ON THE DIVERGENCE OF ALLELES IN NESTED SUBSAMPLES FROM FINITE POPULATIONS

RICHARD R. HUDSON AND NORMAN L. KAPLAN

National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709

Manuscript received July 29, 1985
Revised copy accepted April 9, 1986

ABSTRACT

Within-population variation at the DNA level will rarely be studied by sequencing of loci of *randomly* chosen individuals. Instead, individuals will usually be chosen for sequencing based on some knowledge of their genotype. Data collected in this way require new sampling theory. Motivated by these observations, we have examined the sampling properties of a finite population model with two mutation processes and with no selection or recombination. One mutation process generates new alleles according to an infinite-alleles model, and the other generates polymorphisms at sites according to an infinite-sites model. A sample of *n* genes is considered. The stationary distribution of the number of segregating sites in a subsample from one of the allelic classes in the sample conditional on the allelic configuration of the sample is studied. A recursive scheme is developed to compute the moments of this distribution, and it is shown that the distribution is functionally independent of the number of additional alleles in the sample and their respective frequencies in the sample. For the case in which the sample contains only two alleles, the distribution of the number of segregating sites in a subsample containing both alleles conditional on the sample frequencies of the alleles is studied. The results are applied to the analysis of DNA sequences of two alleles found at the *Adh* locus of *Drosophila melanogaster.* No significant departure from the neutral model is detected.

SEVERAL biochemical methods now exist with which genetic diversity can be studied at the molecular level. Three of the more common types of molecular data are electrophoretic data, restriction enzyme data and nucleotide sequence data. Electrophoretic data are the easiest of the three to obtain, and consequently, this method is useful for surveying large samples. However, this type of data provides only information on protein variation, and even these data are of a limited nature. Restriction enzyme data provide more information about DNA variation, but only a small fraction of the DNA is studied. The best data for determining DNA variation are sequence data, but at present, the amount of work required to sequence a region as small as 1 kb is formidable.

Because of the difficult nature of DNA sequencing methods, the sequencing of large random samples of genes from natural populations is not feasible. A practical alternative, which will certainly be frequently employed for studying DNA variation in natural populations, is a stratified sampling approach. For example, a large sample of genes would first be surveyed electrophoretically in order to identify the electrophoretic alleles at a locus. Random subsamples of genes from all or some of the allelic classes would then be studied in more detail, either by restriction mapping or by DNA sequencing. A scheme such as this was recently used by KREITMAN (1983) to study the *alcohol dehydrogenase* (*Adh*) locus in *Drosophila melanogaster*. It is well known that there are two common electrophoretic alleles at the *Adh* locus: fast (*F*) and slow (*S*). KREITMAN sequenced a 3-kb region including the *Adh* locus of five chromosomes bearing the *F*-allele and six chromosomes bearing the *S*-allele.

To study statistical properties of the stratified sampling scheme just described, we suppose that there are two ongoing mutational processes, one giving rise to electrophoretic variation and the other to nucleotide variation that is not electrophoretically detectable. It is assumed that the region of DNA under study is fully linked and that the evolution of the electrophoretic variation can be modeled with a selectively neutral infinite-alleles model. For the electrophoretically undetectable nucleotide variation, a neutral infinite-sites model without recombination is assumed. The question of how to properly connect the infinite-alleles model and the infinite-sites model for analyzing data obtained by stratified sampling is studied in this paper.

The major focus of this investigation is to determine the stationary distribution of the number of segregating sites in a random subsample of genes from one or several allelic classes conditioned on the allelic configuration of the larger sample. It does not seem possible to find a formula for this distribution, but a set of recursion relations can be established with which one can rapidly compute its moments for any choice of parameters. If the subsample of genes is from one allelic class, then the recursive scheme is particularly simple since the conditional distribution is functionally independent of the number of other allelic classes and their respective sizes.

We now outline the rest of the paper. The first part of the THEORY section contains an analysis for samples of two genes. Even though this case is not of practical interest, its simplicity is useful for demonstrating the kind of results we prove and the method of proof. In the second part of the THEORY section, samples of n genes ($n > 2$) are studied. Essential to our analysis is the coalescent process for the sample as defined by KINGMAN (1982a,b; see also TAVARÉ 1984); thus, a short discussion of this process is included for completeness. Some typical calculations are presented in the NUMERICAL RESULTS. An analysis of KREITMAN'S data in light of our results is also discussed. Finally, the DISCUSSION contains some comments on our assumptions and also some possible applications of our results.

THEORY

Samples of two genes: In this section we study a sample of two genes. This case is the simplest to analyze because there is only one topology for the genealogy of the sample back to the most recent common ancestor (Figure 1a).

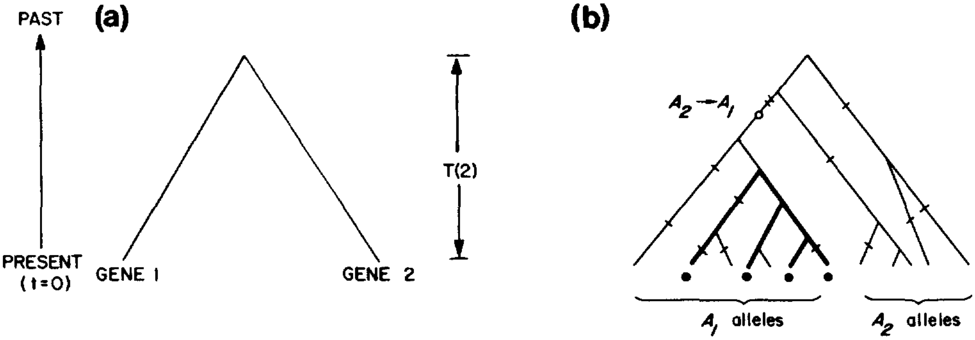

FIGURE 1.—a, The history of a sample of two genes. The time, $T(2)$, is the time back to the most recent common ancestor of the two genes. b, An example of a tree representing the history of a sample of 12 genes. The single type I mutation, $A_2 \rightarrow A_1$, is indicated by a small circle. The allelic configuration is $\{\alpha\} = \{\alpha_1, \cdots, \alpha_{12}\}$ where $\alpha_5 = \alpha_7 = 1$ and all the other $\alpha_i = 0$. Type II mutations are indicated by the small bars on the tree. If the four tips indicated by dots were subsampled for DNA sequencing, then three sites would be segregating in the subsample of four genes. In this case, $n = 12$, $k_1 = 7$, $k_0 = 4$. The total length of the part of the tree drawn in bold lines is $\tilde{T}(k_0)$.

Furthermore, for the neutral Wright-Fisher sampling scheme, the stationary distribution of $T(2)$, the time to the most recent common ancestor of the two genes, is approximately negative exponential with parameter 1, when measured in units of $2N$ generations and where $2N$ is the population size (KINGMAN 1982a,b). The n-coalescent introduced by KINGMAN (1982a,b), which describes the genealogical process for a random sample of n genes, is a more complicated process and is considered in the next section.

