The Amounts of Nucleotide Variation Within and Between Allelic Classes and the Reconstruction of the Common Ancestral Sequence in a Population

Hideki Innan and Fumio Tajima

Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113, Japan

Manuscript received May 2, 1997
Accepted for publication August 8, 1997

ABSTRACT

The amounts of nucleotide variation within and between allelic classes were studied. The expectation and variance of the number of segregating sites and the expectation of the average number of pairwise differences among a sample of DNA sequences were obtained by using the theory of gene genealogy with no recombination. When the ancestral allelic class is unknown, it was found that the amount of variation within an allelic class increases with its frequency in the sample, while the amount of variation between two allelic classes is the largest when the two allelic classes exist equally. On the other hand, if we know the ancestral allelic class, as the frequency of the mutant allelic class increases, the amounts of variation within the mutant allelic class and between two allelic classes increase and the amount of variation within the ancestral allelic class decreases. As an example, we analyzed the polymorphism in the ND5 gene of *Drosophila melanogaster* and constructed the common ancestral sequence with high confidence, suggesting that the pattern of polymorphism within species gives useful information to know the ancestral sequence of the species.

THE amount and pattern of DNA polymorphism in a population have useful information about the mechanism of maintenance of genetic variation, the evolutionary history of the population and so on. There are a number of polymorphic (segregating) sites in DNA sequences sampled from a population. Knowing the ancestral sequence in the population might contribute to answering the above questions. In some cases the outgroup sequence is available. For example, when we analyze the human population, a sequence in chimpanzee is often used as the outgroup sequence. In many cases, however, the assumption that the ancestral sequence is identical to the outgroup sequence is not correct because of back and parallel mutations between the ancestral and outgroup sequences. In this article we show how to reconstruct the ancestral sequence from the pattern of DNA polymorphism.

When there are two nucleotides in a particular site, one of them is the ancestral nucleotide under the infinite site model (KIMURA 1969). WATTERSON and GUESS (1977) have shown by using the infinite allele model (KIMURA and CROW 1964) that the probability that the allele with frequency q is the oldest is q. This suggests that the most frequent nucleotide is likely to be the ancestor. However, this is not necessarily the case. If the frequency of one nucleotide increased recently, then this nucleotide may not be ancestral even when this nucleotide is the most frequent. If we use the pattern of DNA polymorphism at linked sites, however, we can obtain more information about the ancestor. Suppose that there are two nucleotides, say A and T, in a particular site. Then, we can divide DNA sequences into two classes: one class includes sequences with A and the other includes sequences with T in this site. We call such a class an allelic class. We expect that the class with the ancestral nucleotide is more variable than the class with the mutant nucleotide. Our algorithm for reconstructing the ancestral sequence uses both the frequencies of classes and the amounts of nucleotide variation within and between allelic classes.

Recently, SLATKIN (1996) has obtained the expected amount of nucleotide variation within the mutant allelic class under the condition that we know when the mutant appeared. In this paper, we obtain the unconditional expectations and variances of variation within and between allelic classes by modifying the theory developed by HUDSON and KAPLAN (1986).

THE AMOUNTS OF NUCLEOTIDE VARIATION

In this article we consider a random mating population with N diploid individuals. We assume that mutations are selectively neutral (KIMURA 1968, 1983). We use the infinite site model with no recombination (WATTERSON 1975) and the genealogical relationship of DNA sequences (GRIFFITHS 1980; KINGMAN 1982; HUDSON 1983; TAJIMA 1983). We consider the case where n sequences are randomly sampled from the population.

The numbers of segregating sites within and between two allelic classes when we do not know the ancestor:

Corresponding author: Fumio Tajima, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113, Japan.
E-mail: ftajima@biol.s.u-tokyo.ac.jp

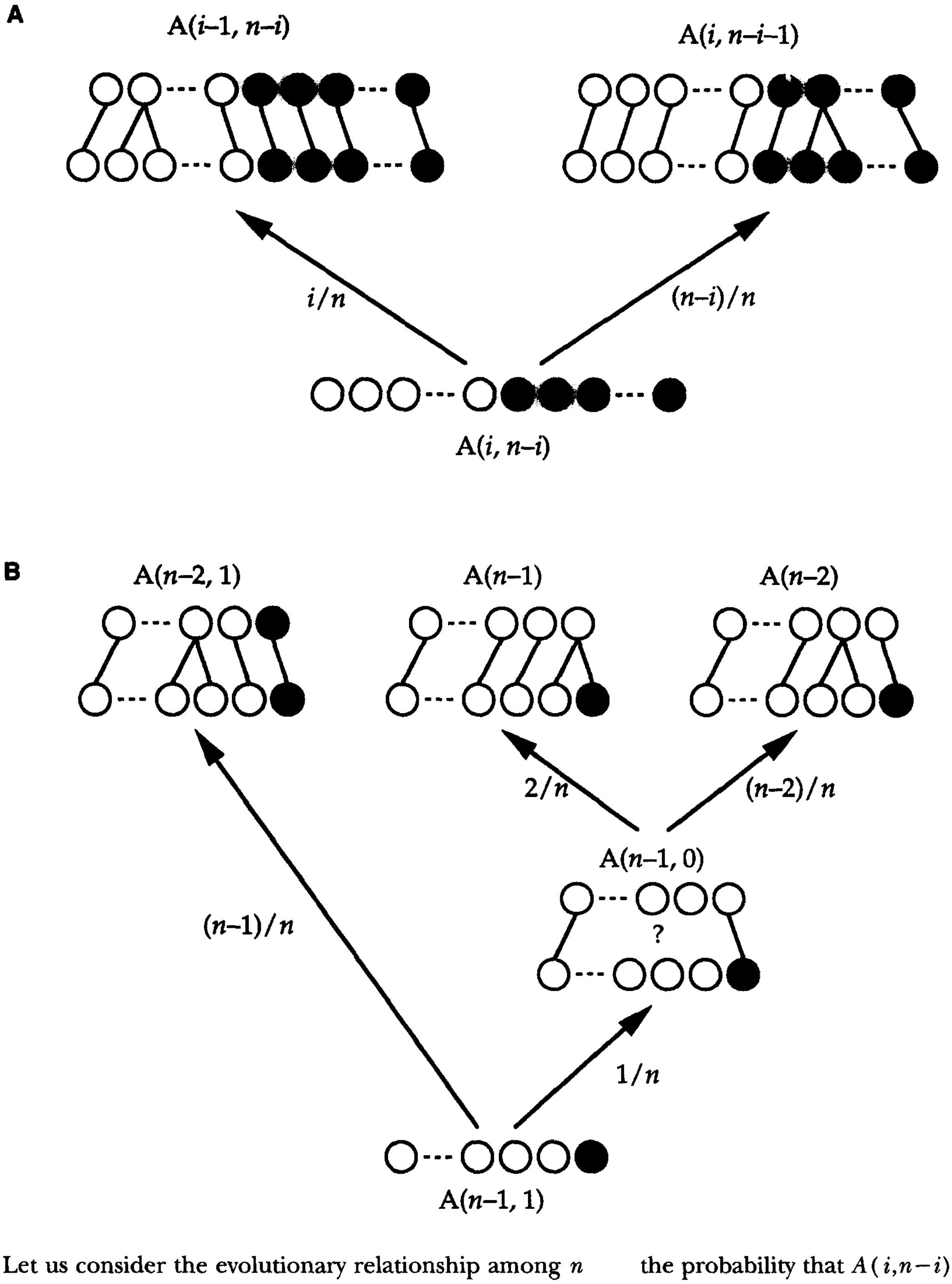

FIGURE 1.—Coalescent scheme in $A(i,n-i)$ when we do not know the ancestral allelic class. ○ represents the A1 allelic class and ● represents the A2 allelic class.

Let us consider the evolutionary relationship among n sequences. Assume that there are two allelic classes, A1 and A2, and that the A1 allelic class consists of i sequences and A2 consists of $n - i$ sequences. Denote this state by $A(i,n-i)$. Now we consider the number of segregating sites within allelic class. Let $S(i,n-i)$ be the expected number of the segregating sites within the A1 allelic class in $A(i,n-i)$. Note that the first number in the parentheses indicates the number of the sequences in the A1 allelic class and the second indicates that in the A2 allelic class. When the n sequences coalesce into $n - 1$ sequences, there are two possible states, $A(i-1,n-i)$ and $A(i,n-i-1)$, to which $A(i,n-i)$ can change (Figure 1A). Here, we consider the probability that $A(i,n-i)$ changes to $A(i-1,n-i)$. Denote this probability by p. Apparently, the probability that $A(i,n-i)$ changes to $A(i,n-i-1)$ is $1 - p$. Since A1 and A2 are selectively neutral, the expected frequency of A1 when the n sequences coalesce into $n - 1$ sequences must be the same as that in $A(i,n-i)$. The expected frequencies of A1 in $A(i,n-i)$, $A(i-1,n-i)$ and $A(i,n-i-1)$ are i/n, $(i-1)/(n-1)$ and $i/(n-1)$, respectively. Then, from

$$p\frac{i-1}{n-1} + (1-p)\frac{i}{n-1} = \frac{i}{n},$$

we have $p = i/n$. HUDSON and KAPLAN (1986) also used this probability. Therefore, for $2 \leq i \leq n-2$, $S(i,n-i)$ can be written as

$$S(i,n-i) = \frac{i}{n} S(i-1,n-i) + \frac{n-i}{n} S(i,n-i-1) + \frac{i\theta}{n(n-1)}, \quad (1)$$

where $\theta = 4N\mu$ (N is the effective population size and μ is the mutation rate per sequence per generation). The third term of the right side of (1) is the expected number of mutations in the A1 allelic class in the time during which the n sequences coalesce into $n-1$ sequences. This recursion was used by HUDSON and KAPLAN (1986) to study the amounts of variation in the nested subsamples from allozyme alleles, although in their report mutations that cause allozyme changes were also considered. In (1), only mutations that cause nucleotide changes are considered.

