# Coalescent Theory for a Partially Selfing Population

**Yun-Xin Fu**

*Human Genetics Center, University of Texas, Houston, Texas 77225*

Manuscript received October 16, 1996
Accepted for publication April 21, 1997

ABSTRACT

A coalescent theory for a sample of DNA sequences from a partially selfing diploid population and an algorithm for simulating such samples are developed in this article. Approximate formulas are given for the expectation and the variance of the number of segregating sites in a sample of $k$ sequences from $n$ individuals. Several new estimators of the important parameters $\theta = 4N\mu$ and the selfing rate $s$, where $N$ and $\mu$ are, respectively, the effective population size and the mutation rate per sequence per generation, are proposed and their sampling properties are studied.

THE coalescent theory (KINGMAN 1982a,b; HUDSON 1983; TAJIMA 1983; and see HUDSON 1991 for a review) is becoming the cornerstone for utilizing the information in a sample of DNA sequences from a population to study its evolution. Although a number of population genetics models have been studied in the framework of coalescence, it remains a common assumption that mating between individuals is at random, despite that inbreeding is not rare in nature. A completely selfing population is an extreme example of inbreeding, while many plant populations exhibit partial self-fertilization, that is, a mixture of selfing and random mating (FRYXELL 1957; WILLSON 1984). The purpose of this article is to develop a coalescent theory for a sample from a partially selfing population.

I shall use WRIGHT's (1969) definition of partial selfing throughout this article, which assumes that in a diploid population of $N$ individuals each individual has probability $s$ (the selfing rate) of being the offspring of a self-fertilization and probability $1 - s$ of being the offspring of a random mating. Furthermore, an individual resulting from selfing has probability $1/N$ of being the offspring of any particular individual of the previous generation, and an individual resulting from random mating is formed by randomly selecting two sequences from the gene pool of $2N$ sequences of the previous generation. We also assume in this article that random genetic drift and partial selfing are the only evolutionary forces in action. In other words, there is no selection, no population subdivision and migration, and no recombination.

Although the coalescent theory is a relatively young branch of population genetics, the concept of identity by descent in classical population genetics theory has essentially the same spirit as the coalescence. Indeed, SLATKIN (1992) tried to link the probability of identical by descent, the inbreeding coefficient, to coalescent time and MILLIGAN (1996) utilized SLATKIN's approach to derive an estimator of the selfing rate. However, "identical by descent" has a fundamental weakness: it deals with only two sequences at a time and by doing so, much information in a sample is wasted. The coalescent approach deals with the whole sample of sequences and allows efficient use of available information (FELSENSTEIN 1992; FU and LI 1993). Furthermore, coalescent algorithms provide highly efficient methods to simulate samples from a population. The coalescent theory and a simulation algorithm developed in this article enable us to find approximate formulas for the expectation and variance of the number of segregating sites in a sample, and consequently to explore estimators of the mutation rate (or population size) and the selfing rate.

*Corresponding author:* Yun-Xin Fu, Human Genetics Center, University of Texas at Houston, 6901 Bertner Ave., Houston, TX 77030. E-mail: fu@hgc.sph.uth.tmc.edu

## THE THEORY

**Definitions:** The essence of coalescent approach is that one starts with the sequences in a sample and traces backward in generations until the most recent common ancestor of the sample is found. To characterize the coalescent process, one need keep track of only the number of ancestral sequences of the sample from a random mating population, but for a partially selfing population, one need keep track of not only the number of ancestral sequences but also how the sequences are packed into ancestors, because the two ancestral sequences in one individual have different fates from two ancestral sequences from different individuals. Let the present generation of a population be designated as generation 0 and the time $t$ generations earlier as generation $t$. Let $k$ be the number of ancestral sequences and $n$ be the number of ancestors into whom these $k$ sequences are packed at generation $t$. We call $(k, n)$ the state of coalescence at generation $t$. In particular, we refer to $(k, n)$ at generation 0, *i.e.*, the generation at which the sample was taken, as the sample con-

figuration. For diploid organisms, as we assume in this paper, there are $k - n$ ancestors whose both sequences are descended in the sample, and there are $2n - k$ ancestors with only one sequence descended in the sample. Obviously we must have $n \leq k \leq 2n$.

An ancestor of the sequences in a sample (and simply referred to as an ancestor hereafter) is an individual at certain generation whose sequence(s) have descendant(s) in the sample. We distinguish two types of ancestors. An ancestor is a *full ancestor* if both of its sequences have descendant(s) in the sample; an ancestor is a *half ancestor* if only one of its sequences has descendant(s) in the sample. For convenience, we also treat the individuals in the sample as ancestors. It follows that when the state of coalescence at generation $t$ is $(k, n)$, there are $k - n$ full ancestors and $2n - k$ half ancestors in that generation.

To describe the coalescent process for a partially selfing population, it is convenient to divide ancestral sequences into groups that will be called "effective sequences." We define each ancestral sequence at generation $t$ in an ancestor that resulted from random mating as *one effective sequence*, and the ancestral sequence(s) in each ancestor that resulted from selfing as *one effective sequence*. Therefore, an effective sequence may represent one or two ancestral sequences. According to the definition, the two sequences of a full ancestor are each an effective sequence if the full ancestor is derived from random mating, but are counted as only one effective sequence if the full ancestor is derived from selfing. The reason for them being counted as only one effective sequence is that they have the same probability of coming from a particular individual of the previous generation as a single sequence resulting from random mating.

For a random mating population, a coalescent event usually means that two ancestral sequences at a generation come from a single ancestral sequence at previous generation (*e.g.*, HUDSON 1991). For a partially selfing population, we need to broaden the definition of a coalescent event and we recognize three different types of coalescent events. First, two effective sequences at generation $t$ coalesce to a full ancestor at generation $t + 1$. That is, two effective sequences come from the two sequences of a full ancestor at the previous generation. Second, two effective sequences coalesce to a half ancestor at the previous generation. Third, the two sequences of a single effective sequence in an individual derived by selfing coalesce to a single ancestral sequence at the previous generation.

We note that when two effective sequences each represent a single sequence, the second type of coalescent event defined above corresponds to the traditional definition of a coalescent event, but the first type of coalescent event defined above is not considered as a coalescent event for a random mating population. The third type of coalescent event is unique to a partially selfing population because it occurs only when a full ancestor is the product of selfing. Each of the first two types of coalescent events involves two effective sequences, while the third type involves only one effective sequence. For brevity, a coalescent event in this article means a coalescent event of either type one or type two, unless the third type of coalescent event is stated explicitly.

**Probability of the number of effective sequences:** Suppose the state of coalescence at generation $t$ is $(k, n)$. Then we have $k - n$ full ancestors at the generation. If $i$ of the $k - n$ full ancestors are results of selfing, the number of effective sequences is $k - i$ according to our definition. Because the probability that a full ancestor is the product of selfing is equal to the selfing rate $s$, the probability that there are $i$ full ancestors resulting from selfing is thus given by the binomial probability

$$\binom{k-n}{i} s^i (1-s)^{k-n-i}, \tag{1}$$

which is also the probability of having $k - i$ effective sequences.