We consider two distinct mutation processes. Type I mutations give rise to new electrophoretically detectable alleles distinct from all alleles currently segregating in the population. Type II mutations give rise to electrophoretically silent nucleotide variation, and each type II mutation is assumed to occur at a site not currently segregating in the population. The random numbers of type I and II mutations that occur in each gene in each generation are assumed to be stochastically independent Poisson distributed random variables with means μ_1 and μ_2, respectively. It follows from these assumptions that, in any individual line of descent, the times between successive type I (type II) mutations are independent random variables, and if time is measured in units of $2N$ generations, then their distributions are approximately negative exponential with parameter $\theta_1/2(\theta_2/2)$, where $\theta_1 = 4N\mu_1(\theta_2 = 4N\mu_2)$.

Let S_1 and S_2 denote the number of type I and type II mutations in the lines of descent of the two genes since their most recent common ancestor. It follows from the assumptions of the model that conditional on $T(2) = t$, S_1 and S_2 are independent Poisson variables with means $\theta_1 t$ and $\theta_2 t$, respectively. The unconditional distribution of $S_1(S_2)$, therefore, is a compound Poisson (JOHNSON and KOTZ 1969, chapter 8) that has a compounding distribution that is a negative exponential with parameter $1/\theta_1(1/\theta_2)$, *i.e.*,

$$P(S_1 = j) = \int_0^\infty \frac{e^{-u}u^j}{j!}\,\frac{1}{\theta_1}\,e^{-\left(\frac{u}{\theta_1}\right)}du \qquad j \geq 0 . \tag{1}$$

If $S_1 = 0$, then the two genes in the sample belong to the same allelic class and thus there is one allelic class of size 2. We denote this allelic configuration by $\{0, 1\}$. Alternatively, if $S_1 > 0$, then necessarily there are two allelic classes of size 1, and we denote this allelic configuration by $\{2\}$.

The calculation of the probabilities of the allelic configurations $\{0, 1\}$ and $\{2\}$ follows directly from (1). Indeed,

$$P(\{0, 1\}) = 1 - P(\{2\}) = P(S_1 = 0) = \frac{1}{1 + \theta_1}.$$

The distribution of S_2 conditional on the allelic configuration of the sample being $\{0, 1\}$ is also straightforward to calculate. For any $j \geq 0$,

$$\begin{aligned} P(S_2 = j \mid \{0, 1\}) &= P(S_2 = j \mid S_1 = 0) \\ &= \frac{E(P(S_2 = j \mid T(2))P(S_1 = 0 \mid T(2)))}{P(S_1 = 0)} \\ &= \int_0^\infty \frac{e^{-u}u^j}{j!} \frac{(\theta_1 + 1)}{\theta_2} e^{-\frac{(\theta_1 + 1)u}{\theta_2}} du. \end{aligned}$$

The distribution of S_2 conditional on the allelic configuration being $\{0, 1\}$ is therefore a compound Poisson for which the compounding distribution is negative exponential with parameter $(1 + \theta_1)/\theta_2$. In exactly the same way, one can show that the distribution of S_2 conditional on the allelic configuration being $\{2\}$ is also a compound Poisson for which the compounding distribution has the density

$$\frac{(1 + \theta_1)}{\theta_1\theta_2}(1 - e^{-\frac{\theta_1}{\theta_2}u})e^{-\frac{u}{\theta_2}}.$$

The fact that in both cases the conditional distribution of S_2 is compound Poisson with only the compounding distribution changing is a direct consequence of the assumption that the two mutational processes are stochastically independent. Conditioning on the allelic configuration of the sample (an event associated with the type I mutation process) only gives information about the type II mutation process by providing information about $T(2)$. Thus, in both cases, the compounding distribution is just the distribution of $\theta_2 T(2)$ conditional on the configuration of the sample.

Samples of n genes: We now consider the general problem for a sample of n genes. To begin, we define the n-coalescent that describes the genealogical process of the sample. For a formal discussion of this process, one should consult the papers by KINGMAN (1982a,b) or the recent review paper by TAVARÉ (1984). It is appropriate for our purposes to think of a realization of the n-coalescent as a binary tree having a node at the top and n tips at the bottom. Each of the n tips is identified with one of the genes in the sample; thus, as we move up the tree, tracing the lineages of the ancestors of the sample, time is measured from the present into the past. There are $n - 1$ nodes in the tree that we label from 1 to $n - 1$, going from the most recent node to the most ancient node. A node is interpreted as a time point in the history of the sample when the most recent common ancestor of two or more genes in the sample occurred. Let $T(j)$ denote the time between the $(n - j)$th and $(n - j + 1)$th node $(2 \leq j \leq n)$, measured in units of $2N$ generations. (For convenience, we define any of the tips to be the *0*th node.) KINGMAN (1982a,b) has shown that, for Wright-Fisher sampling, the $\{T(j)\}$ are independent random variables, and for large N, the stationary distribution of $T(j)$ when measured in units of $2N$ generations is approximately negative exponential with parameter $\binom{j}{2}$. To complete the description of the n-coalescent, it is necessary to specify which pairs of branches of the tree coalesce at the different nodes. Since the sampling is neutral, all pairs of branches are equally likely to coalesce; therefore, at the jth node any pair of branches has probability of $1\bigg/\binom{n-j+1}{2}$ of coalescing.

It follows from the assumptions of the model that, conditional on the lengths of the branches of the tree, the random numbers of type I and II mutations on different branches are stochastically independent. Furthermore, the conditional distributions of the numbers of type I and II mutations on a branch of length t are Poisson with means $\theta_1 t/2$ and $\theta_2 t/2$, respectively. KINGMAN (1982a) has recently shown that, if genes in the sample are grouped so that any two in a group are descended from a single gene, and no type I mutations have occurred in their line of descent, then the stationary distribution of the number of groups and their respective sizes is given by EWENS' (1972) well-known sampling distribution. (This result requires the standard assumption that $N \rightarrow \infty$, $\mu_1 \rightarrow 0$ and $4N\mu_1$ converges to θ_1.) Let $\{\alpha\} = \{\alpha_1, \cdots, \alpha_n\}$ represent the allelic configuration of a sample of size n, where α_i denotes the number of allelic classes of size i. EWENS (1972) showed that, at stationarity,

$$P(\{\alpha\}) = \frac{n!\theta_1^k}{1^{\alpha_1}2^{\alpha_2} \cdots n^{\alpha_n}\alpha_1! \cdots \alpha_n!S_n(\theta_1)}$$

where

$$k = \sum_{i=1}^{n} \alpha_i \text{ is the number of alleles in the sample,}$$

$$n = \sum_{i\geq 1} i\alpha_i \text{ is the size of the sample,}$$

and

$$S_n(\theta_1) = \theta_1(\theta_1 + 1) \cdots (\theta_1 + n - 1).$$

Suppose that the allelic configuration of the sample is $\{\alpha\}$ and that a random subsample of k_0 genes is taken from one of the allelic classes which we assume without loss of generality is of size $k_1(k_0 \leq k_1)$. Our goal is to study the distribution of the number of type II mutations in the random subsample of size k_0, conditional on the allelic configuration of the sample being $\{\alpha\}$ and that the random subsample is chosen from an allelic class of size k_1. For notational purposes we let $S_2(k_0, k_1, \alpha)$ denote a random variable that has this conditional distribution.

