The Structure of Genealogies and the Distribution of Fixed Differences Between DNA Sequence Samples From Natural Populations

Jody Hey

Department of Biological Sciences, Rutgers University, Nelson Laboratories, Piscataway, New Jersey 08855

Manuscript received October 8, 1990
Accepted for publication April 13, 1991

ABSTRACT

When two samples of DNA sequences are compared, one way in which they may differ is in the presence of fixed differences, which are defined as sites at which all of the sequences in one sample are different from all of the sequences in a second sample. The probability distribution of the number of fixed differences is developed. The theory employs Wright-Fisher genealogies and the infinite sites mutation model. For the case when both samples are drawn randomly from the same population it is found that genealogies permitting fixed differences are very unlikely. Thus the mere presence of fixed differences between samples is statistically significant, even for small samples. The theory is extended to samples from populations that have been separated for some time. The relationship between a simple Poisson distribution of mutations and the distribution of fixed differences is described as a function of the time since populations have been isolated. It is shown how these results may contribute to improved tests of recent balancing or directional selection.

COMPARISONS of DNA sequences from different species often make use of the typological view that observed differences are characteristic of the species sampled. In fact, closely related species are expected to share sequence variation to the extent that it has persisted since the time of species divergence (see, *e.g.*, PAMILO and NEI 1988). One way of describing DNA sequence differences between recently diverged populations is to measure the average number of differences found among all possible comparisons of sequences from one population with sequences from the second population. The net divergence (NEI and LI 1979) is equal to this quantity less the average of the variation within each of the two populations. TAKAHATA and NEI (1985) have found the variance of net divergence under standard assumptions: infinite sites model (KIMURA, 1969); no recombination; and Wright-Fisher genealogies.

An alternative descriptor of sequence divergence is the number of fixed differences, which is defined as the number of sites at which all of the sequences in one sample are different from all of the sequences in a second sample. Unlike net divergence, the number of fixed differences is a meristic character of the type used for maximum parsimony reconstruction of evolutionary trees.

The study of fixed differences between samples of DNA sequences was made by considering the genealogical process of gene samples. Under the model and assumptions used by TAKAHATA and NEI (1985), there are two conditions that must be met for fixed differences to occur between two random samples of genes: (1) the genealogy of the two samples must include a clade of all items in one of the samples, and this clade must be exclusive of items in the other sample; and (2) mutations must occur on the branch of the tree that connects the exclusive clade to the other lineages.

Figure 1, in which three possible genealogies are depicted for two samples of three sequences each, illustrates the conditions necessary for fixed differences. In Figure 1A, there is no node that represents the common ancestor of all of sample *A*, exclusive of sample *B*; nor is there a node that represents the common ancestor of all of sample *B*, exclusive of sample *A*. In this case, it is not possible to have a mutation in the genealogy of these samples (*e.g.*, indicated as a tick mark somewhere on the drawing) that is passed on to only sample *A* or only sample *B*. In Figure 1B, node *3* represents a common ancestor for only sample *A* as does node *2* for sample *B*. Any mutations that occurred in the lineage that persisted between node *1* and node *2* or the lineage between node *1* and node *3* would be observed as a fixed difference between the two samples. Figure 1C is similar, except that there is no node that represents the common ancestor for only sample *B*. In this case, mutations in the lineage that persisted between node *2* and node *3* would appear as fixed differences in the sample.

The theory begins with a consideration of two samples of sequences that are both drawn randomly from the same population. The theory is then extended to samples from populations that have been isolated for some period of time. In both cases, expressions for

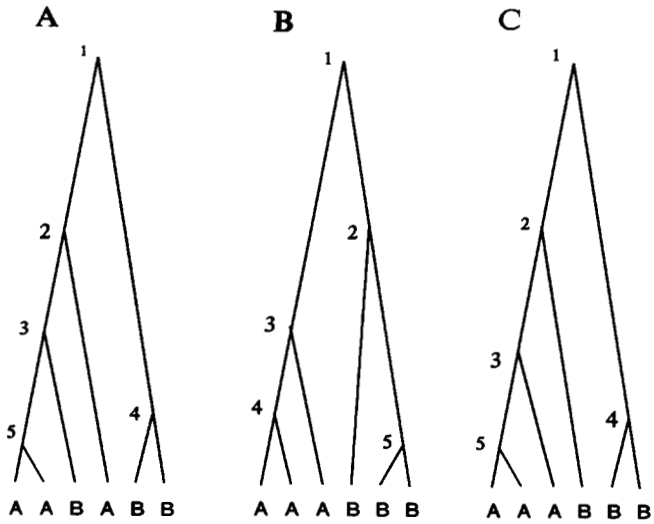

FIGURE 1.—Examples of sample genealogies.

the expectation, the variance and the probability density of the number of fixed differences are found.