When one of allelic classes consists of one sequence, Equation (1) cannot be used. Here, let us consider the allelic state $A(n-1,1)$, where the A2 allelic class consists of one sequence and the remaining $n-1$ sequences belong to A1 allelic class (see Figure 1B). $A(n-1,1)$ changes to $A(n-2,1)$ with probability $(n-1)/n$ and to $A(n-1,0)$ with probability $1/n$. When $A(n-1,1)$ changes to $A(n-2,1)$, the expected number of mutations in the A1 allelic class during allelic state $A(n-1,1)$ is θ/n. On the other hand, in the process of the change from $A(n-1,1)$ to $A(n-1,0)$, there are $\binom{n}{2}$ possible pairs to coalesce, of which $\binom{n-1}{2}$ pairs are those among A1 sequences and $n-1$ pairs are those between A1 and A2 allelic classes. Since $A(n-1,0)$ means that all of the $n-1$ sequences belong to A1 allelic class, if the coalescence occurs among A1 allelic class, A2 must have been changed from A1 by mutation. We denote this state by $A(n-2)$ because $n-2$ out of $n-1$ sequences in $A(n-1,0)$ can contribute to the amount of variation within A1 allelic class. The coalescence between A1 and A2 allelic classes also involves a mutational change from A1 to A2. Since, in this case, all of the $n-1$ sequences in $A(n-1,0)$ can contribute to the amount of variation within A1 allelic class, we denote this state by $A(n-1)$. When $A(n-1,1)$ changes to $A(n-1,0)$, the expected number of mutations in the A1 allelic class is not given by θ/n, because one mutation that results in the allelic change from A1 to A2 must occur. TAJIMA (1983) studied the distribution of coalescent time for $n = 2$ under the condition that k mutations are involved, which was presented by Equation (19) of TAJIMA (1983). Modifying this equation, we obtain the distribution of t_n (time during which n sequences coalesce into $n-1$ sequences) when k mutations are involved. It becomes

$$f(t_n|k) = \left(\frac{B(n)}{2N}\right)^{k+1} t_n^k \exp[-B(n)t_n]/k!, \quad (2)$$

where $B(n) = \binom{n}{2} + (n/2)\theta$. The expectation and variance of t_n for a given value of k are

$$\mathrm{E}(t_n|k) = 2\mathrm{N}(k+1)/B(n), \quad (2a)$$

$$\mathrm{V}(t_n|k) = 4N^2(k+1)/[B(n)]^2. \quad (2b)$$

In the present study, we are interested in a particular segregating site that distinguishes two allelic classes. Since the mutation rate in this particular site is very small, we can assume $\theta = 0$. Then, we have $E(t_n|1) = 8N/n(n-1)$ and $V(t_n|1) = 32N^2/n^2(n-1)^2$, respectively. It should be noted that these values are twice the values for $k = 0$. Assuming that the number of mutations on a branch with length t_n follows the Poisson distribution with mean $t_n\mu$, the expectation and variance of s (the number of mutations on a branch with length t_n for $k = 1$) are

$$\mathrm{E}(s) = \frac{2\theta}{n(n-1)}, \quad (2c)$$

$$\mathrm{V}(s) = \frac{2\theta}{n(n-1)} + \frac{2\theta^2}{n^2(n-1)^2}. \quad (2d)$$

Then, the expected number of the segregating sites within the A1 allelic class during allelic state $A(n-1,1)$ when $A(n-1,1)$ changes to $A(n-1,0)$ is $2\theta/n$.

Considering three allelic states $A(n-2,1)$, $A(n-2)$ and $A(n-1)$ as shown in Figure 1B, we have

$$S(n-1,1) = \frac{n-1}{n}\left[S(n-2,1) + \frac{\theta}{n}\right] + \frac{1}{n}\left[\frac{n-2}{n} S(n-2) + \frac{2}{n} S(n-1) + \frac{2\theta}{n}\right], \quad (3)$$

where

$$S(n) = \sum_{k=1}^{n-1} \frac{1}{k}\theta. \quad (3a)$$

Since

$$S(n-1) = \frac{n-2}{n} S(n-2) + \frac{2}{n} S(n-1) + \frac{\theta}{n}, \quad (4)$$

(3) can be rewritten as

$$S(n-1,1) = \frac{n-1}{n} S(n-2,1) + \frac{1}{n} S(n-1) + \frac{\theta}{n}. \quad (5)$$

Noting $S(1,n-1) = 0$ and using Equations (1) and (5), we have a simple solution

$$S(i,n-i) = \frac{i}{n} S(i). \quad (6)$$

The expectation within the A2 allelic class in $A(i,n-i)$ is obtained by replacing the two numbers in the parentheses.

Next, we consider the number of segregating sites between two allelic classes. The segregating sites between two allelic classes are the sites in which different nucleotides are fixed in the two allelic classes. Let $S_b(i,n-i)$ be the expected number of segregating sites between two allelic classes in $A(i,n-i)$. It should be noted that $S_b(i,n-i) = S_b(n-i,i)$. Then, in the same way as above (see also Figure 1), $S_b(i,n-i)$ is given by

$$S_b(i,n-i) = \frac{i}{n}S_b(i-1,n-i) + \frac{n-i}{n}S_b(i,n-i-1) \quad (2 \le i \le n-2), \quad (7)$$

$$S_b(n-1,1) = \frac{n-1}{n}\left[S_b(n-2,1) + \frac{\theta}{n(n-1)}\right] + \frac{1}{n}\left[\frac{n-2}{n}S_b(n-1) + \frac{2\theta}{n(n-1)}\right] \quad (n \ge 3), \quad (8)$$

where

$$S_b(n) = \frac{\theta}{n-1}. \quad (8a)$$

$S_b(n)$ is the same as the expected number of mutations on one external branch in the genealogy with n sequences, which was obtained by Fu and Li (1993). For the derivation for $S_b(n)$, see appendix a. Substituting (8a) into (8), (8) becomes

$$S_b(n-1,1) = \frac{n-1}{n}S_b(n-2,1) + \frac{2\theta}{n(n-1)} \quad (n \ge 3). \quad (8b)$$

In the case of $n = 2$ ($i = n - i = 1$), since one mutation must occur before the coalescence, according to (2c), $S_b(1,1)$ is given by

$$S_b(1,1) = 2\theta. \quad (9)$$

Then, (8) becomes

$$S_b(n-1,1) = \frac{2}{n}\left(2 + \sum_{k=3}^{n}\frac{1}{k-1}\right)\theta \quad (n \ge 3). \quad (10)$$

From equations (7), (9) and (10), it is shown that

$$S_b(i,n-i) = 2\left[S(n) - \frac{i}{n}S(i) - \frac{n-i}{n}S(n-i)\right]. \quad (11)$$

Note that

$$S(i,n-i) + S(n-i,i) + S_b(i,n-i)/2 = S(n). \quad (12)$$

Modifying (1), (3), (7) and (8), we can also obtain the variances of the number of segregating sites within and between allelic classes, and the results are shown in appendix b.

The average numbers of pairwise differences within and between allelic classes are also obtained, following the scheme in Figure 1. Derivations for the average number of pairwise differences within an allelic class, $K(i,n-i)$, and that between allelic classes, $D(i,n-i)$, are shown in appendix c. It is indicated that $K(i,n-i)$ is a linear function of i for $i \ge 2$ and $D(i,n-i)$ is given by adding $(n-2)\theta/n$ to $S_b(i,n-i)$. Namely, we have

$$K(i,n-i) = \frac{i}{n}\theta, \quad (13)$$

$$D(i,n-i) = S_b(i,n-i) + \frac{n-2}{n}\theta. \quad (14)$$

Table 1 shows numerical examples of the number of segregating sites for $n = 20$ when we do not know the ancestor. The expected value is a linear function of θ, whereas the variance is a quadratic function of θ. It can be seen that the number of segregating sites has a considerable amount of variance when $\theta > 1$. Especially, the variance of the number of segregating sites between two allelic classes, $V_b(i,n-i)$, is large, because the coefficient of θ^2 is always larger than that of θ. The expected values for $1 \le i \le 19$ are plotted in Figure 2A. It is shown that $S(i,n-i)$ increases as i increases, suggesting that an allelic class maintained in higher frequency has a larger amount of variation. $S_b(i,n-i)$ has the highest peak at $i = n - i = 10$ and decreases symmetrically as i departs from 10. In other words, the expected number of segregating sites between two allelic classes is the largest when two allelic classes exist equally. The similar results are also obtained from the numbers of pairwise differences within and between allelic classes (Table 1, Figure 2B). $K(i,n-i)$ increases linearly as i increases, while $D(i,n-i)$ shows a symmetrical decrease from the highest peak at $i = 10$.