Let $\tilde{T}(k_0)$ denote the total time in the history of the subsample of k_0 genes. This random time can be determined from the tree by tracing the lineages of the genes in the subsample and finding the first node where all of these lineages coalesce (see Figure 1b). In view of our remarks at the end of the previous section, the distribution of $S_2(k_0, k_1, \alpha)$ is a compound Poisson for which the compounding distribution is the distribution of $\theta_2\tilde{T}(k_0)$, conditional on the allelic configuration of the sample being $\{\alpha\}$ and that the random subsample is chosen from an allelic class of size k_1. For notational purposes we let $T(k_0, k_1, \alpha)$ denote a random variable that has the conditional distribution of $\tilde{T}(k_0)$.

Let T^* denote the time of the most recent event in the tree regardless of whether it is a most recent common ancestor event (*i.e.*, a node) or a type I mutation. It follows from the construction of the tree that T^* is the minimum of $T(n)$ and the time to the most recent type I mutation, which, because the mutation rate is assumed to be constant, has a negative exponential distribution with parameter $n\theta_1/2$. Thus, T^* itself has a negative exponential distribution with parameter $\binom{n}{2} + n\theta_1/2$. Furthermore, if B denotes the set of trees where the most recent event is a most recent common ancestor, then

$$P(B) = \frac{\frac{n}{2}}{\binom{n}{2} + \frac{n\theta_1}{2}} = \frac{n-1}{n-1+\theta_1}.$$

The most recent event in the tree is important because one can use it to decompose the set of trees for which the allelic configuration is $\{\alpha\}$, *i.e.*, those trees for which type I mutations result in the allelic configuration $\{\alpha\}$. For any α, define

$$A(\alpha) = \{\text{set of trees for which the allelic configuration is } \{\alpha\}\}.$$

Then, clearly,

$$\chi[A(\alpha)] = \chi[A(\alpha)]\chi[B] + \chi[A(\alpha)]\chi[\bar{B}].$$

For any set of trees C, $\bar{C}$ is its complement, and $\chi[C]$ is its indicator function, *i.e.*, $\chi[C]$ is a function defined on the set of all trees such that, if a tree belongs to C, then $\chi[C] = 1$, and if a tree does not belong to C, then $\chi[C] = 0$. On the set B, $\{\alpha^*\}$, the allelic configuration of the ancestors of the sample just before T^* can equal $\{\alpha_2\}$, $\{\alpha_3\}$, $\cdots$, or $\{\alpha_n\}$, where

$$\{\alpha_i\} = \{\alpha_1, \cdots, \alpha_{i-1} + 1, \alpha_i - 1, \cdots, \alpha_n\}, \qquad 2 \le i \le n.$$

Thus,

$$\chi[A(\alpha)] = \sum_{i=2}^{n} \chi[B]\chi[A(\alpha)]\chi[B_i] + \chi[\bar{B}]\chi[A(\alpha)], \tag{2}$$

where B_i denotes the set of trees where $\alpha^* = \alpha_i$, $2 \le i \le n$. Equation (2) is the basic result from which the recursions for the distribution of $T(k_0, k_1, \alpha)$ are derived. Before we do this, however, it is necessary to compute the expectations of each term on the right in (2). It follows from the strong Markov property of the n-coalescent and the stationarity of the genealogical process that $\chi[B]$ and $\chi[B_i]$ are stochastically independent and that $E(\chi[B_i]) = P(\{\alpha_i\})$. Thus,

$$E(\chi[B]\chi[B_i]) = P(B)P(\{\alpha_i\}), \qquad 2 \le i \le n.$$

Also, if the most recent event in the tree is a most recent common ancestor and $\alpha^* = \alpha_i$, then for the allelic configuration of the sample to be $\{\alpha\}$, it is necessary that the most recent common ancestor of the two genes that coalesce at T^* is one of the $(i-1)(\alpha_{i-1}+1)$ ancestral genes that belong to an allelic class of size $i - 1$. Therefore,

$$P(\{\alpha\} \mid \chi[B]\chi[B_i] = 1) = \frac{(i-1)(\alpha_{i-1}+1)}{n-1},$$

and so

$$E(\chi[B]\chi[A(\alpha)]\chi[B_i]) = P(B)P(\{\alpha_i\}) \frac{(i-1)(\alpha_{i-1}+1)}{n-1}. \tag{3}$$

It is a straightforward matter to check that, if the most recent event on the tree is a mutation, then the allelic configuration of the other $n - 1$ ancestors immediately preceding T^* must be $\{\alpha_1\}$, where

$$\{\alpha_1\} = \{\alpha_1 - 1, \alpha_2, \cdots, \alpha_n\}.$$

Thus,

$$E(\chi[\bar{B}]\chi[A(\alpha)]) = P(\bar{B})P(\{\alpha_1\}). \tag{4}$$

This last result is also a direct consequence of the work of GRIFFITHS (1980) concerning the lines of descent process (see also TAVARÉ 1984). It follows from (2), (3) and (4) that for any α,

$$P(\{\alpha\}) = \frac{n-1}{n-1+\theta_1} \sum_{i=2}^{n} P(\{\alpha_i\}) \frac{(i-1)(\alpha_{i-1}+1)}{n-1} \chi[\alpha_i \ge 1] + \frac{\theta_1}{n-1+\theta_1} P(\{\alpha_1\})\chi[\alpha_1 \ge 1], \tag{5}$$

where $\chi[\alpha_i \ge 1] = 1$ if $\alpha_i \ge 1$ and zero otherwise. One should note that all the $P(\{\alpha_i\})$ on the right of (5) are for samples of $n - 1$ genes; therefore (5) can be used to derive EWENS' distribution inductively.

We now turn to the distribution of $T(k_0, k_1, \alpha)$. Suppose $i \ne k_1$. It follows from the strong Markov property of the genealogical process that, on the set of trees where $\chi[B]\chi[A(\alpha)]\chi[B_i] = 1$,

$$T(k_0, k_1, \alpha) = k_0 T^* + T'(k_0, k_1, \alpha_i) \tag{6}$$

where $T'(k_0, k_1, \alpha_i)$ is stochastically independent of T^* and has the same distribution as $T(k_0, k_1, \alpha_i)$. Similarly, where $\chi[\bar{B}]\chi[A(\alpha)] = 1$,

$$T(k_0, k_1, \alpha) = k_0 T^* + T'(k_0, k_1, \alpha_1) \tag{7}$$

where $T'(k_0, k_1, \alpha_1)$ is stochastically independent of T^* and has the same distributions as $T(k_0, k_1, \alpha_1)$. If $i = k_1$, then it is possible that the two genes associated with the branches that coalesce belong to the allelic class that is subsampled, and also to the subsample itself. If C denotes the former event and D the latter, then

$$P(C \mid \chi[B]\chi[A(\alpha)]\chi[B_{k_1}] = 1) = \frac{1}{\alpha_{k_1}}$$

and

$$P(D \mid \chi[C]\chi[B]\chi[A(\alpha)]\chi[B_{k_1}] = 1) = \frac{\binom{k_0}{2}}{\binom{k_1}{2}}.$$

