26:** 77–92.

Communicating editor: R. R. Hudson

APPENDIX

Denote by A_i the ith oldest ancestral DNA sequence to a sample of n sequences as shown in Figure 2. Also denote by $a_{i,n}$ the probability that the common ancestor to the two sequences randomly chosen from a sample of n sequences is A_i, where $1 \leqslant i \leqslant n - 1$. First, we notice that there are $\binom{n}{2}$ pairwise combinations and that only one of them creates the latest ancestor (A_5 in Figure 2), so that we have

$$a_{n-1,n} = 1 \bigg/ \binom{n}{2} = \frac{2}{n(n-1)}. \tag{A1}$$

After this combination we now have $n - 1$ sequences. If we know $a_{i,n-1}$, then $a_{i,n}$ can be given by

$$a_{i,n} = (1 - a_{n-1,n})a_{i,n-1} = \frac{(n+1)(n-2)}{n(n-1)}\, a_{i,n-1}, \tag{A2}$$

where $1 \leqslant i \leqslant n - 2$. From these equations we have

$$a_{i,n} = \frac{2(n+1)}{(i+1)(i+2)(n-1)}. \tag{A3}$$

If we denote $a_{i,2N}$ by a_i, we finally have (6).