**Change of state when there is no coalescent event:** Each ancestral sequence of the sample at generation $t$ has the probability $1/N$ of coming from a particular individual at generation $t + 1$ regardless of whether the sequence is a result of random mating or a result of selfing. However, when the latter is true, the other allele of the same ancestor also comes from the same individual at generation $t + 1$. That is, each effective sequence has the probability $1/N$ of coming from a particular individual of the previous generation. This result implies that once the effective sequences are determined, there is no need to consider if an effective sequence is from selfing or from random mating. Consequently, the probability of no coalescence in one generation between effective sequences, given there are $k - i$ effective sequences at generation $t$, is

$$\frac{N-1}{N}\frac{N-2}{N}\cdots\frac{N-(k-i)+1}{N} = 1 - \frac{(k-i)(k-i-1)}{2N} + O\left(\frac{1}{N^2}\right).$$

Neglecting terms of higher order, *i.e.*, assuming the probability of more than one coalescent event is negligible, we have that the probability of no coalescent event between effective sequences is

$$1 - \frac{(k-i)(k-i-1)}{2N}.$$

The probability of one coalescent event is thus $(k - i)(k - i - 1)/(2N)$.

The state $(k, n)$ can change in one generation even when there is no coalescent event between effective sequences. The value of $n$ will change because all full ancestors resulting from random mating change to two half ancestors at generation $t + 1$; on the other hand,

the value of $k$ may change if a full ancestor resulting from selfing becomes a half ancestor (*i.e.*, a coalescent event of third type) at generation $t + 1$, which has probability $^1/_2$. Given there are $i$ full ancestors resulting from selfing, the probability that $l$ of them become half ancestors in generation $t + 1$ is given by

$$\binom{i}{l}\left(\frac{1}{2}\right)^i,$$

and the number of ancestors at generation $t + 1$ becomes $i + 2(k - n - i) + 2n - k = k - i$. Therefore conditioning on there being $i$ full ancestors from selfing at generation $t$, the probability that there is no coalescence between effective sequences and that state $(k, n)$ changes to $(k - l, k - i)$ at generation $t + 1$ is

$$\left[1 - \frac{(k-i)(k-i-1)}{2N}\right]\binom{i}{l}\left(\frac{1}{2}\right)^i \qquad (2)$$

for $0 \le l \le i \le n$.

**Change of state when there is a coalescent event:** When a coalescent event occurs between two effective sequences, two randomly chosen effective sequences coalesce to an ancestor at generation $t + 1$. Because an effective sequence may represent either one or two sequences, the two effective sequences may represent two, three, or four sequences. Let $p_j (2 \le j \le 4)$ be the probability that they represent $j$ sequences. Then conditioning on there being $i$ full ancestors resulting from selfing, $p_2$, $p_3$ and $p_4$ are equal to

$$\frac{(k-2i)(k-2i-1)}{(k-i)(k-i-1)}, \quad \frac{2i(k-2i)}{(k-i)(k-i-1)}, \quad \frac{i(i-1)}{(k-i)(k-i-1)},$$

respectively. The ancestor of the two effective sequences at generation $t + 1$ may be a half ancestor or a full ancestor. The former event has the probability

$$q_j = \left(\frac{1}{2}\right)^{j-1},$$

because each of the $j - 1$ sequences has probability $^1/_2$ sharing the same ancestral sequence of the first sequence. The latter event has the probability $1 - q_j$. We note that when there are $i$ full ancestors resulting from selfing and there is a coalescent event, the number of ancestors at generation $t + 1$ becomes $k - i - 1$. There are two ways the number of sequences can be reduced from $k$ to $k - l$ given that the two effective sequences represent $j$ sequences. The first is that the two effective sequences coalesce to a full ancestor and $l - j + 1$ of the $i - j + 2$ full ancestors resulting from selfing ($j - 2$ are already involved in coalescence between effective sequences) at generation $t$ become half ancestors at generation $t + 1$. The probability of this event is

$$q_j\binom{i-j+2}{l-j+1}\left(\frac{1}{2}\right)^{i-j+2}.$$

The second is that the two effective sequences coalesce to a half ancestor and $l - j + 2$ of the full ancestors at generation $t$ become half ancestors at generation $t + 1$. The probability of this event is

$$(1 - q_j)\binom{i-j+2}{l-j+2}\left(\frac{1}{2}\right)^{i-j+2}.$$

Therefore, conditioning on there being $i$ full ancestors resulting from selfing, the joint probability that there is a coalescent event and that the state $(k, n)$ changes to $(k - l, k - i - 1)$ is

$$\begin{aligned}
&\frac{(k-i)(k-i-1)}{2N}\sum_{j=2}^{4} p_j\left[q_j\binom{i-j+2}{l-j+1}\left(\frac{1}{2}\right)^{i-j+2}\right.\\
&\quad\left.+ (1-q_j)\binom{i-j+2}{l-j+2}\left(\frac{1}{2}\right)^{i-j+2}\right]\\
&= \frac{1}{2N}\left(\frac{1}{2}\right)^{i+1}\left\{(k-2i)(k-2i-1)\right.\\
&\quad\times\left[\binom{i}{l}+\binom{i}{l-1}\right] + 2i(k-2i)\\
&\quad\times\left[3\binom{i-1}{l-1}+\binom{i-1}{l-2}\right]\\
&\quad\left.+ i(i-1)\left[7\binom{i-2}{l-2}+\binom{i-2}{l-3}\right]\right\} \qquad (3)
\end{aligned}$$

for $0 \le l \le i$.

**Probability of a transition of state:** Putting (1–3) together and noting that to change from $n$ ancestors at generation $t$ to $k - i$ ancestors at generation $t + 1$ through a coalescent event, there must be $i - 1$ full ancestors resulting from selfing at generation $t$, we have that the probability of changing state $(k, n)$ at generation $t$ to state $(k - l, k - i)$ at generation $t + 1$ is

$$\begin{aligned}
p[(k,n)\to(k-l,k-i)] &= \binom{k-n}{i}s^i(1-s)^{k-n-i}\\
&\quad\times\left[1-\frac{(k-i)(k-i-1)}{2N}\right]\binom{i}{l}\left(\frac{1}{2}\right)^i\\
&\quad+\binom{k-n}{i-1}s^{i-1}(1-s)^{k-n-i+1}\frac{1}{2N}\left(\frac{1}{2}\right)^i\\
&\quad\times\left\{(k-2i+2)(k-2i+1)\right.\\
&\quad\times\left[\binom{i-1}{l}+\binom{i-1}{l-1}\right] + 2(i-1)(k-2(i-1))\\
&\quad\times\left[3\binom{i-2}{l-1}+\binom{i-2}{l-2}\right] + (i-1)(i-2)\\
&\quad\left.\times\left[7\binom{i-3}{l-2}+\binom{i-3}{l-3}\right]\right\} \qquad (4)
\end{aligned}$$

for $0 \le l \le i$ and $0 \le i \le k - n + 1$, with the convention that $\binom{i}{j} = 0$ whenever $i < j$ or $i < 0$ or $j < 0$.