The first probability states that any of the allelic classes of size k_1 are equally likely to be subsampled, and the second probability is the chance that the two branches that coalesce are among the lineages of k_0 genes that were subsampled from an allelic class of size k_1. It follows once more from the strong Markov property that when $\chi[B]\chi[A(\alpha)]\chi[B_{k_1}] = 1$,

$$\begin{aligned} T(k_0, k_1, \alpha) = k_0 T^* &+ T'(k_0, k_1, \alpha_{k_1})\chi[\bar{C}] \\ &+ T'(k_0 - 1, k_1 - 1, \alpha_{k_1})\chi[C]\chi[D] + T'(k_0, k_1 - 1, \alpha_{k_1})\chi[C]\chi[\bar{D}] \end{aligned} \tag{8}$$

where $T'(k_0, k_1, \alpha_{k_1})$, $T'(k_0 - 1, k_1 - 1, \alpha_{k_1})$ and $T'(k_0, k_1 - 1, \alpha_{k_1})$ are stochastically independent of T^*, $\chi[C]$ and $\chi[D]$ and have the same distribution as $T(k_0, k_1, \alpha_{k_1})$, $T(k_0 - 1, k_1 - 1, \alpha_{k_1})$ and $T(k_0, k_1 - 1, \alpha_{k_1})$, respectively.
Let

$$H(k_0, k_1, \alpha) = E(e^{-sT(k_0, k_1, \alpha)})$$

denote the moment generating function of $T(k_0, k_1, \alpha)$. It follows from (2), (6), (7) and (8) that

$$H(k_0, k_1, \alpha) = E(e^{-sk_0 T^*})[P(B)(Q_1 + Q_2) + P(\bar{B})Q_3] \tag{9}$$

where

$$Q_1 = \sum_{i \neq k_1} \frac{(i - 1)(\alpha_{i-1} + 1)}{n - 1} \frac{P(\{\alpha_i\})}{P(\{\alpha\})} H(k_0, k_1, \alpha_i),$$

$$\begin{aligned} Q_2 = \frac{(k_1 - 1)(\alpha_{k_1 - 1} + 1)}{n - 1} \frac{P\{\alpha_{k_1}\}}{P\{\alpha\}} \Bigg[& \frac{\alpha_{k_1} - 1}{\alpha_{k_1}} H(k_0, k_1, \alpha_{k_1}) \\ & + \frac{1}{\alpha_{k_1}} \left(\frac{\binom{k_0}{2}}{\binom{k_1}{2}} H(k_0 - 1, k_1 - 1, \alpha_{k_1}) + \left(1 - \frac{\binom{k_0}{2}}{\binom{k_1}{2}} \quad H(k_0, k_1 - 1, \alpha_{k_1}) \right) \right) \Bigg] \end{aligned}$$

and

$$Q_3 = \frac{P(\{\alpha_1\})}{P(\{\alpha\})} H(k_0, k_1, \alpha_1).$$

The right-hand side of (9) can be simplified since

$$\frac{P(\bar{B})P(\{\alpha_1\})}{P(\{\alpha\})} = \frac{\alpha_1}{n},$$

and

$$P(B) \frac{(i-1)\,(\alpha_{i-1}+1)P(\{\alpha_i\})}{n-1 \quad P(\{\alpha\})} = \frac{i\alpha_i}{n}.$$

Thus, (9) can be written as

$$\begin{aligned} H(k_0, k_1, \alpha) = E(e^{-sk_0T^*}) \Bigg[\sum_{i \neq k_1} \frac{i\alpha_i}{n} H(k_0, k_1, \alpha_i) + \frac{k_1(\alpha_{k_1}-1)}{n} H(k_0, k_1, \alpha_{k_1}) \\ + \frac{k_1}{n} \frac{\binom{k_0}{2}}{\binom{k_1}{2}} H(k_0-1, k_1-1, \alpha_{k_1}) + \frac{k_1}{n} \left(1 - \frac{\binom{k_0}{2}}{\binom{k_1}{2}} \right) H(k_0, k_1-1, \alpha_{k_1}) \Bigg]. \end{aligned} \tag{10}$$

The recursion for $H(k_0, k_1, \alpha)$ in (10) can be simplified even further. It is straightforward to show that, for a sample of four genes,

$$H(2, 2, \{2,1\}) = H(2, 2, \{0, 2\}).$$

This demonstrates that, in this special case, H depends only on k_0, k_1 and n and not on the allelic configuration of the other $n - k_1$ genes in the sample. We now show by induction that this property holds in general. Suppose that, for samples of size less than or equal to $n - 1$, $H(k_0, k_1, \alpha)$ only depends on k_0, k_1 and the sample size, *i.e.*, it does not matter how many other allelic classes there are in the sample and what their respective sizes are. In particular, it follows that the $H(k_0, k_1, \alpha_i)$ are all equal to, say, $H_{n-1}(k_0, k_1)$. Thus, equation (10) becomes

$$\begin{aligned} H(k_0, k_1, \alpha) = E(e^{-sk_0T^*}) \Bigg[\left(1 - \frac{k_1}{n}\right) H_{n-1}(k_0, k_1) \\ + \frac{k_1}{n} \left(\frac{\binom{k_0}{2}}{\binom{k_1}{2}} H_{n-1}(k_0-1, k_1-1) + \left(1 - \frac{\binom{k_0}{2}}{\binom{k_1}{2}} \right) H_{n-1}(k_0, k_1-1) \right) \Bigg]. \end{aligned} \tag{11}$$

Equation (11) shows that, for samples of size n, $H(k_0, k_1, \alpha)$ also depends only on k_0, k_1 and n.

Differentiating both sides of (11) leads to recursions for the moments of $T(k_0, k_1, \alpha)$. The moments of $S_2(k_0, k_1, \alpha)$ follow directly from properties of the compound Poisson distribution, *i.e.*,

$$E(S_2(k_0, k_1, \alpha)) = \frac{\theta_2}{2} E(T(k_0, k_1, \alpha)) \tag{12}$$

and

$$\operatorname{Var}(S_2(k_0, k_1, \alpha)) = \frac{\theta_2}{2} E(T(k_0, k_1, \alpha)) + \frac{\theta_2^2}{4} \operatorname{Var}(T(k_0, k_1, \alpha)). \tag{13}$$

Let

$$M(k_0, k_1, n) = E(T(k_0, k_1, \alpha))$$

and

$$L(k_0, k_1, n) = E(T^2(k_0, k_1, \alpha))$$

denote the first and second moments of $T(k_0, k_1, \alpha)$. It follows from (11) that

$$\begin{aligned} M(k_0, k_1, n) = {} & \frac{2k_0}{n(n-1+\theta_1)} + \left(1 - \frac{k_1}{n}\right) M(k_0, k_1, n-1) \\ & + \frac{k_1}{n}\left[\frac{\binom{k_0}{2}}{\binom{k_1}{2}} M(k_0 - 1, k_1 - 1, n - 1)\right. \\ & \left. + \left(1 - \frac{\binom{k_0}{2}}{\binom{k_1}{2}}\right) M(k_0, k_1 - 1, n - 1)\right] \end{aligned} \tag{14}$$

and

$$\begin{aligned} L(k_0, k_1, n) = {} & \frac{4k_0}{n(n-1+\theta_1)} M(k_0, k_1, n) + \left(1 - \frac{k_1}{n}\right) L(k_0, k_1, n-1) \\ & + \frac{k_1}{n}\left[\frac{\binom{k_0}{2}}{\binom{k_1}{2}} L(k_0 - 1, k_1 - 1, n - 1)\right. \\ & \left. + \left(1 - \frac{\binom{k_0}{2}}{\binom{k_1}{2}}\right) L(k_0, k_1 - 1, n - 1)\right]. \end{aligned} \tag{15}$$

Equations (14) and (15) can be solved recursively using the initial conditions

$M(1, k_1, n) = 0$ and $L(1, k_1, n) = 0$, $n \geq 2$, $k_1 \geq 1$. Several sample calculations are presented in the next section.

