The Effect of Change in Population Size on DNA Polymorphism

Fumio Tajima

Department of Biology, Kyushu University, Fukuoka 812, Japan

Manuscript received March 10, 1989
Accepted for publication July 14, 1989

ABSTRACT

The expected number of segregating sites and the expectation of the average number of nucleotide differences among DNA sequences randomly sampled from a population, which is not in equilibrium, have been developed. The results obtained indicate that, in the case where the population size has changed drastically, the number of segregating sites is influenced by the size of the current population more strongly than is the average number of nucleotide differences, while the average number of nucleotide differences is affected by the size of the original population more severely than is the number of segregating sites. The results also indicate that the average number of nucleotide differences is affected by a population bottleneck more strongly than is the number of segregating sites.

THE amount of genetic variation at the DNA level can be measured by the number of segregating sites among DNA sequences sampled (WATTERSON 1975) or by the average number of (pairwise) nucleotide differences between DNA sequences sampled (TAJIMA 1983). The statistical properties of these quantities have been obtained under the assumption that the size of population is constant (WATTERSON 1975; TAJIMA 1983).

The size of population, however, often changes drastically. Although the effects of change in population size on heterozygosity and the number of alleles in a sample have already been studied by NEI, MARUYAMA and CHAKRABORTY (1975), CHAKRABORTY and NEI (1977), MARUYAMA and FUERST (1984, 1985a,b), WATTERSON (1986), the effect of change in population size on the number of segregating sites and the average number of nucleotide differences is not yet known.

Here I examine this problem quantitatively, since the number of segregating sites and the average number of nucleotide differences are more appropriate measures for the amount of DNA polymorphism than heterozygosity and the number of alleles.

THEORY

Assumption: Assume that a mutant is selectively neutral (KIMURA 1968, 1983), and 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 (KIMURA 1969). Also assume that a population consists of diploid individuals, and consider a DNA sequence located on an autosomal chromosome.

General formula: Consider a randomly mating population with discrete and nonoverlapping generations, and let N_t be the effective population size in the tth generation. Denote by v the mutation rate per DNA sequence per generation. Also denote the expected number of segregating sites among n DNA sequences randomly chosen from a population in the tth generation by $S_n(t)$.

The number of segregating sites is the number of sites which are segregating (or polymorphic) among n DNA sequences. On the other hand, the average number of nucleotide differences between DNA sequences is given by

$$\hat{k} = \sum\sum_{i<j} k_{ij} \Big/ \binom{n}{2},$$

where k_{ij} is the number of nucleotide differences between the ith and jth DNA sequences. Therefore, the expectation of the average number of nucleotide differences is equal to the expected number of nucleotide differences between two DNA sequences randomly sampled from a population. Since the number of nucleotide differences between two DNA sequences is equal to the number of segregating sites when n is 2, the expectation of the average number of nucleotide differences is equal to the expected number of segregating sites for $n = 2$, namely

$$E(\hat{k}) = S_2(t).$$

Incidentally, $S_1(t) = 0$ since there is no segregating site when only one DNA sequence is considered.

If we denote the probability, that n DNA sequences randomly sampled from a population in the tth generation are derived from i DNA sequences in the previous generation, by $P_n(i)$ then $S_n(t)$ is given by

$$S_n(t) = \sum_{i=1}^{n} S_i(t-1)P_n(i) + nv, \qquad (1)$$

where the last term in the right side of (1) is the effect of mutations. When n is small, $P_n(i)$ is approximately given by

$$P_n(n) = 1 - \frac{\binom{n}{2}}{2N_{t-1}}, \qquad P_n(n-1) = \frac{\binom{n}{2}}{2N_{t-1}}. \tag{2}$$

and

$$P_n(i) = 0 \quad \text{for} \quad i < n-1$$

(KINGMAN 1982; HUDSON 1983; TAJIMA 1983). Substituting (2) into (1), we have

$$S_n(t) - S_n(t-1) = \frac{\binom{n}{2}}{2N_{t-1}} [S_{n-1}(t-1) - S_n(t-1)] + nv, \tag{3}$$

where $S_1(t) = 0$ as mentioned earlier.