THEORY

Samples from a single population

Genealogical structures: Consider the standard Wright-Fisher model commonly used in coalescent modeling (*i.e.*, a discrete generation model with N diploid individuals in which each generation of $2N$ gene copies is formed by sampling $2N$ times, with replacement, from the previous generation). Assume that recombination does not occur. If a random sample of n gene copies are drawn from the population in generation zero, then the probability that these n were descended from $n - 1$ ancestors t generations prior to generation zero is, to a close approximation,

$$\frac{\binom{n}{2}}{2N}\left[1 - \frac{\binom{n}{2}}{2N}\right]^{t-1}. \qquad (1)$$

This geometric distribution can be closely approximated with an exponential distribution having parameter $\binom{n}{2}/2N$. For the remainder of this paper, the convention of using time in units of $2N$ generations will be followed. Thus the probability that the waiting time until n items have $n - 1$ ancestors is t, in units of $2N$ generations, is approximately equal to

$$\binom{n}{2} e^{-\binom{n}{2}t}. \qquad (2)$$

See HUDSON (1990) for an accessible review of coalescent population genetic modelling.

A bifurcating genealogical tree of a sample of n items has $n - 1$ branch nodes. For convenience the nodes are indexed so that in the time between node i and node $i - 1$ there are i separate lineages. The state of n separate lineages at $t = 0$ is referred to as node n and the root node is referred to as node *1*. Thus, if we consider the construction of coalescent genealogies as a process that extends into the past, node i is formed by the joining of two of $i + 1$ lineages. The distribution of the length of this time interval is exponential with parameter $\binom{i}{2}$.

Consider the case where two samples have been randomly drawn from a single Wright-Fisher population. Let n_A and n_B represent the sizes of samples A and B, respectively, and let $n = n_A + n_B$. If all sample A lineages coalesce such that there is a node i that represents a most recent common ancestor to all n_A lineages, then we say that sample A coalesces at node i; or alternatively, that sample A forms a clade at node i. If that clade includes no lineages of sample B, then we say that the clade is exclusive.

Calculation of the probability density of fixed differences between two samples begins with a combinatorics problem (*i.e.*, how many of all possible genealogies include an exclusive clade of one of the samples), and then proceeds through several nested levels of assessment of conditional probabilities. Assessment of the distribution of the time interval during which mutations can appear as fixed differences between the samples proceeds through three distinct steps.

The probability that one of the samples forms an exclusive clade at node *i*: It is useful to begin by finding the total number of possible genealogies. Consider that any one of $\binom{i+1}{2}$ possible pairs of lineages will join (i.e. have a most recent common ancestor) at node i. Since a sample of n items will have $n - 1$ branch nodes, the total number of possible genealogies is

$$F(n) = \prod_{i=2}^{n} \binom{i}{2} = \frac{n!(n-1)!}{2^{(n-1)}}. \qquad (3)$$

Note this quantity is larger than the number of tree topologies (see, *e.g.*, FELSENSTEIN 1978). For every topology or branching pattern there may be multiple genealogies that are distinct in terms of how the nodes are ordered in time.

Let $P(i)$ be the probability that either of the samples forms an exclusive clade at node i. Consideration of genealogical structures (see Figure 1), shows that if one sample forms an exclusive clade at node i, for $i \geq 3$, then the other sample must necessarily form either an exclusive clade at node *2* or a nonexclusive clade at node *1*. By this reasoning we let $P(i) = 0$ for $i < 3$. It also follows that an exclusive clade for a sample, say sample A, must include exactly $n_A - 2$ branch nodes, not including the clade node. Thus it is not possible for sample A to form an exclusive clade at node i for $i > (n - 1) - (n_A - 2) = n_B + 1$, nor is it possible for sample B to form an exclusive clade at node i,

for $i > n_A + 1$. It follows that $P(i) = 0$ for $i > \max(n_A + 1, n_B + 1)$.