If θ_1 is set equal to zero in (14), then it is straightforward to check that the solution of the resulting system of equations is

$$M(k_0, k_1, n) = \frac{2k_1}{n} \sum_{j=1}^{k_0-1} \frac{1}{j}.$$

Thus, if θ_1 is small,

$$M(k_0, k_1, n) \approx \frac{2k_1}{n} \sum_{j=1}^{k_0-1} \frac{1}{j}. \tag{16}$$

There does not appear to be any such simple approximation for $L(k_0, k_1, n)$.

The approximation in (16) can be interpreted in the following way. If k_0 genes are sampled at random, then the expected value of the total time in the history of the sample is

$$2 \sum_{j=1}^{k_0-1} \frac{1}{j}$$

(WATTERSON 1975). Thus, when θ_1 is small, $M(k_0, k_1, n)$ is approximately equal to the expected value of the total time in the history of a random sample of k_0 genes from a population of size $2Nk_1/n$.

Until now, we have studied the behavior of a subsample from a particular allelic class. The methods developed for dealing with this problem can also be used when one subsamples from different allelic classes. In order to demonstrate the ideas, we consider the following case. Suppose in the sample, there are exactly two allelic classes of size k_1 and l_1 and that subsamples of size k_0 and l_0 are randomly picked from each allelic class. If there are just two allelic classes in the sample of size k_1 and $l_1 (n = k_1 + l_1)$, then it is easier to denote the allelic configuration by $\{k_1, l_1\}$.

Let $T(k_0, k_1, l_0, l_1)$ denote a random variable for which the distribution is the distribution of the total time in the history of the two subsamples conditional on the allelic configuration of the sample being $\{k_1, l_1\}$ and that subsamples of size k_0 and l_0 are randomly chosen from each allelic class. We now develop a system of recursions for $K(k_0, k_1, l_0, l_1)$, the moment generating function of $T(k_0, k_1, l_0, l_1)$. There are three cases to consider.

Case 1: $k_1 > 1$, $l_1 > 1$. The method used to derive (9) is directly applicable. To simplify notation, we shall only indicate which arguments of K change, *e.g.*, $K(k_1 - 1) \equiv K(k_0, k_1 - 1, l_0, l_1)$. Therefore, we have

$$
\begin{aligned}
K(k_0, k_1, l_0, l_1) = E(e^{-s(k_0+l_0)T^*}) \left(\frac{k_1}{k_1 + l_1}\right) \Bigg[& \chi[k_0 = 1]K(k_1 - 1) \\
& + \chi[k_0 > 1]\left(\frac{\binom{k_0}{2}}{\binom{k_1}{2}} K(k_0 - 1, k_1 - 1) + \left(1 - \frac{\binom{k_0}{2}}{\binom{k_1}{2}} \right) K(k_1 - 1) \right) \Bigg] \\
+ E(e^{-s(k_0+l_0)T^*}) \left(\frac{l_1}{k_1 + l_1}\right) \Bigg[& \chi[l_0 = 1]K(l_1 - 1) \\
& + \chi[l_0 > 1]\left(\frac{\binom{l_0}{2}}{\binom{l_1}{2}} K(l_0 - 1, l_1 - 1) + \left(1 - \frac{\binom{l_0}{2}}{\binom{l_1}{2}} \right) K(l_1 - 1) \right) \Bigg].
\end{aligned}
\tag{17}
$$

Case 2: $k_1 = 1$, $l_1 > 1$ or $k_1 > 1$ and $l_1 = 1$. For definiteness we assume that $k_1 = 1$ and $l_1 > 1$. The analogue of (2) is

$$
\begin{aligned}
\chi[A(1, l_1)] = {} & \chi[B]\chi[A(1, l_1)]\chi[\alpha^* = (1, l_1 - 1)] \\
& + \chi[\bar{B}]\chi[A(1, l_1)]\chi[\alpha^* = (l_1 + 1)] + \chi[\bar{B}]\chi[A(1, l_1]\chi[\alpha^* = (1, l_1)].
\end{aligned}
\tag{18}
$$

It follows from (18) and arguments similar to those employed earlier that

$$
\begin{aligned}
K(1, 1, l_0, l_1) = E(e^{-s(1+l_0)T^*}) \Bigg[& \frac{1}{1 + l_1} [\chi[l_0 = 1]K(l_1 - 1) \\
& + \chi[l_0 > \underset{)}{1]}\left(\frac{\binom{l_0}{2}}{\binom{l_1}{2}} K(l_0 - 1, l_1 - 1) + \quad 1 - \frac{\binom{l_0}{2}}{\binom{l_1}{2}} \right) K(l_0, l_1 - 1) \Bigg) \\
& + \frac{1}{\theta_1 + l_1} \frac{1}{l_1 + 1} H(l_0 + 1, l_1 + 1) + \frac{\theta_1}{\theta_1 + l_1} \frac{1}{l_1 + 1} K(1, 1, l_0, l_1) \Bigg].
\end{aligned}
\tag{19}
$$

An interesting feature of (19) is that $H(l_0 + 1, l_1 + 1)$ must be evaluated. This term arises when one considers the behavior of $T(1, 1, l_0, l_1)$ on the set where $\chi[\bar{B}]\chi[A(1, l_1)]\chi[\alpha^* = (l_1 + 1)] = 1$. Another point to note is that the coefficients on the right side of (19) involve θ_1.

Case 3: $k_0 = k_1 = l_0 = l_1 = 1$. This case was dealt with earlier in this section when we considered a sample of two genes, and so we are done.

Typical calculations of the mean and variance of $T(k_0, k_1, l_0, l_1)$ are given in the next section.