If we use the differential equation method, (3) becomes

$$\frac{dS_n(t)}{dt} = \frac{\binom{n}{2}}{2N_t} [S_{n-1}(t) - S_n(t)] + nv. \tag{4}$$

This formula is simpler than (3), and we do not have to assume that n is small in this case. We use (4) instead of (3) in order to obtain $S_n(t)$.

Assume that the population size is constant ($N_t = N$, for $t > 0$). Then, integration of (4) gives

$$S_n(t) = a_n \exp(-a_n t) \int S_{n-1}(t)\exp(a_n t)\, dt + \frac{M}{n-1} + C_n \exp(-a_n t), \tag{5}$$

where

$$M = 4Nv, \qquad a_n = \frac{\binom{n}{2}}{2N},$$

and C_n is the integral constant which can be determined from the initial conditions. Then, we have

$$S_n(t) = b_{n,1} + \sum_{i=2}^{n} b_{n,i}\exp(-a_i t), \tag{6}$$

where

$$b_{n,1} = b_{n-1,1} + \frac{M}{n-1},$$

$$b_{n,i} = \frac{n(n-1)}{(n-i)(n+i-1)} b_{n-1,i}, \quad \text{for} \quad 1 < i < n \tag{7}$$

$$b_{n,n} = S_n(0) - \sum_{i=1}^{n-1} b_{n,i}.$$

$b_{1,1}$ is equal to 0 since $S_1(t)$ is 0, so that we have

$$b_{n,1} = M \sum_{i=1}^{n-1} \frac{1}{i}.$$

$b_{n,i}$ can be obtained by using (7) repeatedly.

For example, when n is 2, from (7) we have

$$b_{2,1} = M \quad \text{and} \quad b_{2,2} = S_2(0) - M.$$

Therefore, we obtain

$$S_2(t) = M + [S_2(0) - M]\exp[-t/(2N)], \tag{8}$$

which is identical with the formula obtained by LI (1977) using a different method. Incidentally, LI (1977) has shown not only the expectation but also the variance and distribution of the number of nucleotide differences between two DNA sequences.

Starting from an equilibrium population: When the population is in equilibrium at time 0, we can simplify (6). Since $S_n(0) = M_0 \sum_{i=1}^{n-1} (1/i)$, where $M_0 = 4N_0 v$ (WATTERSON 1975), (6) becomes

$$S_n(t) = M \sum_{i=1}^{n-1} \frac{1}{i} + (M_0 - M) \sum_{i=1}^{[n/2]} c_{n,i}\exp(-a_{2i}t), \tag{9}$$

where $[n/2]$ is the largest integer which is not greater than $n/2$, and $c_{n,i}$ is given by

$$c_{n,i} = \frac{(n-1)!n!(4i-1)}{(n-2i)!(n+2i-1)!i(2i-1)}. \tag{10}$$

When $n = 2$, we have $c_{2,1} = 1$ from (10). Therefore, we obtain (8).

NUMERICAL EXAMPLE

Starting from an equilibrium population: First, we consider the case where the population is in equilibrium at time 0. Then, $S_n(t)$ is given by (9). Table 1 shows the case where $M_0 = 0$ and $M = 1$. This means that until time 0 the size of the population is so small that there is no genetic variation, but population size becomes large afterwards. In this table the values of $S_n(t)/\sum_{i=1}^{n-1} (1/i)$ are shown, since they are equal to M when the population is in equilibrium. From this table

TABLE 1

Values of $S_n(t)/\sum_{i=1}^{n-1}(1/i)$ obtained by equation 9, where $4N_0v = 0$ and $4Nv = 1$ are assumed