The transition probability (4) and the sample configuration $(k, n)$ completely specify the Markov chain for the coalescence process. When the number of ancestors changes from $n$ to $k - i$ $(i = 0, \ldots, k - n + 1)$, the number of sequences can be any value between $k - i$ and $k$. Therefore the number of states to which $(k, n)$ can change in one generation is thus

$$1 + 2 + \cdots + (k - n + 2) = \frac{(k - n + 2)(k - n + 3)}{2}.$$

We can see that the number of possible states of the Markov chain is considerably larger than that for a random mating population. Among the states, $(k, k)$ is of special importance because once the Markov chain enters this state, it tends to stay there, because from (4) the probability that $(k, k)$ remains unchanged in one generation is

$$1 - \frac{k(k-1)}{2N},$$

which is large unless $k$ is very large. We also have

$$p[(k, n) \to (k, k)] = (1 - s)^{k-n}\left[1 - \frac{k(k-1)}{2N}\right],$$

so the probability of entering the state $(k, k)$ increases with $n$. The probability that state $(k, n)$ remains unchanged in one generation is

$$\left(\frac{1}{2}\right)^{k-n}\left\{s^{k-n}\left[1 - \frac{n(n-1)}{2N}\right] + (k-n)\,s^{k-n-1}(1-s)\,\frac{(k - 2(k-n-1))(k - 2(k-n-1) - 1)}{2N}\right\},$$

which is smaller than $2^{n-k}$.

**Simulation algorithm:** A coalescent algorithm for simulating a genealogy of a sample starts with the sample configuration, then moves backward in time and changes state stochastically along the way until the most recent common ancestral sequence (MRCA) is found. Given that sample configuration is $(k, n)$, the selfing rate is $s$ and the effective population size is $N$, the analysis in the previous sections suggests that we can simulate a sample genealogy by the following steps:

1. Go to step 3 if there is at least one full ancestor, otherwise go to step 2.
2. Simulate the time length before the state $(m, m)$ changes, then change it to the state $(m, m - 1)$ or the state $(m - 1, m - 1)$ randomly and go to step 7.
3. Determine for each full ancestor whether it is an offspring from selfing, and thus determine each effective sequence.
4. Determine if there is a coalescent between effective sequences.
5. If there is coalescent event, randomly select two effective sequences and determine if they coalesce to a full ancestor or a half ancestor.
6. Determine for each full ancestor resulting from selfing whether a coalescent event of the third type occurs.
7. Update information and return to the first step if $k > 1$ and stop otherwise.

We can see from the algorithm that most of the simulation cycles move one generation backward at a time unless the state is $(m, m)$ (*i.e.*, the number of ancestral sequences is equal to the number of ancestors). The time length before the state $(m, m)$ changes follows approximately the exponential distribution

$$\frac{2N}{m(m-1)}\exp\left[-\frac{m(m-1)}{2N}t\right].$$

When it does change, it has equal probability changing to the state $(m, m - 1)$ and the state $(m - 1, m - 1)$.

The above algorithm simulates only the path from the MRCA to the sequences in the sample. To simulate a sample of DNA sequences from a partially selfing population, one can superimpose mutations in each branch of a simulated genealogy, as in the coalescent algorithm for simulating DNA samples from a random mating population. A *C* subroutine that generates a genealogy of a sample of any configuration is available upon request.

Examining a few examples of the simulated genealogies of a sample from a partially selfing population should be interesting, and we give two examples in Figure 1. Figure 1a shows a simulated genealogy of a sample with $(k, n) = (6, 3)$ from a population with low selfing rate $(s = 0.2)$, and Figure 1b from a population with high selfing rate $(s = 0.9)$. In the case of low selfing rate, it happened that among the three full ancestors only one was an offspring of selfing, whose sequences coalesced to a single ancestral sequence at previous generation. After this event, the rest of the genealogy back to the MRCA was similar to that of a sample of five sequences from a completely random mating population. In the case of high selfing rate, it happened that all the three full ancestors were from selfing, in one generation two of them became half ancestors and the third one remained a full ancestor, who was also from selfing, after one more generation the full ancestor became a half ancestor and from then on the genealogy looked similar to that of a sample of three sequences from a random mating population.

It is interesting to see that the number of generations during which there were three ancestral sequences (the so called three-coalescent time) in the case of low selfing rate is considerably larger than in the case of high selfing, and the same is true for the two-coalescent time. Although the stochastic nature of the coalescent process may account for the pattern, it turns out that this is expected, as we shall see later.

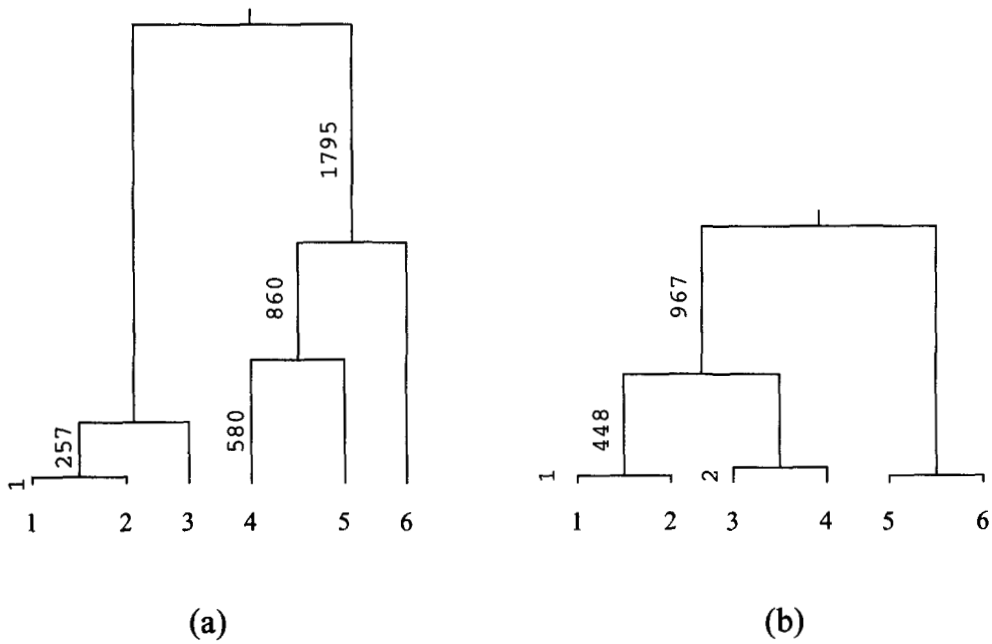


FIGURE 1.—Simulated genealogies of a sample with $(k, n) = (6, 3)$ from a partially selfing population with $N = 1000$. (a) $s = 0.2$ and (b) $s = 0.9$. Branch lengths are in terms of the numbers of generations.

## THE NUMBER OF SEGREGATING SITES

Let $S_{k,n}$ be the number of segregating sites in a sample of configuration $(k, n)$. For a random mating population, it does not matter how many individuals the $k$ sequences are from, and it is well known (WATTERSON 1975) that

$$E(S_{k,n}) = a_k\theta \tag{5}$$

$$\mathrm{Var}(S_{k,n}) = a_k\theta + b_k\theta^2, \tag{6}$$

where $\theta = 4N\mu$ where $\mu$ is mutation rate per sequence per generation, and

$$a_k = 1 + \frac{1}{2} + \cdots + \frac{1}{k-1} \tag{7}$$

$$b_k = 1 + \frac{1}{4} + \cdots + \frac{1}{(k-1)^2}. \tag{8}$$

In the case of a completely selfing population, the two sequences of an full ancestor always come from a single ancestor and the mean time for them to coalesce to a single ancestral sequence is two generations. Since the coalescent between two effective sequences takes many more generations, the mean number $T_{k,n}$ of generations in the sample genealogy is mainly due to the times between coalescent events between effective sequences. Therefore, $T_{k,n}$ is approximately the mean time length of the genealogy of a sample of $n$ sequences from a haploid population with effective population size $N$, that is,

$$T_{k,n} \approx a_n(2N).$$

Assuming the number of mutations in each branch of the sample genealogy follows a Poisson distribution with mean equal to the product of the mutation rate $\mu$ and the branch length (generations), we then have

$$E(S_{k,n}) = \tfrac{1}{2}\, a_n\theta, \tag{9}$$

where $\theta = 4N\mu$.