The numbers of segregating sites within and between allelic classes when we know the ancestor: In this section, let us assume that we know which is the ancestral allelic class. Assume that A1 is the ancestral allelic class and A2 is the new mutant allelic class. In this case, the probabilities that $A(i,n-i)$ changes to $A(i-1,n-i)$ and to $A(i,n-i-1)$ are different from those in the case where the ancestor is unknown. Let us consider the probability, p, that $A(i,n-i)$ changes to $A(i-1,n-i)$. When we do not know the ancestral allelic class, $A(i,n-i)$ changes to $A(i-1,n-i)$ with probability i/n, and to $A(i,n-i-1)$ with $(n-i)/n$, as shown in (1). Here, we know that the A1 allelic class is the ancestor. In $A(i-1,n-i)$, the probability that A1 is ancestral is $(i-1)/(n-1)$ and the corresponding probability in $A(i,n-i-1)$ is $i/(n-1)$ (Watterson and Guess 1977). Therefore, we have

$$p = \frac{\dfrac{i}{n}\dfrac{i-1}{n-1}}{\dfrac{i}{n}\dfrac{i-1}{n-1} + \dfrac{n-i}{n}\dfrac{i}{n-1}} = \frac{i-1}{n-1}. \quad (15)$$

Figure 3 shows the coalescent scheme assuming that A1 is the ancestor. In $A(i,n-i)$ and $A(n-1,1)$, the scheme

TABLE 1

The expected amount of nucleotide variation when the ancestral allelic class is unknown

		No. of segregating sites						Average no. of pairwise differences		
i	$n-i$	$S(i,n-i)$	$V(i,n-i)$	$S(n-i,i)$	$V(n-i,i)$	$S_b(i,n-i)$	$V_b(i,n-i)$	$K(i,n-i)$	$K(n-i,i)$	$D(i,n-i)$
19	1	3.320θ	$3.320\theta + 1.619\theta^2$	0.000θ	$0.000\theta + 0.000\theta^2$	0.455θ	$0.455\theta + 0.803\theta^2$	0.950θ	0.000θ	1.355θ
18	2	3.096θ	$3.096\theta + 1.621\theta^2$	0.100θ	$0.100\theta + 0.045\theta^2$	0.704θ	$0.704\theta + 1.244\theta^2$	0.900θ	0.100θ	1.604θ
17	3	2.874θ	$2.874\theta + 1.600\theta^2$	0.225θ	$0.225\theta + 0.119\theta^2$	0.898θ	$0.898\theta + 1.526\theta^2$	0.850θ	0.150θ	1.798θ
16	4	2.655θ	$2.655\theta + 1.559\theta^2$	0.367θ	$0.367\theta + 0.216\theta^2$	1.053θ	$1.053\theta + 1.709\theta^2$	0.800θ	0.200θ	1.953θ
15	5	2.439θ	$2.439\theta + 1.498\theta^2$	0.521θ	$0.521\theta + 0.329\theta^2$	1.176θ	$1.176\theta + 1.828\theta^2$	0.750θ	0.250θ	2.076θ
14	6	2.226θ	$2.226\theta + 1.421\theta^2$	0.685θ	$0.685\theta + 0.453\theta^2$	1.273θ	$1.273\theta + 1.903\theta^2$	0.700θ	0.300θ	2.173θ
13	7	2.017θ	$2.017\theta + 1.329\theta^2$	0.858θ	$0.858\theta + 0.584\theta^2$	1.346θ	$1.346\theta + 1.949\theta^2$	0.650θ	0.350θ	2.246θ
12	8	1.812θ	$1.812\theta + 1.223\theta^2$	1.037θ	$1.037\theta + 0.718\theta^2$	1.397θ	$1.397\theta + 1.976\theta^2$	0.600θ	0.400θ	2.297θ
11	9	1.611θ	$1.611\theta + 1.107\theta^2$	1.223θ	$1.223\theta + 0.852\theta^2$	1.428θ	$1.428\theta + 1.990\theta^2$	0.550θ	0.450θ	2.328θ
10	10	1.414θ	$1.414\theta + 0.982\theta^2$	1.414θ	$1.414\theta + 0.982\theta^2$	1.438θ	$1.438\theta + 1.994\theta^2$	0.500θ	0.500θ	2.338θ

$S(i,n-i)$, $S(n-i,i)$ and $S_b(i,n-i)$ were calculated using Equations (6) and (11); $V(i,n-i)$, $V(n-i,i)$ and $V_b(i,n-i)$ were calculated using Equations (B1) and (B4); and $K(i,n-i)$, $K(n-i,i)$ and $D(i,n-i)$ were calculated using Equations (13) and (14).

is nearly the same as that of Figure 1, while in $A(1,n-1)$ the next coalescence necessarily occurs in the mutant allelic class.

First, we consider the number of segregating sites in the ancestral allelic class. Let $S_a(i,n-i)$ be the expected number of segregating sites within the ancestral allelic class. Following the scheme in Figure 3, A and B, $S_a(i,n-i)$ is given by

$$S_a(i,n-i) = \frac{i-1}{n-1} S_a(i-1,n-i) + \frac{n-i}{n-1} S_a(i,n-i-1) + \frac{i\theta}{n(n-1)} \quad (2 \leq i \leq n-2), \qquad (16)$$

$$S_a(n-1,1) = \frac{n-2}{n-1} S_a(n-2,1) + \frac{1}{n-1} S(n-1) + \frac{\theta}{n}, \qquad (17)$$

where $S(n)$ is given by (3a).

Second, we consider the mutant allelic class. Let $S_m(i,n-i)$ be the expected number of segregating sites within the mutant allelic class. Note that the number of sequences in the mutant allelic class is the second number in the parentheses. Then, following the relationship in Figure 3, A and C, $S_m(i,n-i)$ is written as

$$S_m(i,n-i) = \frac{i-1}{n-1} S_m(i-1,n-i) + \frac{n-i}{n-1} S_m(i,n-i-1) + \frac{(n-i)\theta}{n(n-1)} \quad (2 \leq i \leq n-2), \qquad (18)$$

$$S_m(1,n-1) = S_m(1,n-2) + \frac{\theta}{n} = \sum_{k=3}^{n} \frac{1}{k}\theta \quad (n \geq 3). \qquad (19)$$

Third, the number of segregating sites between two allelic classes is considered. Denote the expected number of segregating sites between two allelic classes by $S_b^*(i,n-i)$. Then, we have the following recursions to calculate $S_b^*(i,n-i)$:

$$S_b^*(i,n-i) = \frac{i-1}{n-1} S_b^*(i-1,n-i) + \frac{n-i}{n-1} S_b^*(i,n-i-1) \quad (2 \leq i \leq n-2), \qquad (20)$$

$$S_b^*(1,n-1) = S_b^*(1,n-2) + \frac{\theta}{n(n-1)} = \left(\frac{5}{2} - \frac{1}{n}\right)\theta \quad (n \geq 2), \qquad (21)$$

$$S_b^*(n-1,1) = \frac{n-2}{n-1} S_b^*(n-2,1) + \frac{1}{n-1} S_b(n) + \frac{\theta}{n(n-1)} = \frac{1}{n-1}\left(2 + \sum_{k=3}^{n} \frac{2k-1}{k(k-1)}\right)\theta \quad (n \geq 3), \qquad (22)$$

where $S_b(n)$ is given by (8a). Modifying the Equations (16)-(22), we can also obtain the variances of the numbers of segregating sites and the expected numbers of pairwise differences within and between allelic classes, and the results are shown in APPENDIXES D and E, respectively.

Table 2 shows numerical examples of the numbers of segregating sites and the expected numbers of pairwise differences within the ancestral allelic class, within the mutant allelic class and between them for $n = 20$. Even when we know the ancestral allelic class, the numbers of segregating sites have a considerable amount of variance. The expected numbers of segregating sites are plotted in Figure 4A. It is shown that $S_a(i,n-i)$ decreases and $S_m(i,n-i)$ increases as $n-i$ increases (i decreases). Like $S_m(i,n-i)$, $S_b^*(i,n-i)$ also increases as $n-i$ increases. The expected numbers of pairwise