$\frac{t}{2N}$	Sample size (n)					
	2	5	10	20	50	100
0.0	0.000	0.000	0.000	0.000	0.000	0.000
0.1	0.095	0.109	0.146	0.198	0.283	0.349
0.2	0.181	0.202	0.253	0.317	0.407	0.469
0.3	0.259	0.282	0.337	0.403	0.488	0.545
0.4	0.330	0.353	0.407	0.471	0.550	0.601
0.5	0.393	0.416	0.468	0.527	0.599	0.646
0.6	0.451	0.472	0.521	0.575	0.641	0.683
0.7	0.503	0.523	0.567	0.617	0.677	0.715
0.8	0.551	0.568	0.609	0.655	0.709	0.743
0.9	0.593	0.610	0.647	0.688	0.737	0.768
1.0	0.632	0.647	0.681	0.718	0.763	0.791
1.2	0.699	0.711	0.739	0.769	0.806	0.829
1.4	0.753	0.763	0.786	0.811	0.841	0.860
1.6	0.798	0.806	0.825	0.846	0.870	0.885
1.8	0.835	0.841	0.857	0.874	0.894	0.906
2.0	0.865	0.870	0.883	0.896	0.913	0.923
2.5	0.918	0.921	0.929	0.937	0.947	0.953
3.0	0.950	0.952	0.957	0.962	0.968	0.972
3.5	0.970	0.971	0.974	0.977	0.981	0.983
4.0	0.982	0.982	0.984	0.986	0.988	0.990
4.5	0.989	0.989	0.990	0.992	0.993	0.994
5.0	0.993	0.994	0.994	0.995	0.996	0.996
6.0	0.998	0.998	0.998	0.998	0.998	0.999
7.0	0.999	0.999	0.999	0.999	0.999	0.999
8.0	1.000	1.000	1.000	1.000	1.000	1.000

$S_n(t)$ is the expected number of segregating sites among a sample of n DNA sequences. Especially, $S_2(t)$ is equal to the expectation of the average number of (pairwise) nucleotide differences between DNA sequences sampled.

we can see that the amount of variation increases very slowly, especially in the case of $n = 2$. For example, it takes $1.4N$ generations until this number becomes half of the maximum value. On the other hand, in the case of $n = 100$ it takes only $0.5N$ generations. In fact, from (9) we can see that the larger is the sample size, the more quickly the number of segregating sites increases.

Table 2 shows the case where the size of population suddenly becomes one hundredth at time 0. In this case the number of segregating sites declines more rapidly than the average number of nucleotide differences. Again, the larger is the sample size, the more quickly the number of segregating sites decreases.

Bottleneck effect: In this section we consider the case where the size of the population becomes small, but the population recovers the original size T generations later. Figure 1 shows this process. At time 0 the population is assumed to be in equilibrium, so that $S_n(t)$ for $0 < t < T$ can be computed, using (9). After then, $S_n(t)$ is computed, using (6) with (7), since the population is no more in equilibrium. It should be noted that M is replaced with M_0 in these formulae.

Figure 2 gives several examples in which the population size is assumed to become one hundredth of the

TABLE 2

Values of $S_n(t)/\sum_{i=1}^{n-1}(1/i)$ obtained by equation 9, where $4N_0v = 100$ and $4Nv = 1$ are assumed

$\frac{t}{2N}$	Sample size (n)					
	2	5	10	20	50	100
0.0	100.00	100.00	100.00	100.00	100.00	100.00
0.1	90.58	89.17	85.55	80.36	72.01	65.45
0.2	82.05	80.00	74.99	68.60	59.72	53.53
0.3	74.34	72.06	66.63	60.11	51.65	46.04
0.4	67.36	65.07	59.67	53.40	45.57	40.51
0.5	61.05	58.84	53.70	47.83	40.65	36.10
0.6	55.33	53.27	48.47	43.06	36.52	32.40
0.7	50.16	48.25	43.84	38.88	32.94	29.22
0.8	45.48	43.74	39.69	35.18	29.79	26.43
0.9	41.25	39.66	35.98	31.88	26.99	23.95
1.0	37.42	35.97	32.63	28.91	24.49	21.73
1.2	30.82	29.63	26.88	23.83	20.20	17.95
1.4	25.41	24.44	22.18	19.68	16.71	14.87
1.6	20.99	20.19	18.34	16.29	13.86	12.35
1.8	17.36	16.71	15.20	13.52	11.53	10.29
2.0	14.40	13.86	12.62	11.25	9.62	8.61
2.5	9.13	8.80	8.05	7.22	6.23	5.62
3.0	5.93	5.73	5.28	4.77	4.17	3.80
3.5	3.99	3.87	3.59	3.29	2.92	2.70
4.0	2.81	2.74	2.57	2.39	2.17	2.03
4.5	2.10	2.06	1.95	1.84	1.71	1.62
5.0	1.67	1.64	1.58	1.51	1.43	1.38
6.0	1.25	1.24	1.21	1.19	1.16	1.14
7.0	1.09	1.09	1.08	1.07	1.06	1.05
8.0	1.03	1.03	1.03	1.03	1.02	1.02
9.0	1.01	1.01	1.01	1.01	1.01	1.01
10.0	1.00	1.00	1.00	1.00	1.00	1.00

$S_n(t)$ is the expected number of segregating sites among a sample of n DNA sequences. Especially, $S_2(t)$ is equal to the expectation of the average number of (pairwise) nucleotide differences between DNA sequences sampled.