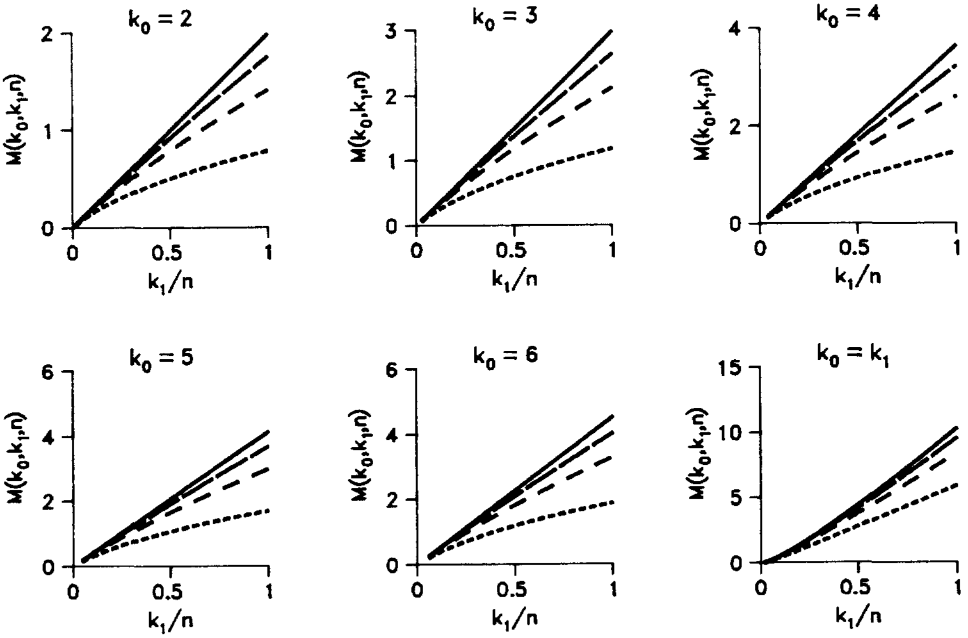

FIGURE 2.—$M(k_0, k_1, n)$, the expectation of $T(k_0, k_1, \alpha)$, plotted as a function of k_1/n for several values of k_0 and θ_1. The solid curves (———) are for $\theta_1 = 0.02$, the long-dash (— — —) curves for $\theta_1 = 0.3$, the short-dash (- - - -) curves for $\theta_1 = 1$, and the dotted curves (····) for $\theta_1 = 5$. In every case, $n = 100$.

One interesting consequence of the results of this section is that the conditional mean of the number of segregating sites in the subsample which distinguish the two allelic classes can also be computed. Segregating sites which distinguish the two allelic classes are those which are not segregating within the allelic classes but between them. It is not difficult to check that the conditional expectation of the size of the region of the tree in which such mutations must occur equals $E(T(k_0, k_1, l_0, l_1)) - E(T(k_0, k_1, \{k_1, l_1\})) - E(T(l_0, l_1, \{k_1, l_1\}))$. Thus, the conditional mean of the number distinguishing mutations can be computed. Its variance, however, cannot be computed directly because we have not considered the joint distributions of $T(k_0, k_1, l_0, l_1)$, $T(k_0, k_1, \{k_1, l_1\})$ and $T(l_0, l_1, \{k_1, l_1\})$. It is possible to develop a system of recursions for the moments of the number of distinguishing mutations by using arguments similar to those presented here, but we shall not pursue this issue.

NUMERICAL RESULTS AND AN APPLICATION

In Figure 2, values of $M(k_0, k_1, n)$, the expectation of $T(k_0, k_1, \alpha)$, obtained by solving the recursion (14), are plotted as a function of k_1/n, for $n = 100$ and several values of k_0 and θ_1. It can be seen from this figure that, for θ_1 less than 0.3, the simple approximation given by (16) is quite good. The ratio of the standard deviation to the mean of $T(k_0, k_1, \alpha)$ is plotted as a function of k_1/n in Figure 3. We note that these ratios are large, typically greater than 0.5. Larger values of k_0 result in lower values of this ratio. In Figure 4 the

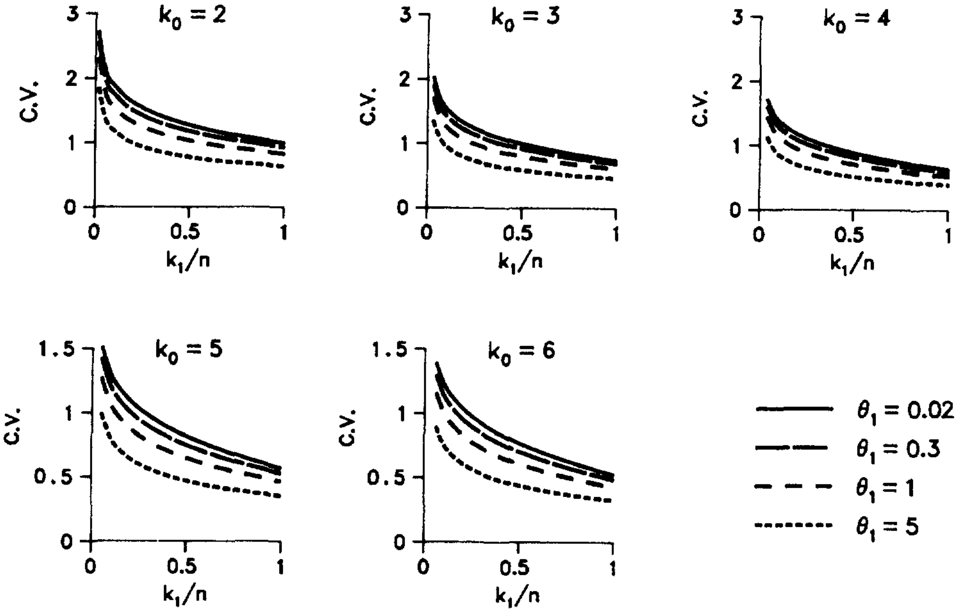

FIGURE 3.—The coefficient of variation of $T(k_0, k_1, \alpha)$ plotted as a function of k_1/n for several values of k_0 and θ_1. In every case, $n = 100$.

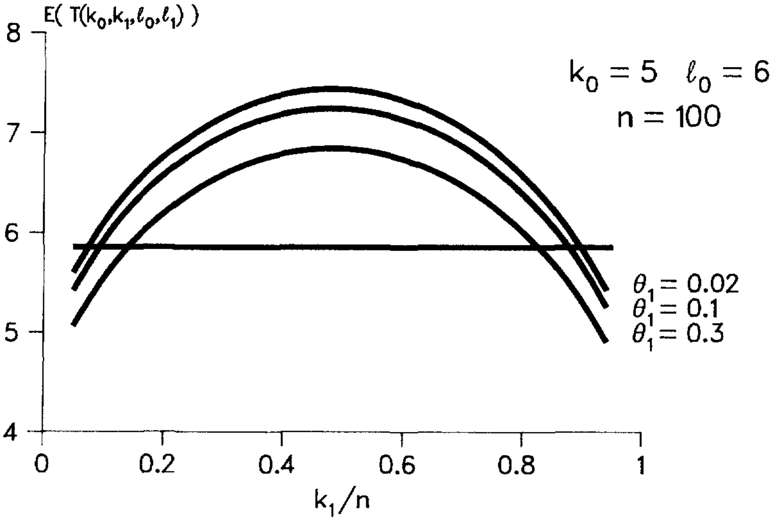

FIGURE 4.—$E(T(k_0, k_1, l_0, l_1))$ plotted as a function of k_1/n for several values of θ_1. The expected time in the history of a completely random sample of 11 genes is 5.86. The horizontal line shows this value for comparison with the conditional expected times.

expectation of $T(k_0, k_1, l_0, l_1)$ is plotted as a function of k_1/n, for $k_0 = 5$, $l_0 = 6$, and several values of θ_1. The expected time in the history of a completely random sample of size 11 is also shown for comparison. The conditional expected time has a maximum at intermediate values of k_1/n. For $\theta_1 = 0.1$ and k_1/n near 0.5, the conditional expected time is about 20% greater than the unconditional expected time. For $k_0 = 3$, and $l_0 = 8$, the curves are similar,