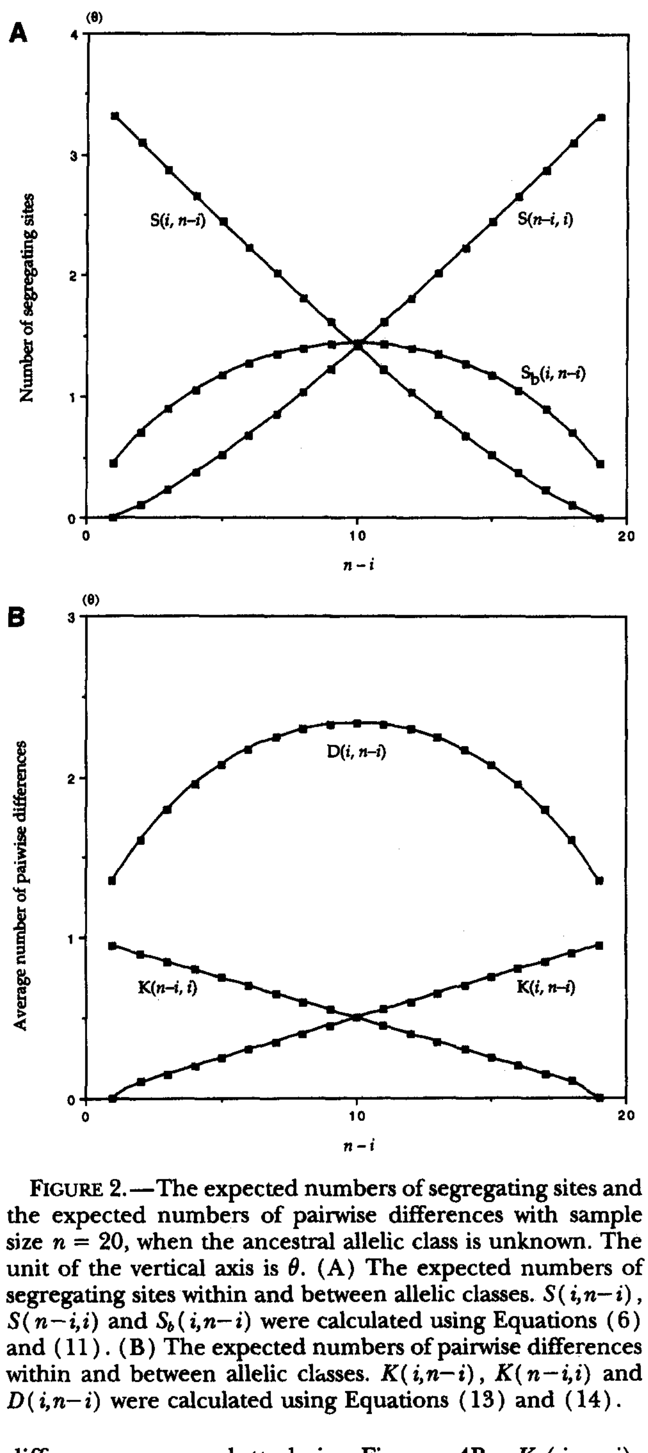

FIGURE 2.—The expected numbers of segregating sites and the expected numbers of pairwise differences with sample size n = 20, when the ancestral allelic class is unknown. The unit of the vertical axis is θ. (A) The expected numbers of segregating sites within and between allelic classes. $S(i,n-i)$, $S(n-i,i)$ and $S_b(i,n-i)$ were calculated using Equations (6) and (11). (B) The expected numbers of pairwise differences within and between allelic classes. $K(i,n-i)$, $K(n-i,i)$ and $D(i,n-i)$ were calculated using Equations (13) and (14).

differences are plotted in Figure 4B. $K_a(i,n-i)$, $K_m(i,n-i)$ and $D^*(i,n-i)$ indicate the average numbers of pairwise differences within the ancestral allelic class, within the mutant allelic class and between them, respectively. Like the number of segregating sites, as $n-i$ increases, $K_m(i,n-i)$ and $D^*(i,n-i)$ increase while $K_a(i,n-i)$ decreases.

RECONSTRUCTION OF THE COMMON ANCESTRAL SEQUENCE

To explain the algorithm to reconstruct the ancestral sequence, we use nucleotide polymorphism in a mitochondrial gene as an example because recombination is a very rare event in mitochondrial DNA. RAND and KANN (1996) reported the nucleotide polymorphism in the mitochondrial gene ND5 of *Drosophila melanogaster*. A total of 21 segregating sites were detected in the 1515-bp region of 59 sequences. In all the polymorphic sites, there were two nucleotides. Using Equations (16)–(22), we calculated the expected numbers of segregating sites within and between two segregating nucleotides (allelic classes) for each site (Table 3). In the left side of Table 3, the ancestral nucleotide is assumed to be the nucleotide in *D. simulans*, and in the right side the alternative nucleotide is assumed to be ancestral. For example, nucleotides G and A are segregating at nucleotide position 161, where only one sequence has A and the remaining 58 sequences have G. The observed number of segregating sites within the allelic class having G is 20, while there is no segregating site within the allelic class having A because it is a unique polymorphism. The number of segregating sites between the two allelic classes is also zero, excluding nucleotide position 161. Note that site 161 was excluded and the information on the remaining 20 segregating sites was used. Since the *D. simulans* sequence has G in the corresponding site, we assume that G is the ancestor in the left side of Table 3. The expected numbers of segregating sites within G, within A and between G and A are 19.288, 0.000 and 0.712, respectively. Alternatively, in the right side of Table 3, A is assumed to be ancestral. The expected numbers of segregating sites within A, within G and between them are 0.000, 11.205 and 8.795, respectively. Comparing the both sides in Table 3, the observed numbers of segregating sites are close to the expected values when G is assumed to be the ancestor, as shown in the left side of Table 3. In the other sites, according to this way, the expected numbers of segregating sites within the ancestral allelic class (nucleotide), within the mutant allelic class and between them are obtained, and the results are shown in Table 3.

Here, we consider the probability that each nucleotide is ancestral in each site. As an example, again, we consider nucleotide position 161 where G and A are segregating. We need to know the probabilities that G is the ancestor and A is the ancestor. Assume that the probability that the other two nucleotides, T and C, are the ancestor is extremely low. Denote the probability that G is the ancestor and the probability that A is the ancestor by Anc{G} and Anc{A}, respectively. Let X, Y and Z be the observed numbers of segregating sites within the ancestral allelic class, within the mutant allelic class and between them, respectively. In the case of nucleotide position 161, $X + Y + Z = 20$, and 59 sequences are partitioned into 58 and one sequences. We denote this probability by $P\{X+Y+Z=20;\ (58,1)\}$. Under this condition, Anc{G} and Anc{A} can be obtained. Let $P\{(58^*,1)\,|\,(58,1)\}$ and $P\{(1^*,58)\,|\,(58,1)\}$ be the probability that 58 sequences are the ancestor

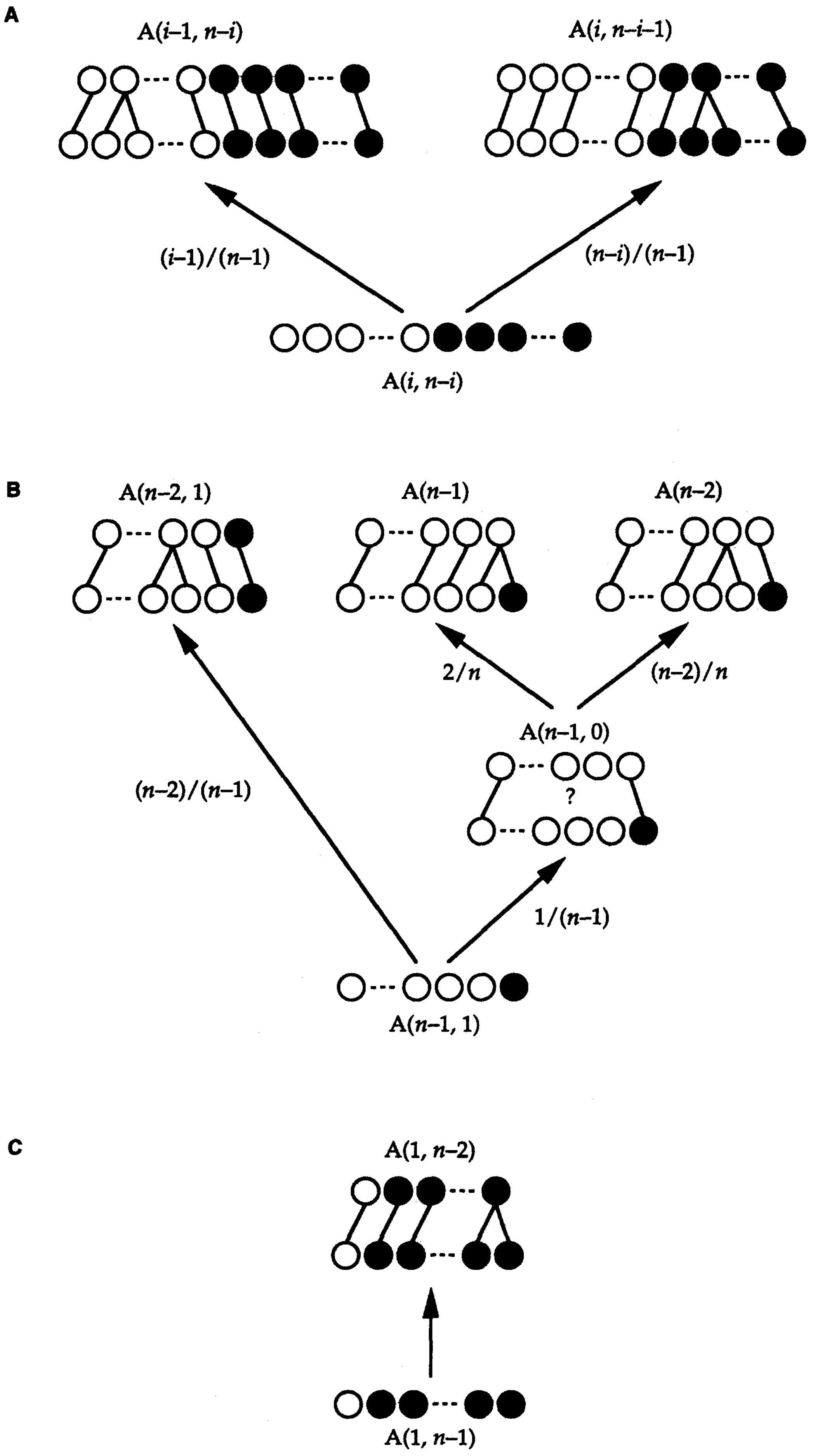

FIGURE 3.—Coalescent scheme in $A(i,n-i)$ when we know the ancestral allelic class. ○ represents the ancestral allelic class and ● represents the mutant allelic class.