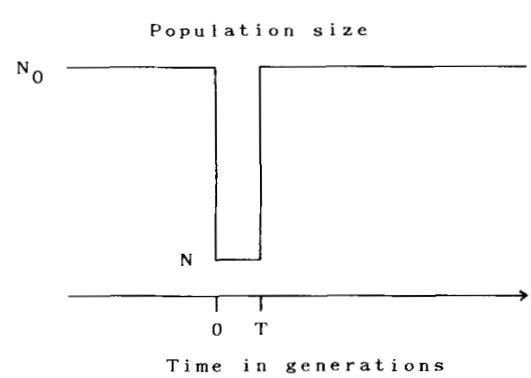

FIGURE 1.—The bottleneck model.

original size. For the values of T, $0.4N$, N, and $2N$ are used. In all the cases examined, larger reduction of $S_n(t)$ is observed when n is larger, but the bottleneck effect continues longer in the case where n is smaller. In other words, the average number of nucleotide differences is affected by the bottleneck of population size more strongly than is the number of segregating sites.

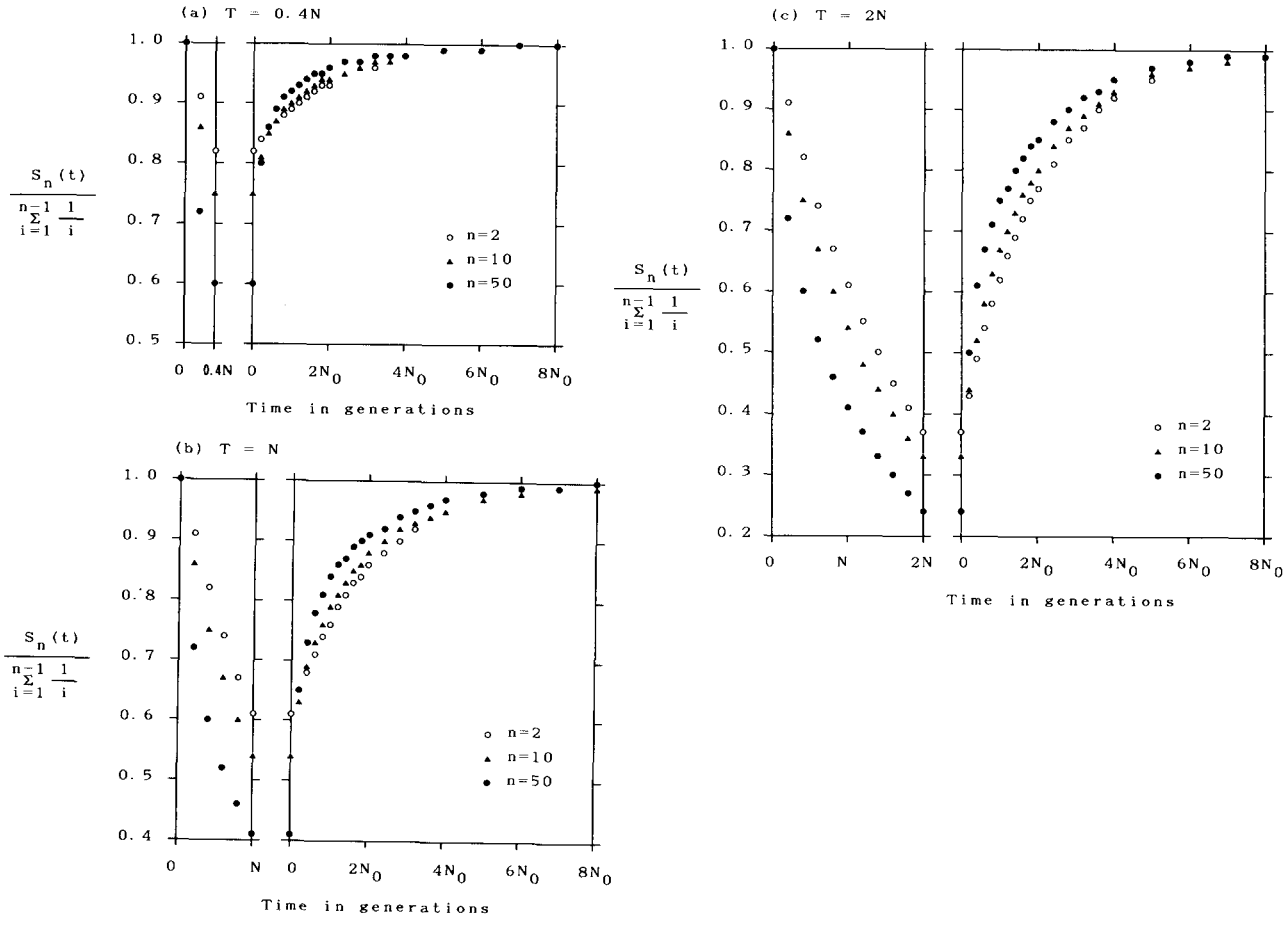

FIGURE 2.—Relationship between $S_n(t)/\sum_{i=1}^{n-1}(1/i)$ and the number of generations after the recovery of population size. $S_n(t)$ is the expected number of segregating sites among a sample of n DNA sequences. Especially, $S_2(t)$ is equal to the expectation of the average number of (pairwise) nucleotide differences between DNA sequences sampled. The bottleneck model is shown in Figure 1. The durations (T) of bottleneck are (a) $0.4N$, (b) N, and (c) $2N$ generations. $4N_0v = 1$ and $4Nv = 0.01$ are assumed. When points ● and ▲ (and ○) are close to each other, only point ● is plotted in order to avoid confusion. Point ○ is eliminated when it is close to point ▲.

DISCUSSION

In this paper the formulae for computing the expected number of segregating sites and the expectation of the average number of nucleotide differences among DNA sequences sampled from a population, which is not in equilibrium, have been developed. The results obtained indicate that the number of segregating sites is influenced by the size of current population more strongly than is the average number of nucleotide differences, while the average number of nucleotide differences is affected by the size of original population more severely than is the number of segregating sites. The relationship between the two numbers is quite similar to the relationship between heterozygosity and the number of alleles. In fact heterozygosity and the number of alleles obtained from the infinite allele model are equivalent to the average number of nucleotide differences and the number of segregating sites obtained from the infinite site model, respectively.