For a partially selfing population, $E(S_{k,n})$ must lie between (5) and (9). We note that $T_{k,n}$'s satisfy the recurrent equation

$$T_{k,n} = k + \sum_{i=0}^{k-n+1} \sum_{l=0}^{i} p[(k, n) \to (k-l, k-i)]\, T_{k-l,k-i}. \tag{10}$$

In principle, one can solve these linear equations for the value of $T(k, n)$. It is simple to do so in the case of two sequences because

$$T_{2,2} = \frac{N-1}{N}\, T_{2,2} + \frac{1}{2N}\, T_{2,1} + 2$$

$$T_{2,1} = (1-s)\,\frac{N-1}{N}\, T_{2,2} + (1-s)\,\frac{1}{2N}\, T_{2,1} + s\,\frac{1}{2}\, T_{2,1} + 2,$$

and from which we have

$$T_{2,1} = (1-s)(4N) + 4s$$

$$T_{2,2} = (1-s)(4N) + s(2N) + 2s.$$

Neglecting the terms $2s\mu$ and $4s\mu$, we have

$$E(S_{2,1}) = (1-s)\theta$$

$$E(S_{2,2}) = [\tfrac{1}{2}\, s + (1-s)]\theta,$$

which can be written as

$$E(S_{2,1}) = [\tfrac{1}{2}\, a_1 s + (1-s)\, a_2]\theta$$

$$E(S_{2,2}) = [\tfrac{1}{2}\, a_2 s + (1-s)\, a_2]\theta.$$

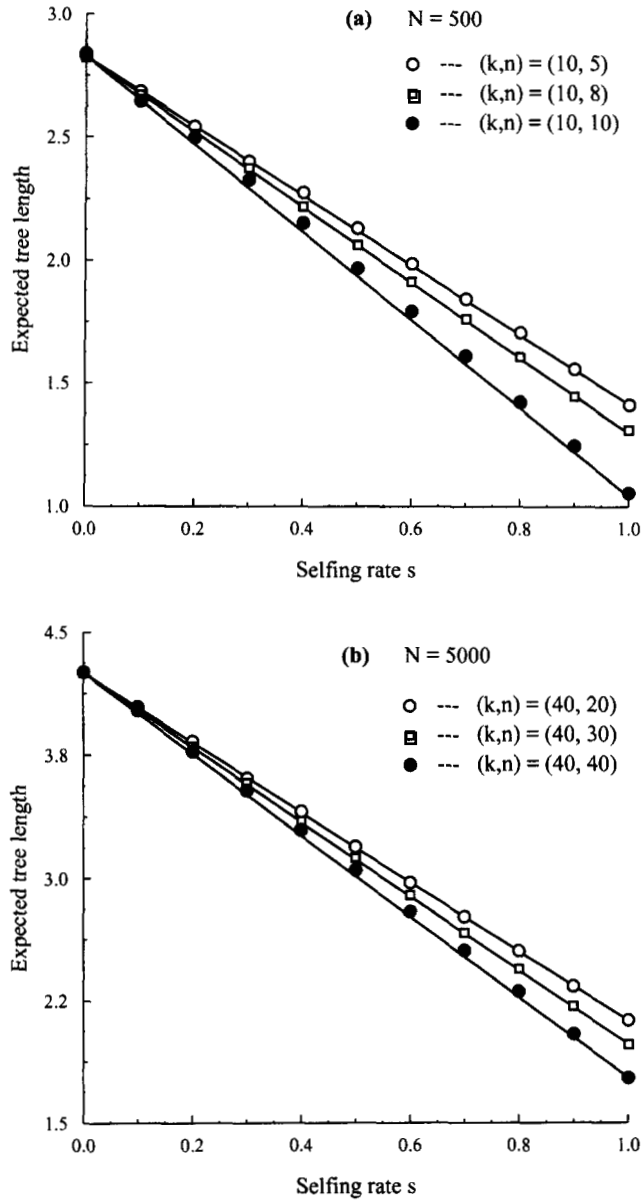


FIGURE 2.—Mean values of tree length ($T_{k,n}/(4N)$) with respect to the selfing rate $s$. Points are simulation results and lines are given by Equation 12. 20,000 genealogies were simulated for each value of $s$ for each sample configuration.

These two equations, together with (5) and (9), suggest the following equation:

$$E(S_{k,n}) = [\tfrac{1}{2}\, a_n s + a_k(1 - s)]\theta. \qquad (11)$$

To verify this equation analytically does not appear simple, we thus turn to computer simulations for help. In our simulations, we used the algorithm developed earlier to generate genealogies of a sample of a given configuration. For each simulated genealogy, we recorded the value of $T_{k,n}$ and obtained the mean value of $T_{k,n}$ after many replicates. Note that for the purpose of verifying (11), there is no need to superimpose mutations to the genealogies, because if we show

$$T_{k,n}/(4N) = \tfrac{1}{2}\, a_n s + a_k(1 - s), \qquad (12)$$

then (11) must be correct. Figure 2 shows examples of $T_{k,n}$ for several sample configurations and effective population sizes. We can see that (12) and simulations agree quite well, suggesting that (11) is an excellent approximation if not entirely correct.

Since there is a linear relationship between $E(S_{k,n})$ and the selfing rate $s$, it is tempting to suggest

$$\mathrm{Var}(S_{k,n}) = \left[ a_n \frac{\theta}{2} + b_n \left(\frac{\theta}{2}\right)^2 \right] s + [a_k\theta + b_k\theta^2](1 - s), \qquad (13)$$

but simulations show this equation overestimates the variance of $S_{k,n}$ for a partially selfing population. Since we assume the number of mutations in each branch of the sample genealogy follows a Poisson distribution, the variance of $S_{k,n}$ is equal to

$$V(S_{k,n}) = E(S_{k,n}) + \frac{\mathrm{Var}(T_{k,n})}{(4N)^2}\theta^2.$$

Because $E(T_{k,n})$ is a linear function of $s$, it is likely that the coefficient, $\mathrm{Var}(T_{k,n}/(4N^2))$, of $\theta^2$ is a quadratic function of $s$. I find that subtracting $b_n s(1 - s)\theta^2/4$ from (13) results an excellent approximation to the variance (see Figure 3). Therefore, I suggest using the following equation:

$$\begin{aligned}\mathrm{Var}(S_{k,n}) &= \left[ a_n \frac{\theta}{2} + b_n \left(\frac{\theta}{2}\right)^2 \right] s + [a_k\theta + b_k\theta^2] \\ &\quad \times (1 - s) - \frac{b_n}{4} s(1 - s)\theta^2 \\ &= \left[\frac{1}{2}\, a_n s + a_k(1 - s)\right]\theta \\ &\quad + \left[ b_k(1 - s) + \frac{b_n}{4} s^2 \right]\theta^2. \qquad (14)\end{aligned}$$

Figure 3 also shows that $\mathrm{Var}(T_{k,n})/(4N^2)$ differs little among different values of $n$. This reflects the fact that $b_n$ converge rapidly, and the difference between $b_n$ and $b_k$ is rather small.