TABLE 2

The expected amount of nucleotide variation when the ancestral allelic class is known

		No. of segregating sites						Average no. of pairwise differences		
i	$n-i$	$S_a(i,n-i)$	$V_a(i,n-i)$	$S_m(i,n-i)$	$V_m(i,n-i)$	$S^*_b(i,n-i)$	$V^*_b(i,n-i)$	$K_a(i,n-i)$	$K_m(i,n-i)$	$D^*(i,n-i)$
19	1	3.385θ	$3.385\theta + 1.603\theta^2$	0.000θ	$0.000\theta + 0.000\theta^2$	0.350θ	$0.350\theta + 0.517\theta^2$	0.972θ	0.000θ	1.273θ
18	2	3.221θ	$3.221\theta + 1.606\theta^2$	0.081θ	$0.081\theta + 0.016\theta^2$	0.529θ	$0.529\theta + 0.842\theta^2$	0.943θ	0.081θ	1.466θ
17	3	3.056θ	$3.056\theta + 1.601\theta^2$	0.175θ	$0.175\theta + 0.039\theta^2$	0.678θ	$0.678\theta + 1.093\theta^2$	0.913θ	0.117θ	1.623θ
16	4	2.889θ	$2.889\theta + 1.589\theta^2$	0.278θ	$0.278\theta + 0.065\theta^2$	0.809θ	$0.809\theta + 1.296\theta^2$	0.882θ	0.151θ	1.757θ
15	5	2.722θ	$2.722\theta + 1.569\theta^2$	0.386θ	$0.386\theta + 0.094\theta^2$	0.928θ	$0.928\theta + 1.466\theta^2$	0.850θ	0.184θ	1.876θ
14	6	2.553θ	$2.553\theta + 1.541\theta^2$	0.498θ	$0.498\theta + 0.123\theta^2$	1.039θ	$1.039\theta + 1.609\theta^2$	0.817θ	0.215θ	1.984θ
13	7	2.382θ	$2.382\theta + 1.504\theta^2$	0.613θ	$0.613\theta + 0.152\theta^2$	1.144θ	$1.144\theta + 1.730\theta^2$	0.783θ	0.244θ	2.082θ
12	8	2.210θ	$2.210\theta + 1.458\theta^2$	0.730θ	$0.730\theta + 0.181\theta^2$	1.244θ	$1.244\theta + 1.833\theta^2$	0.747θ	0.273θ	2.173θ
11	9	2.036θ	$2.036\theta + 1.402\theta^2$	0.849θ	$0.849\theta + 0.208\theta^2$	1.342θ	$1.342\theta + 1.921\theta^2$	0.711θ	0.301θ	2.258θ
10	10	1.860θ	$1.860\theta + 1.336\theta^2$	0.969θ	$0.969\theta + 0.233\theta^2$	1.438θ	$1.438\theta + 1.994\theta^2$	0.672θ	0.328θ	2.338θ
9	11	1.680θ	$1.680\theta + 1.259\theta^2$	1.091θ	$1.091\theta + 0.256\theta^2$	1.532θ	$1.532\theta + 2.054\theta^2$	0.632θ	0.354θ	2.413θ
8	12	1.498θ	$1.498\theta + 1.170\theta^2$	1.214θ	$1.214\theta + 0.277\theta^2$	1.627θ	$1.627\theta + 2.102\theta^2$	0.590θ	0.379θ	2.484θ
7	13	1.312θ	$1.312\theta + 1.068\theta^2$	1.339θ	$1.339\theta + 0.296\theta^2$	1.723θ	$1.723\theta + 2.138\theta^2$	0.546θ	0.403θ	2.551θ
6	14	1.122θ	$1.122\theta + 0.951\theta^2$	1.463θ	$1.463\theta + 0.311\theta^2$	1.820θ	$1.820\theta + 2.163\theta^2$	0.499θ	0.427θ	2.615θ
5	15	0.925θ	$0.925\theta + 0.817\theta^2$	1.589θ	$1.589\theta + 0.324\theta^2$	1.922θ	$1.922\theta + 2.174\theta^2$	0.449θ	0.451θ	2.677θ
4	16	0.722θ	$0.722\theta + 0.663\theta^2$	1.715θ	$1.715\theta + 0.334\theta^2$	2.029θ	$2.029\theta + 2.172\theta^2$	0.396θ	0.473θ	2.736θ
3	17	0.507θ	$0.507\theta + 0.484\theta^2$	1.842θ	$1.842\theta + 0.342\theta^2$	2.145θ	$2.145\theta + 2.154\theta^2$	0.338θ	0.496θ	2.793θ
2	18	0.273θ	$0.273\theta + 0.271\theta^2$	1.970θ	$1.970\theta + 0.345\theta^2$	2.278θ	$2.278\theta + 2.114\theta^2$	0.273θ	0.517θ	2.847θ
1	19	0.000θ	$0.000\theta + 0.000\theta^2$	2.098θ	$2.098\theta + 0.346\theta^2$	2.450θ	$2.450\theta + 2.040\theta^2$	0.000θ	0.539θ	2.900θ

$S_a(i,n-i)$, $S_m(i,n-i)$ and $S^*_b(i,n-i)$ were calculated using Equations (16)–(22); $V_a(i,n-i)$, $V_m(i,n-i)$ and $V^*_b(i,n-i)$ were calculated using Equations (D1)–(D8); $K_a(i,n-i)$, $K_m(i,n-i)$ and $D^*(i,n-i)$ were calculated using Equations (E1)–(E7).

and the probability that one sequence is the ancestor, respectively, when 59 sequences are partitioned into 58 and one sequences. $P\{(X,Y,Z) = (20,0,0) \mid (58^*,1)\}$ denotes the probability that $(X,Y,Z) = (20,0,0)$ when 58 sequences are assumed to be ancestral and $P\{(X,Y,Z) = (0,20,0) \mid (1^*,58)\}$ denotes the probability that $(X,Y,Z) = (0,20,0)$ when one sequence is assumed to be ancestral. Then, Anc{G} and Anc{A} at nucleotide position 161 are given by

$$\mathrm{Anc}\{G\} = \frac{P\{X,Y,Z) = (20,0,0) \mid (58^*,1)\} P\{(58^*,1) \mid (58,1)\}}{P\{X + Y + Z = 20;\ (58,1)\}},$$

$$\mathrm{Anc}\{A\} = \frac{P\{(X,Y,Z) = (0,20,0) \mid (1^*,58)\}\, P\{(1^*,58) \mid (58,1)\}}{P\{X + Y + Z = 20;\ (58,1)\}}.$$

It is known that the probability that an allelic class with frequency q is ancestral is q (WATTERSON and GUESS 1977). Therefore, $P\{(58^*,1) \mid (58,1)\} = 58/59$ and $P\{(1^*,58) \mid (58,1)\} = 1/59$. Since Anc{G} + Anc{A} = 1, Anc{G} and Anc{A} can be rewritten as, respectively,

$$\mathrm{Anc}\{G\} = \frac{58P\{(X,Y,Z) = (20,0,0) \mid (58^*,1)\}}{58P\{(X,Y,Z) = (20,0,0) \mid (58^*,1)\} + P\{(X,Y,Z) = (0,20,0) \mid (1^*,58)\}},$$

$$\mathrm{Anc}\{A\} = \frac{P\{(X,Y,Z) = (0,20,0) \mid (1^*,58)}{58P\{(X,Y,X) = (20,0,0) \mid (58^*,1)\} + P\{(X,Y,Z) = (0,20,0) \mid (1^*,58)\}},$$

$P\{(X,Y,Z) = (20,0,0) \mid (58^*,1)\}$ and $P\{(X,Y,Z) = (0,20,0) \mid (1^*,58)\}$ were obtained by computer simulations with 10,000 replicates. For each replicate of the simulation, a genealogical tree with 20 mutations was constructed by following HUDSON *et al.* (1994). Namely, the number of mutations on a genealogical tree was fixed at 20 and the parameter θ was not used. $P\{(X,Y,Z) = (20,0,0) \mid (58^*,1)\}$ and $P\{(X,Y,Z) = (0,20,0) \mid (1^*,58)\}$ turned out to be 0.733 and 0.002, respectively. Therefore, we have Anc{G} = 0.99995 and Anc{A} = 0.00005.

Following this way, these probabilities were obtained for the other 20 segregating sites (Table 3). In 18 sites, the ancestral nucleotide was determined at the 99% level. Out of these sites, in 13 sites the expected ancestral nucleotide is the same as that of *D. simulans*, while in the remaining five sites the ancestor is different from that of *D. simulans*. At nucleotide positions 240, 813 and 1122, we could not determine the ancestral nucleotide at the 99% level. This is partly because parallel mutations occurred in *D. melanogaster* lineage. Since recombination must be a very rare event in the mitochondrial DNA, parallel mutations should be necessary to interpret the observed pattern of polymorphism (existence of "four gametes": HUDSON and KAPLAN 1985). It is most plausible to assume that one parallel mutation occurred at the nucleotide position 1122, because we find no four gametes if nucleotide position 1122 is excluded. This parallel mutation results in the contribution of this site (nucleotide position 1122) to the num-