Recently, TAJIMA (1989) has developed a statistical method for testing the neutral mutation hypothesis by using the average number of nucleotide differences and the number of segregating sites. This method, however, assumes that a population is in equilibrium. As he has indicated, we must consider whether the population used is in equilibrium or not when we apply this method. In fact, if the population experienced a bottleneck recently, then this method may falsely reject the neutral hypothesis. This might be avoided, however, if we apply this method for several types of DNA polymorphism separately; for example, coding region *vs.* noncoding region, nucleotide polymorphism *vs.* insertion/deletion polymorphism, mitochondrial DNA *vs.* nuclear DNA, and so on.

I thank B. S. WEIR and two anonymous reviewers for their valuable suggestions and comments.

LITERATURE CITED

CHAKRABORTY, R., and M. NEI, 1977 Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. Evolution **31:** 347–356.

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

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

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

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

MARUYAMA, T., and P. A. FUERST, 1984 Population bottlenecks and nonequilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. Genetics **108:** 745–763.

MARUYAMA, T., and P. A. FUERST, 1985a Population bottlenecks and nonequilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. Genetics **111:** 675–689.

MARUYAMA, T., and P. A. FUERST, 1985b Population bottlenecks and nonequilibrium models in population genetics. III. Genic homozygosity in populations which experience periodic bottlenecks. Genetics **111:** 691–703.

NEI, M., T. MARUYAMA and R. CHAKRABORTY, 1975 The bottleneck effect and genetic variability in populations. Evolution **29:** 1–10.

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

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

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

WATTERSON, G. A., 1986 The homozygosity test after a change in population size. Genetics **112:** 899–907.

Communicating editor: B. S. WEIR