We therefore have approximate formulas for the mean and variance of $S_{k,n}$. For practical purposes, (11) and (14) may be sufficiently accurate, but since the number of segregating sites in a sample is an important quantity, it will be useful to obtain exact results for $E(S_{k,n})$ and $\mathrm{Var}(S_{k,n})$ in the future.

## ESTIMATION OF $\theta$ AND $s$

Equation 11 can be used for estimating the mutation parameter $\theta$ and the selfing rate $s$ as follows. Let $(i, j)$ and $(l, m)$ be different, *i.e.*, at least different for one component. We then have

$$\begin{pmatrix} E(S_{i,j}) \\ E(S_{l,m}) \end{pmatrix} = \begin{pmatrix} \tfrac{1}{2}\, a_j & a_i \\ \tfrac{1}{2}\, a_m & a_l \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \qquad (15)$$

where $\theta_1 = s\theta$ and $\theta_2 = (1 - s)\theta$. Solving these equations yields

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} = \frac{2}{a_j a_l - a_i a_m}\begin{pmatrix} a_l & -a_i \\ \frac{-1}{2}a_m & \frac{1}{2}a_j\end{pmatrix}\begin{pmatrix} E(S_{i,j})\\ E(S_{l,m})\end{pmatrix}. \quad (16)$$

Since

$$\theta = \theta_1 + \theta_2$$

$$s = \frac{\theta_1}{\theta_1 + \theta_2},$$

we have

$$\theta = \frac{2}{a_j a_l - a_i a_m}\left[\left(a_l - \frac{1}{2}a_m\right)E(S_{i,j}) - \left(a_i - \frac{1}{2}a_j\right)E(S_{l,m})\right] \quad (17)$$

$$s = \frac{a_l E(S_{i,j}) - a_i E(S_{l,m})}{(a_l - {}^1/_2 a_m)\,E(S_{i,j}) - (a_i - {}^1/_2\, a_j)\,E(S_{l,m})}. \quad (18)$$

The implication of these equations is that when a sample of configuration $(k, n)$ is taken, one can resample the sample in two different ways, and when substituting $E(S_{i,j})$ and $E(S_{l,m})$ by appropriate sample means, we obtain an estimator of $\theta$ and an estimator of $s$. A simple scheme is found by setting $(i, j) = (2, 2)$ and $(l, m) = (2, 1)$, *i.e.*, only two sequences are considered at a time. Let $\bar{S}_{2,2}$ and $\bar{S}_{2,1}$ be, respectively, the average numbers of nucleotide differences between and within individuals. Then substituting $\bar{S}_{2,2}$ for $E(S_{2,2})$ and $\bar{S}_{2,1}$ for $E(S_{2,1})$, we have the following pair of estimators:

$$\hat{\theta}_m = 2\bar{S}_{2,2} - \bar{S}_{2,1} \quad (19)$$

$$\hat{s}_m = 2\,\frac{\bar{S}_{2,2} - \bar{S}_{2,1}}{2\bar{S}_{2,2} - \bar{S}_{2,1}}. \quad (20)$$

We note that $\hat{\theta}_m$ is identical to MILLIGAN's (1996) estimator of $\theta$ while $\hat{s}_m$ differs from MILLIGAN's (1996) estimator of $s$ by a factor $N/(N-1)$. Since $N/(N-1)$ is usually very close to 1, $\hat{s}_m$ is essentially the same estimator as MILLIGAN's. Because of (11), $\hat{\theta}$ is nearly unbiased estimator.

An alternative scheme of estimation is to use the pair $S_{k,n}$ and $\bar{S}_{2,1}$, that is, the number of segregating sites in the sample and the mean number of nucleotide differences within individual. From (17) and (18), we have

$$\hat{\theta}_a = \frac{2}{a_n}\left[S_{k,n} - \left(a_k - \frac{1}{2}a_n\right)\bar{S}_{2,1}\right] \quad (21)$$

$$\hat{s}_a = \frac{S_{k,n} - a_k\bar{S}_{2,1}}{S_{k,n} - (a_k - {}^1/_2 a_n)\,\bar{S}_{2,1}}. \quad (22)$$

It is easy to show using (11) that $\hat{\theta}_a$ is nearly unbiased. Using simulated samples, we found (11) is indeed unbiased. However, our simulation study shows that $\hat{s}_a$ is not a very reliable estimator, so we shall not discuss it further.

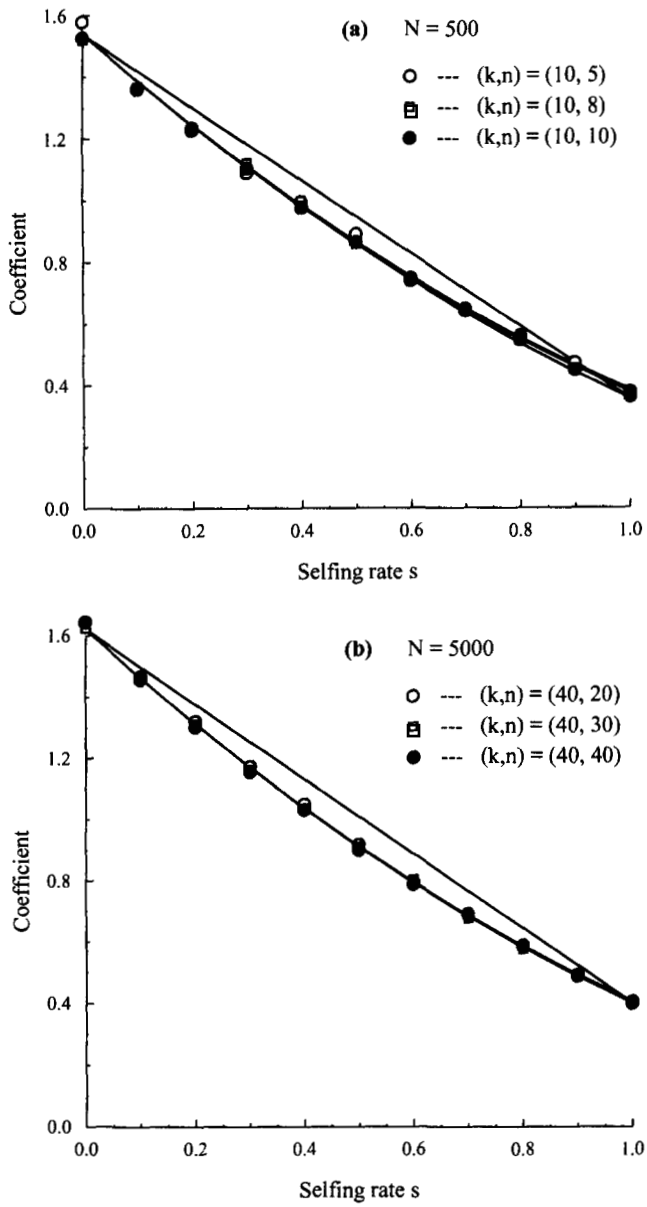


FIGURE 3.—Coefficient of $\theta^2$ in $V(S_{k,n})$. Points are simulation results, the straight lines are given by $b_n/4s + b_k(1 - s)$ and the curves are given by $b_k(1 - s) + b_n s^2/4$. 20,000 genealogies were simulated for each value of $s$ for each sample configuration.