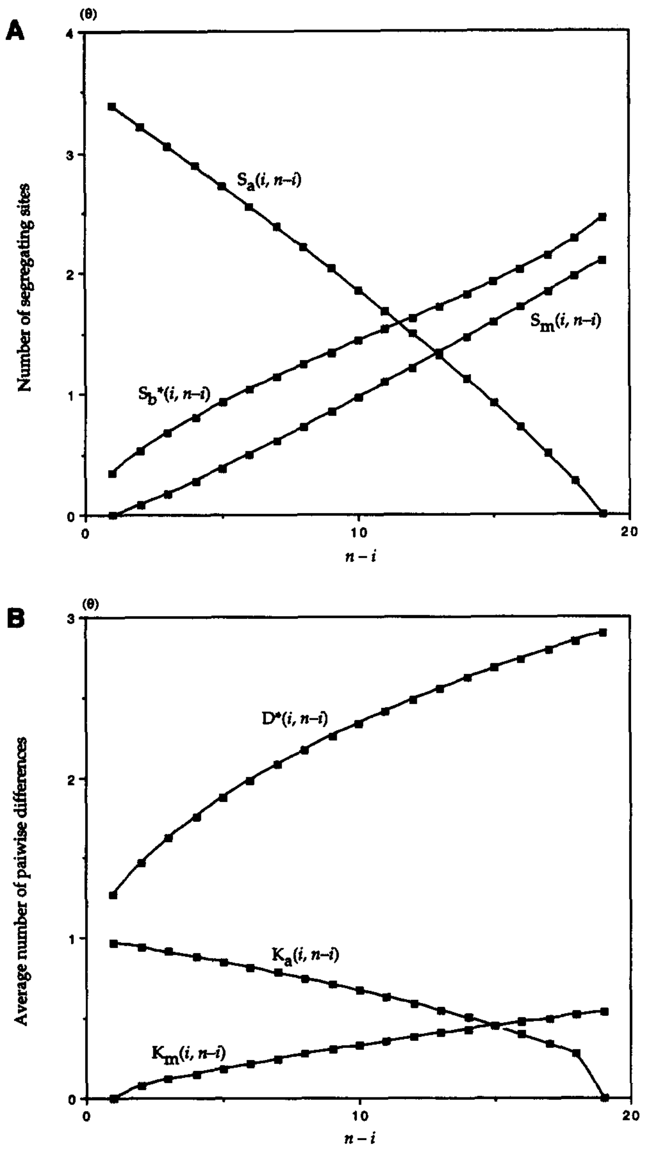

FIGURE 4.—The expected numbers of segregating sites and the expected numbers of pairwise differences with sample size $n = 20$, when the ancestral allelic class is known. The unit of the vertical axis is θ. (A) The expected numbers of segregating sites within the ancestral allelic class, within the mutant allelic class and between them. $S_a(i,n-i)$, $S_m(i,n-i)$ and $S_b^*(i,n-i)$ were calculated using Equations (16)-(22). (B) The expected numbers of pairwise differences within the ancestral allelic class, within the mutant allelic class and between them. $K_a(i,n-i)$, $K_m(i,n-i)$ and $D^*(i,n-i)$ were calculated using Equations (E1)-(E7).

bers of segregating sites within both of two allelic classes at nucleotide positions 813 and 840, where $X + Y + Z = 21$ although the total number of segregating sites is 20. Of course, at nucleotide position 1122, $X + Y + Z$ exceeds 20 ($X + Y + Z = 22$). At nucleotide position 240, since A and G are polymorphic in intermediate frequency (A: 32/59; G: 27/59), it is difficult to determine the ancestral nucleotide with high level of confidence although the observed number of segregating sites is more consistent with the expected value when G is assumed to be ancestral than is A.

Next, the nucleotide distance between *D. simulans* and *D. melanogaster* is examined (Figure 5). When we have 59 *D. melanogaster* sequences and one *D. simulans* sequence, we can estimate the average number of nucleotide differences between two species. This is illustrated in Figure 5A and the average number of nucleotide differences between two species is 6.119. Note that only the 21 polymorphic sites in *D. melanogaster* are considered here, although there are 69 differences between two species that are fixed in *D. melanogaster.* In this report, we have reconstructed the common ancestral sequence of 59 *D. melanogaster* sequences. Using this information, the nucleotide distance between two species can be reconsidered (Figure 5B). The number of nucleotide differences between *D. simulans* and *D. melanogaster* ancestral sequence is 6.005. Note that the *D. melanogaster* ancestral sequence is given by two possible nucleotides in each site with their probabilities. The average number of pairwise differences between *D. melanogaster* and its ancestral sequence is 1.667. The sum of these two values (6.005 and 1.667) is 7.672, which is larger than the average number of nucleotide differences between two species. It is suggested that there may exist a few undetected mutations between two species resulted from back and/or parallel mutations.

DISCUSSION

In this article we obtained the expected amounts of variation within and between two allelic classes with no recombination by using the theory of gene genealogy. First, we considered the case where we do not know the ancestral allelic class. It was found that the amount of variation (the number of segregating sites and the average number of pairwise differences) within an allelic class increases with its frequency (Figure 2). On the other hand, the amount of variation between two allelic classes is the largest when two allelic classes exist equally. Second, the case where the ancestor is known was considered. As shown in Figure 4, the amount of variation within the mutant allelic class increases as its frequency increases. The amount of variation within the ancestral allelic class also increases as its frequency increases, but the amount of increase is larger than that of the mutant allelic class. The amount of variation between two allelic classes increases as the frequency of the mutant allelic class increases. A notable difference between the two cases is the amount of variation between two allelic classes.

The mitochondrial DNA polymorphism in *D. melanogaster* ND5 gene region (RAND and KANN 1996) was analyzed. The common ancestral sequence of 59 *D. melanogaster* sequences was reconstructed. The common ancestral sequence was given by two possible nucleotides in each site with their probabilities as shown in

TABLE 3

Polymorphism in the ND5 gene of *Drosophila melanogaster*

		Out group: *D. simulans*										Out group: not *D. simulans*[a]									
Pos.	Poly.	Anc.[b]	i	$n-i$	X	$S_a(i,n-i)$	Y	$S_m(i,n-i)$	Z	$S_b^*(i,n-i)$	P	Anc.[b]	i	$n-i$	X	$S_a(i,n-i)$	Y	$S_m(i,n-i)$	Z	$S_b^*(i,n-i)$	P
161	G/A	G	58	1	20	19.288	0	0.000	0	0.712	1.000	A	1	58	0	0.000	20	11.205	0	8.795	0.000
218	T/C	T	58	1	20	19.288	0	0.000	0	0.712	1.000	C	1	58	0	0.000	20	11.205	0	8.795	0.000
222	A/G	A	58	1	19	19.288	0	0.000	1	0.712	0.999	G	1	58	0	0.000	19	11.205	1	8.795	0.001
240	A/G	A	32	27	8	9.729	12	4.857	0	5.414	0.118	G	27	32	12	8.244	8	5.852	0	5.904	0.882
529	T/C	T	58	1	19	19.288	0	0.000	1	0.712	0.999	C	1	58	0	0.000	19	11.205	1	8.795	0.001
657	A/G	A	58	1	20	19.288	0	0.000	0	0.712	1.000	G	1	58	0	0.000	20	11.205	0	8.795	0.000
682	T/C	T	58	1	19	19.288	0	0.000	1	0.712	0.999	C	1	58	0	0.000	19	11.205	1	8.795	0.001
687	G/A	A	1	58	0	0.000	20	11.205	0	8.795	0.000	G	58	1	20	19.288	0	0.000	0	0.712	1.000
813	T/C	T	51	8	16	16.871	5	1.218	0	2.912	0.937	C	8	51	5	2.684	16	10.211	0	8.105	0.063
840	A/G	A	57	2	19	19.675	2	0.134	0	1.191	0.999	G	2	57	2	0.470	19	11.541	0	8.990	0.001
930	A/G	A	58	1	19	19.288	0	0.000	1	0.712	0.999	G	1	58	0	0.000	19	11.205	1	8.795	0.001
1053	G/A	A	7	52	0	2.232	19	9.934	1	7.834	0.002	G	52	7	19	16.468	0	0.974	1	2.558	0.998
1062	C/T	C	58	1	20	19.288	0	0.000	0	0.712	1.000	T	1	58	0	0.000	20	11.205	0	8.795	0.000
1122	A/G	A	36	23	15	12.041	7	4.473	0	5.486	0.945	G	23	36	7	7.777	15	7.319	0	6.904	0.055
1134	C/T	T	1	58	0	0.000	20	11.205	0	8.795	0.000	C	58	1	20	19.288	0	0.000	0	0.712	1.000
1181	G/A	G	58	1	20	19.288	0	0.000	0	0.712	1.000	A	1	58	0	0.000	20	11.205	0	8.795	0.000
1239	G/A	A	2	57	0	0.447	20	10.991	0	8.561	0.000	G	57	2	20	18.738	0	0.128	0	1.135	1.000
1306	G/A	G	58	1	20	19.288	0	0.000	0	0.712	1.000	A	1	58	0	0.000	20	11.205	0	8.795	0.000
1367	A/G	A	58	1	20	19.288	0	0.000	0	0.712	1.000	G	1	58	0	0.000	20	11.205	0	8.795	0.000
1442	T/C	T	52	7	19	16.468	0	0.974	1	2.558	0.998	C	7	52	0	2.232	19	9.934	1	7.834	0.002
1479	T/G	G	1	58	0	0.000	20	11.205	0	8.795	0.000	T	58	1	20	19.288	0	0.000	0	0.712	1.000

Data for ND5 polymorphism in *Drosophila melanogaster* is from Rand and Kann (1996) and the *D. simulans* sequence is from Figure 1 in Rand *et al.* (1994). *X*, *Y* and *Z* are the observed numbers of segregating sites within the ancestral allelic class, within the mutant allelic class and between two allelic classes, respectively. $S_a(i,n-i)$, $S_m(i,n-i)$ and $S_b^*(i,n-i)$ are the expected numbers of segregating sites within the ancestral allelic class, within the mutant allelic class and between two allelic classes, respectively. Pos., position; Poly., polymorphism; Anc., ancestral;

[a] Not *D. simulans* indicates the sequence which is not common to *D. simulans* in the polymorphic site in *D. melanogaster*.

[b] Hypothetical ancestral nucleotide.

Table 3. Among the total of 21 polymorphic sites detected in the investigated region, in 18 sites the ancestral nucleotide was determined at the 99% level. It was suggested that the pattern of polymorphism within species gives useful information to know the ancestral sequence of the species.