Calculation of $P(i)$ requires enumeration of the possible genealogical structures involving $n - i$ nodes of index greater than or equal to i that include an exclusive clade at node i. In the case of an exclusive clade of sample A lineages, $n_A - 1$ of the $n - i$ nodes join pairs of sample A lineages and $(n - i) - (n_A - 1) = n_B - i + 1$ join pairs of sample B lineages.

The calculation of the number of genealogies that are possible between node n and node i, with an exclusive clade of sample A at node i, includes three components: the number of ways that n_A items can coalesce,

$$F(n_A); \tag{4A}$$

the number of ways that sample B lineages can coalesce through $n_B - i + 1$ nodes,

$$\frac{F(n_B)}{F(n_B - (n_B - i + 1))} = \frac{F(n_B)}{F(i - 1)}; \tag{4B}$$

and the number of ways that $n_A - 2$ nodes and $n_B - i + 1$ nodes can be ordered in time,

$$\binom{n - i - 1}{n_A - 2}. \tag{4C}$$

Between node n and node i there are a total of $F(n)/F(i)$ possible genealogical constructions. The product of quantities (4A), (4B), and (4C) is the number that contain exclusive clades of sample A lineages that coalesce at node i. Calculation of $P(i)$ also includes the quantity $F(i)$ since there are this many different genealogies between nodes i and 1, for every partial genealogy between node n and node i. In summary, the number of genealogies that include an exclusive clade of sample A at node i is

$$\binom{n - i - 1}{n_A - 2}\frac{F(n_A)F(n_B)F(i)}{F(i - 1)} = \binom{n - i - 1}{n_A - 2}\frac{i(i - 1)n_A!^2 n_B!^2}{2^{(n-1)}n_A n_B}. \tag{5}$$

For an exclusive clade of sample B lineages,

$$\binom{n - i - 1}{n_B - 2}$$

replaces (4C) in (5). Then

$$P(i) = F(n)^{-1}\left(\binom{n - i - 1}{n_A - 2} + \binom{n - i - 1}{n_B - 2}\right)\frac{i(i - 1)n_A!^2 n_B!^2}{2^{(n-1)}n_A n_B}. \tag{6}$$

The number of genealogies that include an exclusive clade of either sample is

$$\sum_{i=3}^{\max(n_A+1,n_B+1)}\left(\binom{n - i - 1}{n_A - 2} + \binom{n - i - 1}{n_B - 2}\right) \times \frac{i(i - 1)n_A!^2 n_B!^2}{2^{(n-1)}n_A n_B} \tag{7}$$

The probability that the branch from node i connects with node j: Let $P(j|i)$ represent the probability that node i is directly connected to node j, where $i - 1 \geq j \geq 1$. There are i lineages in the time interval between nodes i and $i - 1$, and there are $\binom{i}{2}$ possible pairings among them. Since only $i - 1$ of these possible pairings include the lineage that originates at node i, we can describe the probability that node i connects directly with node $i - 1$:

$$\frac{i - 1}{\binom{i}{2}} = \frac{2}{i}.$$

Similarly,

$$\left(1 - \frac{2}{i}\right)\frac{2}{i - 1},$$

is the probability that node i connects directly with node $i - 2$. Then

$$P(j|i) = \frac{2}{j + 1}\prod_{k=0}^{i-j-2}\left(1 - \frac{2}{i - k}\right) = \frac{2j}{i(i - 1)}. \tag{8}$$

The probability that the length of the branch, L, is l: Since the distribution of lengths between any pair of nodes i and j is exponential with parameter $\binom{i}{2}$, the expected length between nodes i and j is

$$E(L|i, j) = \frac{1}{\binom{i}{2}} + \frac{1}{\binom{i - 1}{2}} + \ldots + \frac{1}{\binom{j + 1}{2}} = \frac{2}{j} - \frac{2}{i}, \quad \text{for} \quad j > 1. \tag{9A}$$