One problem with estimators $\hat{s}_m$ and $\hat{\theta}_m$ is that they do not guarantee positive values as they ought to do. Table 1 shows the estimated probabilities of having negative values of $\hat{s}_m$ and $\hat{\theta}_a$ for several sample configurations. The chance of having a negative value for $\hat{\theta}_a$ is not high but the same cannot be said for $\hat{s}_m$. Note that in the case of random mating, one may expect that $\hat{s}_m$ takes negative values about half of the time because the expected values of $\bar{S}_{2,2}$ and $\bar{S}_{2,1}$ are both equal to zero. Table 1 shows that $\hat{s}_m$ is more likely to be negative than positive. An examination of its empirical distribution shows that it is skewed to the left although its mean is zero.

**TABLE 1**

**Percentage of samples resulting in negative $\hat{s}_m$ and $\hat{\theta}_a$ with $N = 5000$ and $\theta = 5$**

| $(k, n)$ | $s$ | Negative $\hat{s}_m$ | Negative $\hat{\theta}_a$ |
|---|---|---|---|
| (20, 10) | 0.0 | 54.5 | 3.0 |
| | 0.2 | 33.9 | 1.7 |
| | 0.4 | 16.9 | 0.8 |
| | 0.6 | 5.5 | 0.3 |
| (40, 20) | 0.0 | 52.6 | 1.9 |
| | 0.2 | 25.0 | 1.0 |
| | 0.4 | 7.2 | 0.3 |
| | 0.6 | 1.1 | 0.1 |
| (80, 40) | 0.0 | 52.1 | 1.7 |
| | 0.2 | 15.9 | 0.6 |
| | 0.4 | 2.0 | 0.1 |

Each row is based on 20,000 independently simulated samples.

In fact there is also no guarantee that $\hat{\theta}_m$ is positive because $\overline{S}_{2,2}$ can be smaller than half of $\overline{S}_{2,1}$ although the chance is small. Our simulations show that this does happen and when it happens, $\hat{s}_m$ can be substantially larger than 1. Since $\overline{S}_{2,2}$ is expected to be larger than $\overline{S}_{2,1}$ when there is partial selfing, having inequality $\overline{S}_{2,2} < \overline{S}_{2,1}/2$ suggests that $s = 0$ instead of $s = 1$. Therefore, I propose to use the following estimator of $s$:

$$\hat{s}_f = \begin{cases} 0, & \text{if } \overline{S}_{2,2} < \overline{S}_{2,1} \\ \hat{s}_m, & \text{otherwise.} \end{cases} \tag{23}$$

To remedy $\hat{\theta}_a$, we note that when $\hat{\theta}_a$ is negative, there must be at least one segregating site in the sample, which suggests that $\theta$ must be larger than 0. Therefore, we can use

$$\hat{\theta}_f = \begin{cases} \hat{\theta}_a & \text{if } \hat{\theta}_a > 0 \\ S_{k,n}/a_k & \text{otherwise} \end{cases} \tag{24}$$

as an estimator of $\theta$, that is, estimating $\theta$ by $\hat{\theta}_a$, if $\hat{\theta}_a > 0$, and estimating $\theta$ by WATTERSON's estimate otherwise.

Equation 11 provides yet another estimator of $\theta$ as

$$\hat{\theta}_w = \frac{S_{k,n}}{{}^1/_2 a_n \hat{s}_f + a_k(1 - \hat{s}_f)}. \tag{25}$$

Our simulations show that this estimator is also nearly unbiased. A variation of this estimator is to substitute $\hat{s}_f$ by $\hat{s}_m$ given by (20), and the performance of the resulting estimator is nearly the same as $\hat{\theta}_w$, although $\hat{s}_m$ may be less than 0 or larger than 1.

Figure 4 shows the sampling variances of the three estimators of $\theta$ for a sample of size 20 and 60, respectively. It is clear that MILLIGAN's estimator $\hat{\theta}_m$ has the largest variance among the three, $\hat{\theta}_w$ has the smallest variance; when $s$ is small, $\hat{\theta}_w$ has smaller variance than $\hat{\theta}_f$ but the two converge when $s$ is large, and the speed of convergence increases with $\theta$. Since all the three estimators are nearly unbiased, the one with smallest variance should be preferred. I thus recommend that $\hat{\theta}_w$ be the first choice for estimating $\theta$.

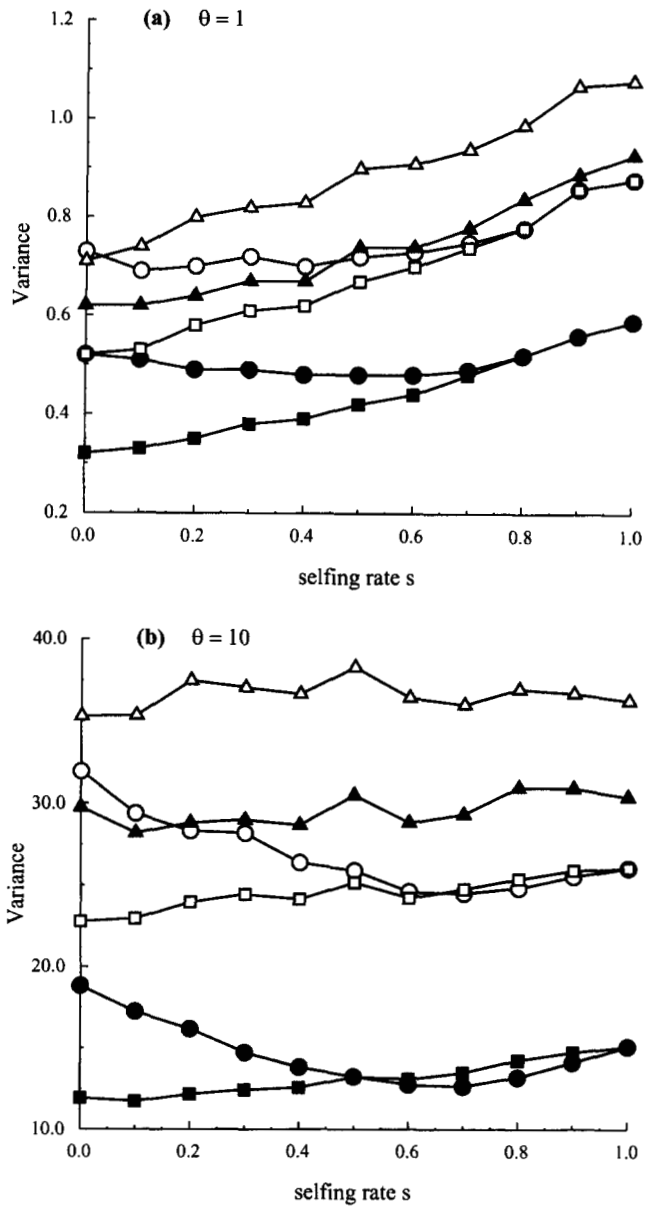


FIGURE 4.—Variances of the three estimators of $\theta$. Curves with unfilled symbols (○, □, and △) correspond to samples of size 20, and curves with filled symbols (●, ■, and ▲) correspond to samples of size 60. Curves with circles, squares and triangles correspond to $\hat{\theta}_w$, $\hat{\theta}_f$ and $\hat{\theta}_m$, respectively. Results for each parameter set are based on 20,000 independent samples from a population with $N = 5000$.

Figure 5 shows that the means and variances of the two estimators of $s$ for several parameter sets. One can see (Figure 5a) that $\hat{s}_m$ is slightly biased downward. It is interesting to see (Figure 5c) that $\hat{s}_f$ is overall less biased than $\hat{s}_m$. The major improvement of $\hat{s}_f$ over $\hat{s}_m$ is the variance, which is significantly smaller than that of $\hat{s}_m$, particularly when $s$ is small. Therefore, $\hat{s}_f$ is recommended as the first choice.