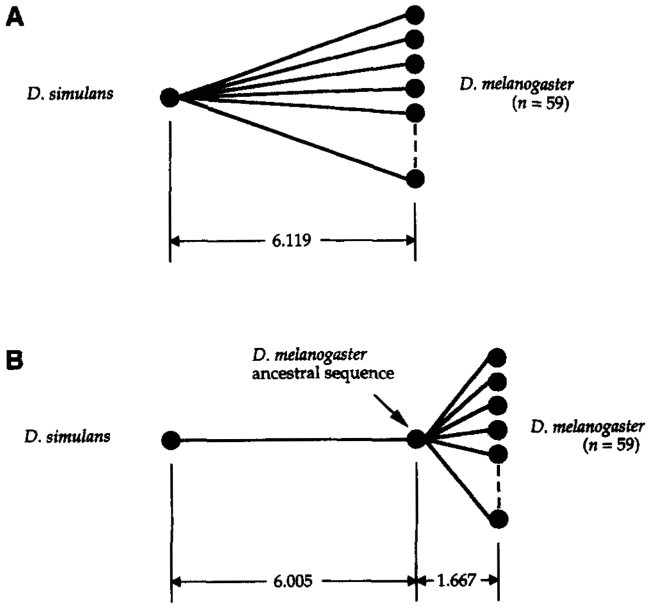

Figure 5.—The relationship between *D. melanogaster* and *D. simulans*. The number of pairwise nucleotide differences is indicated.

The nucleotide distance between *D. melanogaster* and *D. simulans* was estimated by two methods (Figure 5). The first method is simply averaging the number of pairwise differences between two species (Figure 5A), while the second one utilizes the information on the reconstructed common ancestral sequence of *D. melanogaster* (Figure 5B). The distance estimated by the first method was 6.119 and the distance estimated by the second method was 7.672, indicating that the distance from the second method is ~1.25 times larger than the first one. This difference might have been caused by back and/or parallel mutations between *D. simulans* and *D. melanogaster* ancestral sequence. Undetected mutations can be revealed by using the information on the ancestral sequence. Our result indicates that the nucleotide distance obtained without considering the intraspecies variation may lead to underestimation.

In nuclear genes, it is known that recombinations frequently occur and that the effect of recombination on the pattern of polymorphism is not small. If recombination occurs, it is difficult to know the correct amounts of variation within and between allelic classes. One recombination between two allelic classes can decrease the amount of variation between two allelic classes and increase the amounts of variation within both allelic classes, because the several segregating sites may be counted within both allelic classes. For example, consider the case where A and T are polymorphic in the first site, and G and C are polymorphic in the second site. If a recombination occurs between the two sites, four gametes (A-G, A-C, T-G and T-C) are formed. If we count the numbers of segregating sites within A, within T and between A and T, it turns out that $(X, Y, Z) = (1, 1, 0)$. However, actual number of mutations occurred is one and, if no recombination occurred, (X, Y, Z) should be $(1, 0, 0)$ or $(0, 1, 0)$. From such data, it is difficult to know the correct number of mutations that occurred within a particular allelic class. Frequent recombinations lead to more obscurity of the difference of the amounts of variation within two allelic classes. Therefore, recombinations can result in a decrease of the power to determine the ancestral sequence.

One of our motivations to promote this study is to develop a statistical test for detecting natural selection from the pattern of intraspecific variation. It is possible to know whether the neutral hypothesis holds by investigating the amounts of variation within and between allelic classes. The statistic G can be introduced to measure the degree of deviation from the neutral expectation, which is given by

$$G = \frac{[X - S_a(i,n-i)]^2}{V_a(i,n-i)} + \frac{[Y - S_m(i,n-i)]^2}{V_m(i,n-i)} + \frac{[Z - S_b^*(i,n-i)]^2}{V_b^*(i,n-i)}. \quad (23)$$

We can see the degree of deviation site by site, and we can search more likely sites where natural selection is acting. The idea has some similarity to the haplotype test developed by HUDSON *et al.* (1994). By computer simulation, the distribution of G was investigated (data not shown). It was found that the distribution depends on i and n and that no consensus distribution was detected. The only way to know the confidence limits of the distribution is to conduct computer simulation, but this test may be conservative because of a large amount of variance. Recombination can also affect the power of this test. It may be difficult to test the neutral hypothesis by using the amounts of variation within and between allelic classes.

This work was supported in part by a grant-in-aid from the Ministry of Education, Science, Sports and Culture of Japan.

LITERATURE CITED

FU, Y.-X., and W.-H. LI, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

GRIFFITHS, R. C., 1980 Lines of descent in the diffusion approximation of neutral Wright-Fisher models. Theoret. Popul. Biol. **17:** 37–50.

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

HUDSON, R. R., and N. L. KAPLAN, 1985 A statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

HUDSON, R. R., and N. L. KAPLAN, 1986 On the divergence of alleles in nested subsamples from finite populations. Genetics **113:** 1057–1076.

HUDSON, R. R., K. BAILEY, J. KWIATOWSKI and F. J. AYALA, 1994 Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics **136:** 1329–1340.

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, Cambridge.

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

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

RAND, D. M., and L. M. KANN, 1996 Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and human. Mol. Biol. Evol. **13:** 735–748.

RAND, D. M., M. DORFSMAN and L. M. KANN, 1994 Neutral and nonneutral evolution of Drosophila mitochondrial DNA. Genetics **138:** 741–756.

SLATKIN, M., 1996 Gene genealogies within mutant allelic classes. Genetics **143:** 579–587.

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

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

WATTERSON, W. A., and H. A. GUESS, 1977 Is the most frequent allele the oldest? Theoret. Popul. Biol. **11:** 141–160.

Communicating editor: G. B. GOLDING

APPENDIX A

According to the coalescent scheme in Figure 1, we can have $S_b(n)$. $S_b(n)$ is the same as the expected number of the segregating sites on an external branch in a sample of n sequences without outgroup (FU and LI 1993). When n sequences coalesce into $n-1$ sequences, a sequence coalesces with one of the remaining $n-1$ sequences with probability $(n-1)/\binom{n}{2}$ and does not coalesce with probability $\binom{n-1}{2}/\binom{n}{2}$. Then we have following recursions:

$$S_b(n) = \frac{n-2}{n} S_b(n-1) + \frac{\theta}{n(n-1)}, \quad (A1)$$

$$S_b(2) = \theta. \quad (A1a)$$

Using the above two equations, we have

$$S_b(n) = \frac{\theta}{n-1}, \quad (A2)$$

which is equivalent to that obtained by FU and LI (1993). In the same way, the variance of the number of segregating sites, $V_b(n)$, is given by

$$V_b(n) = \frac{n-2}{n} V_b(n-1) + \frac{\theta}{n(n-1)} + \frac{\theta^2}{n^2(n-1)^2}$$

$$+\frac{n-2}{n}\,[S_b(n-1)]^2-\left[\frac{n-2}{n}\,S_b(n-1)\right]^2=\frac{n-2}{n}$$

$$\times\, V_b(n-1)+\frac{\theta}{n(n-1)}+\frac{(2n-3)\theta^2}{n(n-1)^2(n-2)}, \quad \text{(A3)}$$

$$V_b(2)=\theta+\theta^2. \quad \text{(A3a)}$$

Using the above two equations, we have

$$V_b(n)=\frac{\theta}{n-1}$$

$$+\frac{1}{n(n-1)}\left[2+\sum_{k=3}^{n}\frac{2k-3}{(k-1)(k-2)}\right]\theta^2. \quad \text{(A4)}$$

APPENDIX B

The variances of the numbers of segregating sites within and between allelic classes can be derived by modifying Equations (1), (3), (7) and (8), when the ancestral allelic class is unknown. In $A(i,n-i)$, denote the variances within an allelic class and between allelic classes by $V(i,n-i)$ and $V_b(i,n-i)$, respectively. Then, $V(i,n-i)$ can be given by

$$V(i,n-i)=\frac{i}{n}\,V(i-1,n-i)+\frac{n-i}{n}\,V(i,n-i-1)$$

$$+\frac{i\theta}{n(n-1)}+\frac{i\theta^2}{n^2(n-1)^2}+\frac{2\binom{i}{2}\theta^2}{n^2(n-1)^2}$$

$$+\frac{i}{n}\,[S(i-1,n-i)]^2+\frac{n-i}{n}\,[S(i,n-i-1)]^2$$

$$-\left[\frac{i}{n}\,S(i-1,n-i)+\frac{n-i}{n}\,S(i,n-i-1)\right]^2$$

$$=\frac{i}{n}\,V(i-1,n-i)+\frac{n-i}{n}\,V(i,n-i-1)$$

$$+\frac{i\theta}{n(n-1)}+\frac{i^2\theta^2}{n^2(n-1)^2}+\frac{i(n-i)}{n^2}$$

$$\times\,[S(i-1,n-i)-S(i,n-i-1)]^2$$

$$(2\le i\le n-2), \quad \text{(B1)}$$

$$V(n-1,1)=\frac{n-1}{n}\,V(n-2,1)+\frac{1}{n}\,V(n-1)+\frac{\theta}{n}$$

$$+\frac{\theta^2}{n^2}+\frac{n-1}{n}\,[S(n-2,1)]^2+\frac{1}{n}\,[S(n-1)]^2$$

$$-\left[\frac{n-1}{n}\,S(n-2,1)+\frac{1}{n}\,S(n-1)\right]^2 \quad \text{(B2)}$$

where

$$V(n)=\sum_{k=1}^{n-1}\frac{1}{k}\,\theta+\sum_{k=1}^{n-1}\frac{1}{k^2}\,\theta^2, \quad \text{(B2a)}$$

and $V_b(i,n-i)$ can be given by

$$V_b(i,n-i)=\frac{i}{n}\,V_b(i-1,n-i)+\frac{n-i}{n}\,V_b(i,n-i-1)$$

$$+\frac{i(n-i)}{n^2}\,[S_b(i-1,n-i)-S_b(i,n-i-1)]^2$$

$$(2\le i\le n-2), \quad \text{(B3)}$$

$$V_b(n-1,1)=\frac{n-1}{n}\,V_b(n-2,1)+\frac{1}{n}\,V_b(n)+\frac{\theta}{n(n-1)}$$

$$+\frac{\theta^2}{n^2(n-1)^2}+\frac{n-1}{n}\,[S_b(n-2,1)]^2+\frac{1}{n}\,[S_b(n)]^2$$

$$-\left[\frac{n-1}{n}\,S_b(n-2,1)+\frac{1}{n}\,S_b(n)\right]^2, \quad \text{(B4)}$$

where $V_b(n)$ is given by (A4). Note that the variance of the number of mutations occurred on a branch during allelic state $A(i,n-i)$ is

$$\frac{\theta}{n(n-1)}+\frac{\theta^2}{n^2(n-1)^2}$$

and the covariance of the numbers of mutations between two branches in this time duration is

$$\frac{\theta^2}{n^2(n-1)^2}$$

(see Tajima 1983).