TABLE 1

Mean and standard variation of the time in the history of a subsample of genes

				$E(T(k_0, k_1, \alpha))$		$E(T(l_0, l_1, \alpha))$		$E(T(k_0, k_1, l_0, l_1))$	
k_0	l_0	k_1/n	θ_1	$n = 50$	$n = 200$	$n = 50$	$n = 200$	$n = 50$	$n = 200$
5	6	0.1	0.1	0.41 (0.54)	0.41 (0.51)	3.93 (2.18)	3.94 (2.18)	5.90	5.93
		0.2	0.02	0.83 (0.93	0.83 (0.90)	3.62 (2.23)	3.62 (2.22)	6.74	6.74
			0.1	0.82 (0.89)	0.82 (0.86)	3.50 (2.10)	3.51 (2.09)	6.56	6.59
			0.3	0.79 (0.81)	0.79 (0.79)	3.26 (1.85)	3.27 (1.84)	6.17	6.21
		0.3	0.1	1.22 (1.17)	1.22 (1.15)	3.08 (1.99)	3.08 (1.98)	6.96	6.99

The standard deviations of the times are in parentheses under the mean times.

except that the left end of the curves are somewhat higher and the right ends are lower.

The curves shown in Figures 2, 3 and 4 were generated with n equal to 100. Similar curves were generated with n equal to 30, and the resulting curves were almost indistinguishable from those shown in Figures 2, 3 and 4. This suggests that, given k_1/n, the mean and variance of $T(k_0, k_1, \alpha)$ and the mean of $T(k_0, k_1, l_0, l_1)$ are nearly independent of n. This point is further illustrated in Table 1, where values of $M(k_0, k_1, n)$, the variance of $T(k_0, k_1, \alpha)$, and the expectation of $T(k_0, k_1, l_0, l_1)$ are given for several values of θ_1, k_1/n, and for $n = 50$ and $n = 200$. The values of these quantities are nearly the same with $n = 50$ as with $n = 200$.

To illustrate the use of our results, we consider the set of DNA sequences of the *Adh* locus of *D. melanogaster* obtained by KREITMAN (1983). The *Adh* locus of *D. melanogaster* has two common alleles, *F* and *S*. KREITMAN chose to sequence five chromosomes bearing the *F*-allele, and six chromosomes bearing the *S*-allele. The chromosomes were obtained from a variety of locations worldwide. The frequency of the *F*-allele in natural populations has been found to vary with latitude (OAKESHOTT *et al.* 1981). This clinal pattern may be due to population subdivision and drift, founder effects or selection.

To calculate the expected numbers of segregating sites within each allelic class and distinguishing the two classes, we must first estimate θ_2. Equation (12) suggests the following estimator of θ_2:

$$\hat{\theta}_2 = 2S/E(T(k_0, k_1, l_0, l_1)),$$

where S equals the observed number of segregating sites in the sample. In order to calculate the expectation of $T(k_0, k_1, l_0, l_1)$, it is necessary to specify values for θ_1, k_1 and n.

Table 1 indicates that the value of $E(T(k_0, k_1, l_0, l_1))$ is not very sensitive to the value of θ_1 providing θ_1 is small, or to the value of n providing n is

TABLE 2

Comparison of observed and expected numbers of segregating sites in the samples of *Adh* sequences

			Segregating sites		
	k_1/n	$\hat{\theta}_2$	Within *F*-allelic class	Within *S*-allelic class	Distinguishing
Expected[a]	0.1	14.2	2.9 (4.0)	28.0 (16.4)	11.1
	0.2	12.7	5.2 (5.9)	22.3 (14.1)	14.5
	0.3	12.0	7.3 (7.4)	18.5 (12.6)	16.2
Observed[b]			15	30	2
Observed (adjusted)[b]			10	26	6

[a] The expected numbers of segregating sites in each category are calculated as described in the NUMERICAL RESULTS with $n = 200$. The standard deviation of the number of segregating sites is in parentheses.

[b] The observed numbers are for the data of KREITMAN (1983). The observed (adjusted) numbers were obtained from the same data as described in the NUMERICAL RESULTS.

reasonably large. Thus, we assume that $\theta_1 = 0.1$, which is an average estimate of this mutational parameter from electrophoretic data for enzyme loci such as *Adh* (LEWONTIN 1974), and we assume that $n = 200$. Estimates of θ_2 were calculated for three representative values of k_1/n: 0.1, 0.2 and 0.3. These estimates of θ_2, as well as the expected numbers of sites segregating among the *F*-allelic class, among the *S*-allelic class and the expected number of distinguishing sites between the two classes are given in Table 2. The observed numbers of segregating sites in each of these categories are also given in Table 2. It is clear from the table that these values do not agree very well with the corresponding predictions.

Since five of the polymorphic sites segregate both within the *F*-alleles and within the *S*-alleles, at least one of our assumptions is violated. Under the infinite-site model without recombination, sites cannot segregate within two different allelic classes. Either back mutations, recurrent mutations or recombination events have occurred. KREITMAN (1983) described two inferred recombination events, and HUDSON and KAPLAN (1985) demonstrated that many more recombination events probably took place. For illustrative purposes only, we now attempt to "undo" two putative recombination events to obtain sequences less influenced by recombination. Suppose that the adult intron of the sequence, designated by KREITMAN as Fl-2S, was introduced into this sequence by recombination or conversion and that, before the recombination event, the adult intron was like that of any of the other *S*-allele adult introns. Similarly, suppose that the 5′ end of Fl-F was derived from a recombination event with an *S*-allele and that, before that recombination event, the 5′ sequence of Fl-F was like that of any of the other *F*-allele sequences. By undoing these two putative recombination events, the number of sites segregating within both allelic classes is zero. The resulting numbers of segregating sites in the three categories are shown in Table 2 on the line labeled "observed (adjusted)." This adjustment improves the agreement between the observed and predicted values. Given the high variances of these quantities (see Figure 3), it is unlikely that the differences between these observed and predicted numbers of sites are significant. It is interesting to note that, whether adjusted or not, the observed numbers differ from the expected by having too many sites segregating within the allelic classes, and not enough sites distinguishing the two allelic classes. It is, of course, possible that a proper analysis of the model with recombination would lead to a different conclusion.

DISCUSSION

With the advent of new molecular techniques for DNA sequencing and restriction-site mapping, additional data about DNA variation within and between electrophoretic allelic classes are becoming available. Since these new methods are laborious when compared to electrophoresis, it is natural to study genetic variation in populations with a stratified sampling scheme. For example, one might first survey a large sample of genes electrophoretically and then sequence or determine restriction enzyme maps for random subsamples from the different allelic classes. In this paper we have developed for a strictly neutral model with no recombination, a recursion scheme that allows us to study properties of the stationary distribution of the number of segregating sites in a subsample from a single allelic class conditional on the allelic configuration of the larger sample. We find that, for a subsample from a single allelic class, the number of segregating sites in the subsample depends on the frequency (k_1/n) of that allele in the full sample, but not on the number of other alleles or their frequencies. If θ_1 is small, a remarkably simple approximation is quite accurate; namely, the expected number of segregating sites in a subsample of size k_0, from an allelic class of frequency k_1/n in the full sample, is k_1/n times the expected number of segregating sites in a random sample of size k_0. The conditional distribution of the number of segregating sites in random subsamples from more than one allelic class is much more difficult to study; therefore, we have only dealt with the case in which there are two alleles in the large sample. This case is of interest because it is suitable for analyzing the data of KREITMAN (1983) as discussed below.