Figure 5 also shows that increasing either sample size or the value of $\theta$ reduces the variance of estimation.

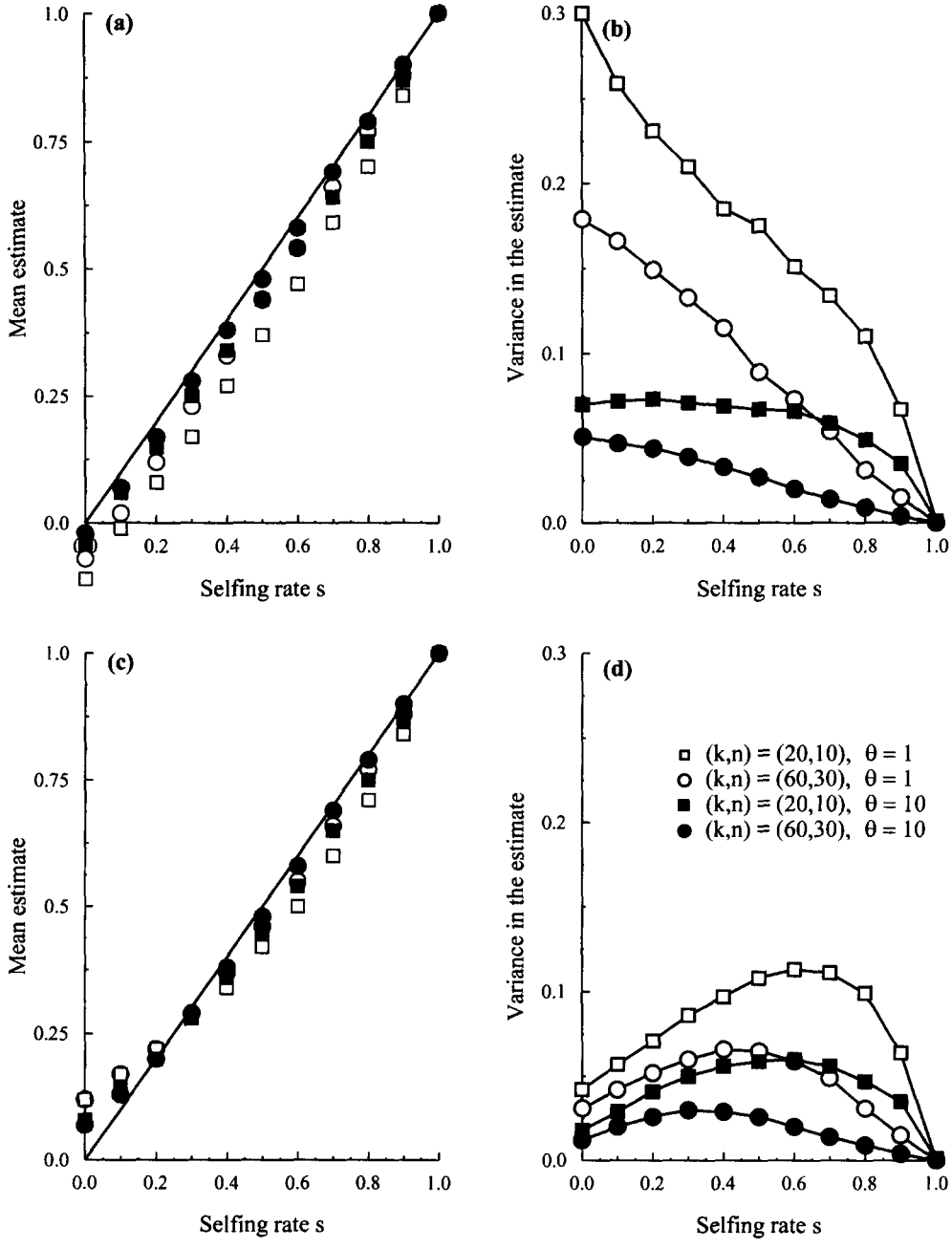


FIGURE 5.—The mean and variance of estimators of $s$. (a and b) Results of $\hat{s}_m$. (c and d) Results of $\hat{s}_f$. All the panels use the sample legends as d. Results for each parameter set are based on 20,000 independent samples from a population with $N = 5000$.

Comparing the variances of $\hat{s}_f$ in the two cases $(k, n) = (60, 30), \theta = 1$ and $(k, n) = (20, 10), \theta = 10$, it is clear that increasing sample size is a more effective way of reducing the sampling variance of $\hat{s}_f$.

It should be pointed out that the variances of $\hat{s}_m$ shown in Figure 5 are substantially smaller than those shown in Figure 5 of MILLIGAN (1996). There are two reasons for this discrepancy. First, MILLIGAN's results were based on independent samples of two sequences, which inflates the difference between sequences from different individuals in a larger sample due to shared common ancestry. Second, MILLIGAN (1996) appeared to use the same number of between and within individual pairs to compute $\overline{S}_{2,2}$ and $\overline{S}_{2,1}$ ($\gamma_b$ and $\gamma_w$ in MILLIGAN's notation) and because there are many more between individual pairs than within individual pairs, doing so necessarily results in loss of information and thus inflates the variance of estimation. MILLIGAN (1996) recognized the limitations of his simulations and predicted that the true variance of $\hat{s}_m$ may be substantially smaller, as is indeed the case shown here.

Finally we would like to examine how sample configurations affect the estimation of $\theta$ and $s$. Given the number ($k$) of sequences in a sample, one may obtain these $k$ sequences from $k/2$ individuals, or $k/2 + 1$, ..., or $k$ individuals. If one obtains the sample from $k/2$ individuals (assuming $k$ is a even number), then both alleles of an individual would have been se-

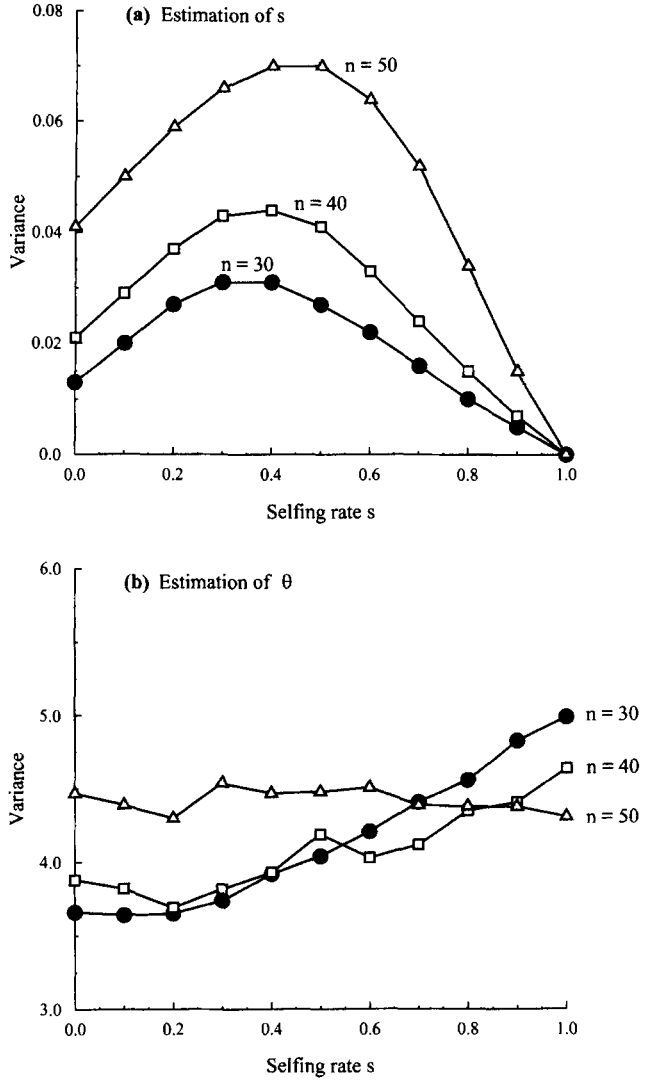


FIGURE 6.—Relationship between the variance of an estimator and sample configuration $(k, n)$ with $k = 60$, $N = 5000$ and $\theta = 5$. (a) The variance of $\hat{s}_f$. (b) The variance of $\hat{\theta}_w$. 10,000 independent samples were generated for each value of $s$ and each value of $n$.