Equation 9A does not apply for $j = 1$ because in this case the sample that did not form an exclusive clade at node i necessarily forms an exclusive clade at node 2 (see Figure 1). When $j = 1$, the total length to be considered includes twice the distance between nodes 1 and 2. Thus

$$E(L|i,1) = E(L|i, 2) + 2 = 3 - \frac{2}{i}. \tag{9B}$$

The same reasoning for the variance yields

$$V(L|i,j) = \frac{1}{\binom{i}{2}^2} + \frac{1}{\binom{i-1}{2}^2} + \ldots + \frac{1}{\binom{j+1}{2}^2},$$

for $j > 1$, (10A)

and

$$V(L|i,1) = V(L|i,2) + 4. \quad (10B)$$

Let $P(L = l|i,j)$ be the probability that the length of the branch between node i and node j is l. Since the length is the sum of $i - j$ independent random variables, the probability can be calculated by taking the convolution of $i - j$ exponential distributions. From TAKAHATA and NEI (1985), for $j > 1$,

$$P(L = l|i,j) = \left(\prod_{r=j+1}^{i}\binom{r}{2}\right)\sum_{r=j+1}^{i}\frac{e^{-\binom{r}{2}l}}{\prod_{\substack{s=j+1\\s\neq r}}^{i}\binom{s}{2}-\binom{r}{2}}. \quad (11)$$

When $j = 1$, L will include twice the distance between nodes *1* and *2*. If we let z be an instance of a random variable representing the time between nodes *1* and *2* then

$$P(L = l|i,1) = \int_0^{l/2} P(L = l - 2z|i,2)e^{-z}\,dz$$

$$= \int_0^{l} P(L = z|i,2)\frac{e^{-(l-z)/2}}{2}\,dz.$$

This can also be represented

$$\left(\sum_{r=2}^{i}\rho(r)\right)\sum_{r=2}^{i}\frac{e^{-\rho(r)l}}{\prod_{\substack{s=2\\s\neq r}}^{i}\rho(s)-\rho(r)}, \quad (12)$$

where

$$\rho(r) = \binom{r}{2}, \quad \text{for } r > 2;$$

$$\rho(r) = \frac{1}{2}, \quad \text{for } r = 2.$$

The distribution of fixed differences: Because the neutral mutation process is isolated from the genealogical process in coalescent models, the probability density of the number of mutations given l, $P(M = m|l)$, will vary depending on what mutation model is used. For the question at hand, we use the infinite sites model (KIMURA 1969), in which every mutation is assumed to occur at a different site and is distinguishable from all other mutations. The distribution of the number of mutations is Poisson so that

$$P(M = m|l) = \frac{e^{-\mu l}(\mu l)^m}{m!}, \quad (13)$$

where μ is the neutral mutation rate per $2N$ generations (*i.e.*, $2N$ times the neutral mutation rate per generation). Furthermore, for a branch between nodes i and j,

$$E(M|i,j) = \mu E(L|i,j), \quad (14)$$

and

$$V(M|i,j) = \mu E(L|i,j) + \mu^2 V(L|i,j). \quad (15)$$

Combining the components of genealogical structure and mutation yields a compound distribution that represents the probability that samples A and B are separated by m mutations:

$$P_{n_A,n_B}(M=m) = \int_0^\infty P(M=m|l)$$

$$\sum_{i=3}^{\max(n_A+1,n_B+1)} P(i)\sum_{j=1}^{i-1}P(j|i)P(L=l|i,j)\,dl$$

$$= \mu^m \sum_{i=3}^{\max(n_A+1,n_B+1)} P(i)\sum_{j=1}^{i-1}P(j|i)\prod_{r=j+1}^{i}\rho(r)$$

$$\sum_{r=j+1}^{i}\left((\mu+\rho(r))^{m+1}\prod_{\substack{s=j+1\\s\neq r}}^{i}\rho(s)-\rho(r)\right)^{-1},$$

for $m > 0$. (16)

Note that $\mu^m/(\mu + \rho(r))^{m+1}$ is a monotonically decreasing function of m, and thus so is $P_{n_A,n_B}(M = m)$. This means that the mode of this distribution is always zero.

It is clear that there are two distinct causes for an absence of fixed differences between the samples. One occurs when no mutations happen on the branch that separates the genealogies of the two samples. The probability of this can be calculated using expression (16) with $m = 0$. The second cause of zero mutations is when neither sample has a genealogy exclusive of the other (*i.e.*, no exclusive clades). The probability of this occurring is equal to

$$1 - \sum_{i=3}^{\max(n_A+1,n_B+1)} P(i).$$

When $m = 0$, this quantity should be added to expression (16) to find the total probability of zero mutations. To further distinguish these types of events it may be useful to calculate the probability of m mutations *given that the genealogy includes a clade of one of the samples.* This can be calculated simply by replacing $F(n)$ in expression (6) with expression (7) and then using this conditional form of $P(i)$ in (16).