Whether the genetic variation detected by protein electrophoresis is primarily the result of mutation and drift of selectively neutral alleles, or is a consequence of selective forces, is a question that has remained essentially unanswered despite nearly two decades of effort to resolve it. In the past the only available information for answering this question was the number of alleles and their frequencies at different loci in various samples from populations, and consequently, methods were developed to test the neutral model based on these quantities (see FUERST, CHAKRABORTY and NEI 1977; WATTERSON 1978; EWENS 1979). These tests do not have great power and so the results have been ambiguous. The methods that we have developed here to analyze sequence data and restriction enzyme map data in conjunction with electrophoretic data do not, in themselves, provide a formal statistical procedure for testing the neutral model. They are, however, useful in making an informal assessment of the fit of the data to predictions under the neutral model and predictions under certain alternative selective models.

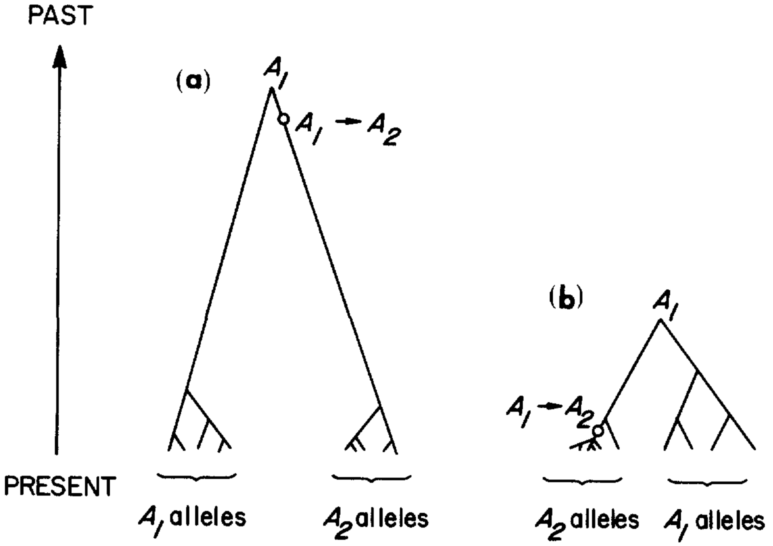

FIGURE 5.—Two histories illustrating the effects of selection. a, If the two alleles, A_1 and A_2, have been maintained for a long time by some form of balancing selection, the number of nucleotide site differences between an A_1 allele and an A_2 allele would be very large compared to the number of sites differing between two A_1 alleles or two A_2 alleles. b, If a newly arisen allele (A_2) rises to high frequency rapidly due to selection, there would be very few nucleotide site differences between A_2 alleles.

We consider two alternative hypotheses to the neutral model. Under the first it is assumed that two alleles are maintained in a population by some form of balancing selection. Unless selection is very weak or the population size small, such a polymorphism is expected to last indefinitely, as long as the selection forces are maintained. In this case a genealogy of a sample of genes is probable, such as that shown in Figure 5a, and so the number of segregating sites that distinguish the two alleles would be expected to be large compared to the numbers of sites that segregate within each allelic class. Another alternative hypothesis to neutrality is that one of the two alleles in the population arose recently by mutation and was rapidly driven to moderate or high frequency by selection. The genealogy of a sample in this case would probably look like that in Figure 5b, and consequently, one would expect few segregating sites within the subsample of genes from the new allelic class. These two alternative hypotheses are overly simple and perhaps naive, but they do give some idea of the effect that selection can have.

To illustrate these ideas we consider the set of 11 DNA sequences obtained by KREITMAN (1983). In Table 2, the observed, the observed (adjusted) and the predicted numbers of segregating sites are presented for sequences within the *F*-allelic class, within the *S*-allelic class and distinguishing the two classes. The number of differences distinguishing the classes is less than the expected number under the neutral model, and the number of segregating sites within each class is greater than or nearly equal to the expected number under the neutral model. Thus, the departures from the neutral predictions are, in fact, in the opposite directions from the departures expected under both of the alternative hypotheses described above.

It is possible that additional recombination events could obscure the effects of selection (HUDSON and KAPLAN 1985). If we study a shorter region around the *F-S* polymorphism, then recombination should be less important. Hence, we consider just intron 3, exon 4 and the 3′ untranslated region. In this region there are 15 polymorphic nucleotide sites, not counting the *F-S* polymorphism, and so our estimate of θ_2 for this region is $2(15)/(6.6) = 4.55$, assuming $k_1/n = 0.2$ (see NUMERICAL RESULTS). For $\theta_2 = 4.55$, the expected numbers of segregating sites within the *F*-allelic class, within the *S*-allelic class and distinguishing the classes are 1.90, 7.93 and 5.17, respectively. The observed numbers are 2, 11 and 2, respectively. Thus, in this case too, there is no indication that the numbers of segregating sites within the *F*-allelic class is too small, nor that the number of sites that distinguish the two classes is too large.

In conclusion, we are unable to detect any significant departure from stationary neutral expectations in the sample of *Adh* sequences. We do not find that the *F*-allele sequences are too similar to each other, nor do we find that the *F*-allele sequences differ too much from the *S*-allele sequences. It is, of course, possible that a more refined analysis could provide the power to do so. Thus, our analysis suggests that *Adh* is not a locus with two very old alleles maintained in the population by some form of balancing selection. We note that the possibility of three or more alleles being maintained by selection has not been rejected. (In this case, however, at least one of the alleles would have to be identical, at the protein level, to one of the other alleles.) Clearly, it would be desirable to characterize properties of sample genealogies under models with recombination, as well as more realistic models with selection.

LITERATURE CITED

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Pop. Biol. **3:** 87–112.

EWENS, W. J., 1979 *Mathematical Population Genetics.* Springer Verlag, New York.

FUERST, P. A., R. CHAKRABORTY and M. NEI, 1977 Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. Genetics **86:** 455–483.

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

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

JOHNSON N. L. and S. KOTZ, 1969 *Discrete Distributions.* Houghton Mifflin, Boston.

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

KINGMAN, J. F. C., 1982b The coalescent. Stochastic Processes Appl. **13:** 235–248.

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster.* Nature **304:** 412–417.

LEWONTIN, R. C., 1974 *The Genetic Basis of Evolutionary Change.* Columbia University Press, New York.

OAKSHOTT, J. G., J. B. GIBSON, P. R. ANDERSON, W. R. KNIBB, D. G. ANDERSON and G. K. CHAMBERS, 1981 Alcohol dehydrogenase and glycerol-3-phosphate dehydrogenase clines in *Drosophila melanogaster* on three continents. Evolution **36:** 86–96.

TAVARÉ, S., 1984 Line-of-descent and genealogical processes, and their application in population genetics models. Theor. Pop. Biol. **26:** 119–164.

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

WATTERSON, G. A., 1978 Heterosis or neutrality? Genetics **85:** 789–814.

Communicating editor: W. J. EWENS