quenced, while if the sample is from $k$ individuals, one individual would have contributed only one sequence. Figure 6 shows how sample configurations affect the variances in the estimation of $\theta$ and $s$. Figure 6a shows that sampling both alleles of each individual is the best scheme for the purpose of estimating $s$. This appears logical because doing so increases the number of within individual comparisons and thus reduces the variance of $\overline{S}_{2,1}$, which is much more variable than $\overline{S}_{2,2}$ unless $s$ is close to 1. Figure 6b shows the variance of $\hat{\theta}_w$ and it is clear that it is better to sample the sequences from fewer individuals when $s$ is small, obviously due to smaller variance in the estimate of $s$. However, it is better to sample more individuals when $s$ is close to 1, because $\hat{s}_f$ has very small variance and increasing number of individuals is equivalent to increasing the sample size. Overall, the differences among the variances of $\hat{\theta}_w$ are not as substantial as those among the variances of $\hat{s}_f$ for different sample configurations.

## DISCUSSION

The coalescent approach is a powerful way to study the evolution of a population. The development of the coalescent theory and its applications have been centered around animal populations in which random mating is often assumed. With the advent of fast and inexpensive techniques for obtaining DNA sequences, DNA samples from plant populations will become more abundant in the future. The coalescent theory for partially selfing populations in this article, as well as in NORDBORG and DONNELLY (1997), is a step forward in providing a powerful modern framework for studying plant populations. The coalescent algorithm for generating samples of DNA sequences enables one to obtain efficiently a large number of simulated samples for an empirical study of a partially selfing population.

We derived in this article the transition probabilities of sample configurations in two successive generations, and our simulation algorithm is based on these transition probabilities. A reviewer suggested that a simpler, but an approximate, algorithm can be derived from the fact that coalescence in selfing individuals is rapid and the number of generations required can be neglected. The algorithm by NORDBORG and DONNELLY (1997) is essentially based on such approach. However, deriving a simulation algorithm from transition probabilities of sample configurations may be necessary (and safer) when a more complex genetic model, for example involving recombination and selection, is studied. Therefore, the framework established in this article should benefit further theoretical investigations on partially selfing populations.

The number of segregating sites in a sample is the simplest quantity observable, yet highly informative. The first fruit of the coalescent theory and the algorithm developed in this article is the equations for the mean and variance of the number of segregating sites. Given the complexity of these two quantities in a number of population genetic models, such as those involving recombination or population subdivision, it is a pleasant surprise that the mean and variance of the number of segregating sites are simple functions of $\theta$ and $s$, and when $s = 0$, they reduce to WATTERSON's (1975) classical results for a random mating population.

We took advantage of the mean and the variance of the number of segregating sites and the coalescent algorithm to develop two new estimators of $\theta$ and a new estimator of $s$. The best estimator of $\theta$ found in this article is the one that is analogous to WATTERSON's (1975) estimator of $\theta$ for a random mating population. Our new estimator of $s$, which is a modification of MILLIGAN's estimator, not only provides an estimate that is meaningful but has substantially smaller variance. As far as estimating the selfing rate is concerned, this study shows that one can be more optimistic than MILLIGAN (1996), who concluded that a relatively large

number of individuals (>100) is needed to achieve reasonable accuracy in the estimation of $s$. We show that the selfing rate $s$ can be estimated reasonably well even with a sample of 20 sequences from 10 individuals. Since it is found in this article that increasing sample size is a more effective way of reducing the variance in the estimate of $s$ than increasing sequence length, I recommend that obtaining a large random sample be assigned a high priority when designing an experiment.

It should be noted that the estimators of $\theta$ and $s$ developed in this article are based on a few summary statistics of a sample and thus do not make full use of available information. Consequently they are unlikely to be the best estimators that can be developed, although they are easy to compute and reasonably accurate. Since significantly better estimators of $\theta$ than WATTERSON's have been found (*e.g.,* FU 1994a,b; KUHNER *et al.* 1995; GRIFFITHS and TAVARÉ 1995) in the case of random mating populations, developing more efficient estimators of $\theta$ and $s$ for a sample from a partially selfing population by making full use of available information in the sample should be worthy of further effort.

I thank Dr. M. UYENOYAMA and a reviewer for their comments and suggestions. I also thank Drs. NORDBORG and DONNELLY for sending me a copy of their manuscript while this manuscript was being reviewed. This research was supported in part by National Institutes of Health grant R29 GM-50428.

## LITERATURE CITED

FELSENSTEIN, J., 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregation sites as compared to phylogenetic estimates. Genet. Res. **56:** 139–147.

FRYXELL, P. A., 1957 Mode of reproduction of higher plants. Bot. Rev. **23:** 135–233.

FU, Y. X., 1994a Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. Genetics **138:** 1375–1386.

FU, Y. X., 1994b A phylogenetic estimator of effective population size or mutation rate. Genetics **136:** 685–692.

FU, Y. X., and W. H. LI, 1993 Maximum likelihood estimation of population parameters. Genetics **134:** 1261–1270.

GRIFFITHS, R. C., and S. TAVARÉ, 1995 Unrooted genealogical tree probabilities in the ininitely-many-sites model. Math. Biosci. **127:** 77–98.

HUDSON, R. R., 1982 Testing the constant-rate neutral allele model with protein sequence data. Evolution **37:** 203–217.

HUDSON, R. R., 1991 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

KINGMAN, J. F. C., 1982a The coalescent. Stochastic Processes and Their Applications. **13:** 235–248.

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

KUHNER, M. K., Y. YAMATO and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

MILLIGAN, B. G., 1996 Estimating long-term mating systems using DNA sequence. Genetics **142:** 619–627.

NORDBORG, M., and P. DONNELLY, 1997 The coalescent process with selfing. Genetics **146:** 1185–1195.

SLATKIN, M., 1991 Inbreeding coefficients and coalescent times. Genet. Res. **58:** 167–175.

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

WATTERSON, G. A., 1975 On the number of segregation sites. Theor. Popul. Biol. **7:** 256–276.

WILLSON, M. F., 1984 Mating patterns in plants, pp. 261–276 in *Perspectives on Plant Population Ecology,* edited by R. DIRZO and J. SARUKHAN. Sinauer Associates, Sunderland, MA.

WRIGHT, S., 1969 *Evolution and the Genetics of Populations, The Theory of Gene Frequencies.* Vol. 2, The University of Chicago Press, Chicago.

Communicating editor: M. K. UYENOYAMA