APPENDIX C

Denote the expected numbers of pairwise differences within and between allelic classes in $A(i,n-i)$ by $K(i,n-i)$ and $D(i,n-i)$, respectively. To obtain $K(i,n-i)$, we also follow the scheme in Figure 1 and modify Equations (1) and (3). Here, we must consider the case where two sequences in the A1 allelic class coalesce with probability $1/\binom{i}{2}$, when $A(i,n-i)$ changes to $A(i-1,n-i)$. Then, we have

$$K(i,n-i)=\frac{i}{n}\left[\frac{1}{\binom{i}{2}}\,0+\frac{\binom{i}{2}-1}{\binom{i}{2}}\,K(i-1,n-i)\right]$$

$$+\frac{n-i}{n}\,K(i,n-i-1)+\frac{2\theta}{n(n-1)}=\frac{(i+1)(i-2)}{n(i-1)}$$

$$\times\,K(i-1,n-i)+\frac{n-i}{n}\,K(i,n-i-1)+\frac{2\theta}{n(n-1)}$$

$$(2\le i\le n-2), \quad \text{(C1)}$$

$$K(n-1,1)=\frac{n-1}{n}\left[\frac{\binom{n-1}{2}-1}{\binom{n-1}{2}}\,K(n-2,1)+\frac{2\theta}{n(n-1)}\right]$$

$$+\frac{1}{n}\left[\frac{\binom{n}{2}-1}{\binom{n}{2}}\,\theta+\frac{4\theta}{n(n-1)}\right]$$

$$=\frac{n-3}{n-2}\,K(n-2,1)+\frac{(n+1)\theta}{n(n-1)}=\frac{n-1}{n}\,\theta. \quad \text{(C2)}$$

Using equations (C1) and (C2), we have a simple solution

$$K(i,n-i) = \frac{i}{n}\theta. \qquad \text{(C3)}$$

In the same way, by modifying Equations (7) and (8), the expected number of pairwise differences between two allelic classes, $D(i,n-i)$, is written as

$$D(i,n-i) = \frac{i}{n} D(i-1,n-i) + \frac{n-i}{n} D(i,n-i-1) + \frac{2\theta}{n(n-1)} \quad (2 \leq i \leq n-2), \qquad \text{(C4)}$$

$$D(n-1,1) = \frac{n-1}{n} D(n-2,1) + \frac{\theta}{n} + \frac{2\theta}{n(n-1)} = \frac{1}{n}\left(4 + \sum_{k=3}^{n} \frac{k+1}{k-1}\right)\theta. \qquad \text{(C5)}$$

From Equations (11), (C4) and (C5), $D(i,n-i)$ is rewritten as

$$D(i,n-i) = S_b(i,n-i) + \frac{n-2}{n}\theta. \qquad \text{(C6)}$$

APPENDIX D

We consider the variances of the numbers of segregating sites within and between allelic classes when the ancestral allelic class is known. Denote the variances within the ancestral allelic class, within mutant allelic class and between them by $V_a(i,n-i)$, $V_m(i,n-i)$ and $V_b^*(i,n-i)$, respectively. These are obtained by following the scheme in Figure 2 and by modifying (B1)-(B4). Namely, we have

$$V_a(i,n-i) = \frac{i-1}{n-1} V_a(i-1,n-i) + \frac{n-i}{n-1} V_a(i,n-i-1) + \frac{i\theta}{n(n-1)} + \frac{i^2\theta^2}{n^2(n-1)^2} + \frac{(i-1)(n-i)}{(n-1)^2} \times [S_a(i-1,n-i) - S_a(i,n-i-1)]^2 \quad (2 \leq i \leq n-2), \qquad \text{(D1)}$$

$$V_a(n-1,1) = \frac{n-2}{n-1} V_a(n-2,1) + \frac{1}{n-1} V(n-1) + \frac{\theta}{n} + \frac{\theta^2}{n^2} + \frac{n-2}{n-1} [S_a(n-2,1)]^2 + \frac{1}{n-1} [S(n-1)]^2 - \left[\frac{n-2}{n-1} S_a(n-2,1) + \frac{1}{n-1} S(n-1)\right]^2, \qquad \text{(D2)}$$

$$V_m(i,n-i) = \frac{i-1}{n-1} V_m(i-1,n-i) + \frac{n-i}{n-1} V_m(i,n-i-1) + \frac{n-i}{n(n-1)}\theta + \frac{(n-i)^2}{n^2(n-1)^2}\theta^2 + \frac{(i-1)(n-i)}{n-1} \times [S_m(i-1,n-i) - S_m(i,n-i-1)]^2 \quad (2 \leq i \leq n-2), \qquad \text{(D3)}$$

$$V_m(1,n-1) = V_m(1,n-2) + \frac{\theta}{n} + \frac{\theta^2}{n^2} = \sum_{k=3}^{n} \frac{1}{k}\theta + \sum_{k=3}^{n} \frac{1}{k^2}\theta^2, \qquad \text{(D4)}$$

$$V_b^*(i,n-i) = \frac{i-1}{n-1} V_b^*(i-1,n-i) + \frac{n-i}{n-1} V_b^*(i,n-i-1) + \frac{(i-1)(n-i)}{(n-1)^2} [S_b^*(i-1,n-i) - S_b^*(i,n-i-1)]^2 \quad (2 \leq i \leq n-2), \qquad \text{(D5)}$$

$$V_b^*(1,n-1) = V_b^*(1,n-2) + \frac{\theta}{n(n-1)} + \frac{\theta^2}{n^2(n-1)^2}, \qquad \text{(D6)}$$

$$V_b^*(n-1,1) = \frac{n-2}{n-1} V_b^*(n-2,1) + \frac{1}{n-1} V_b(n) + \frac{\theta}{n(n-1)} + \frac{\theta^2}{n^2(n-1)^2} + \frac{n-2}{n-1} [S_b^*(n-2,1)]^2 + \frac{1}{n-1} [S_b(n)]^2 - \left[\frac{n-2}{n-1} S_b^*(n-2,1) + \frac{1}{n-1} S_b(n)\right]^2, \qquad \text{(D7)}$$

$$V_b^*(1,1) = 2\theta + 2\theta^2. \qquad \text{(D8)}$$

APPENDIX E

Denote the expected numbers of pairwise differences within the ancestral allelic class, within the mutant allelic class and between them by $K_a(i,n-i)$, $K_m(i,n-i)$ and $D^*(i,n-i)$, respectively. Modifying (C1)-(C5), it is easy to have the following recursions:

$$K_a(i,n-i) = \frac{(i+1)(i-2)}{i(n-1)} K_a(i-1,n-i) + \frac{n-i}{n-1} \times K_a(i,n-i-1) + \frac{2\theta}{n(n-1)} \quad (2 \leq i \leq n-2), \qquad \text{(E1)}$$

$$K_a(n-1,1) = \frac{n(n-3)}{(n-1)^2} K_a(n-2,1) + \frac{n+2}{n(n-1)}\theta = \frac{n}{(n-1)(n-2)} \left[\sum_{k=3}^{n} \frac{(k+2)(k-2)}{k^2}\right]\theta, \qquad \text{(E2)}$$

$$K_m(i,n-i) = \frac{i-1}{n-1} K_m(i-1,n-i) + \frac{(n-i-2)(n-i+1)}{(n-1)(n-i-1)} K_m(i,n-i-1) + \frac{2\theta}{n(n-1)} \quad (2 \leq i \leq n-2), \qquad \text{(E3)}$$

$$K_m(1,n-1) = \frac{n(n-3)}{(n-1)(n-2)} K_m(1,n-2) + \frac{2\theta}{n(n-1)} = \frac{n}{(n-2)} \sum_{k=3}^{n} \frac{2(k-2)}{k^2(k-1)}\,\theta, \qquad \text{(E4)}$$

$$D^*(i,n-i) = \frac{i-1}{n-1} D^*(i-1,n-i) + \frac{n-i}{n-1} \times D^*(i,n-i-1) + \frac{2\theta}{n(n-1)} \quad (2 \leq i \leq n-2), \qquad \text{(E5)}$$

$$D^*(n-1,1) = \frac{n-2}{n-1}\left[D^*(n-2,1) + \frac{2\theta}{n(n-1)} \right. + \frac{n+2}{n(n-1)}\,\theta = \frac{1}{n-1}\left[n + 2\sum_{k=3}^{n} \frac{1}{k} \right]\theta, \qquad \text{(E6)}$$

$$D^*(1,n-1) = D^*(1,n-2) + \frac{2\theta}{n(n-1)} = \left(3 - \frac{2}{n}\right)\theta. \qquad \text{(E7)}$$