The expectation and variance of M can be found without resort to (16):

$$E(M) = \mu E(L)$$
$$= \sum_{i=3}^{\max(n_A+1,n_B+1)} P(i) \sum_{j=1}^{i-1} P(j|i)E(M|i, j); \quad (17)$$

and

$$V(M) = \sum_{i=3}^{\max(n_A+1,n_B+1)} P(i) \sum_{j=1}^{i-1} P(j|i)V(M|i, j)$$
$$+ \sum_{i=3}^{\max(n_A+1,n_B+1)} P(i) \sum_{j=1}^{i-1} P(j|i)E(M|i, j)^2$$
$$- \left(\sum_{i=3}^{\max(n_A+1,n_B+1)} P(i) \sum_{j=1}^{i-1} P(j|i)E(M|i, j) \right)^2. \quad (18)$$

Samples from isolated populations

Genealogies and mutations during isolation: Consider the case when samples are taken from two populations that have been isolated for some time T (in units of $2N$ generations), and assume that both populations are the same size as the ancestral population. For sample A, the probability that n_A genes sampled at time zero are descended from exactly n_{A_T} ancestral genes T units of time ago, is given by TAKAHATA and NEI (1985). For $n_A > n_{A_T} > 1$,

$$P_{n_A}(n_{A_T}|T) = \left(\prod_{r=n_{A_T}+1}^{n_A} \binom{r}{2} \right) \sum_{r=n_{A_T}}^{n_A} \frac{e^{-\binom{r}{2}T}}{\prod_{\substack{s=n_{A_T} \\ s\neq r}}^{n_A} \binom{s}{2} - \binom{r}{2}}. \quad (19A)$$

For $n_{A_T} = n_A$ and $n_{A_T} = 1$ the quantities are

$$P_{n_A}(n_A|T) = 1 - \int_0^T \binom{n_A}{2} e^{-\binom{n_A}{2}T} dt = e^{-\binom{n_A}{2}T}, \quad (19B)$$

and

$$P_{n_A}(1|T) = \int_0^T P(L = l|n_A, 1)\, dl$$
$$= \left(\prod_{r=2}^{n_A} \binom{r}{2} \right) \sum_{r=2}^{n_A} \frac{1 - e^{-\binom{r}{2}T}}{\binom{r}{2} \prod_{\substack{s=2 \\ s\neq r}}^{n_A} \binom{s}{2} - \binom{r}{2}}, \quad (19C)$$

respectively. Quantity (19C) can also be represented as

$$\phi(n_A) \sum_{r=2}^{n_A} \frac{1 - e^{-\binom{r}{2}T}}{\binom{r}{2} \psi(n_A, r)},$$

where

$$\phi(n_A) = \prod_{r=2}^{n_A} \binom{r}{2}$$

and

$$\psi(n_A, r) = \prod_{\substack{s=2 \\ s\neq r}}^{n_A} \binom{s}{2} - \binom{r}{2}.$$

For those cases when one sample coalesces completely between the time of sampling and T, there will be an interval during which mutations can occur and appear as fixed differences. The probability that this interval, Λ, takes on the value λ can, in the case of sample A, be described by the truncated distribution

$$P_{n_A,T}(\Lambda = \lambda|T) = \frac{P(L = T - \lambda|n_A, 1)}{P_{n_A}(1|T)}, \quad \text{for } 0 \le \lambda \le T. \quad (20)$$

The numerator, $P(L = T - \lambda|n_A, 1)$, is calculated with expression (12) with the exception that $\binom{r}{2}$ is used rather than $\rho(r)$. The first two moments of this distribution are

$$E_{n_A,T}(\Lambda|T) = \frac{T - \sum_{r=2}^{n_A} \binom{r}{2}^{-1} + \phi(n_A) \sum_{r=2}^{n_A} e^{-\binom{r}{2}T} \Big/ \binom{r}{2}^2 \psi(n_A,r)}{P_{n_A}(1|T)} \quad (21)$$

and

$$E_{n_A,T}(\Lambda^2|T) = \frac{\left(T - \sum_{r=2}^{n_A} \binom{r}{2}^{-1}\right)^2 + \sum_{r=2}^{n_A} \binom{r}{2}^{-2} - 2\phi(n_A) \sum_{r=2}^{n_A} e^{-\binom{r}{2}T} \Big/ \binom{r}{2}^3 \psi(n_A, r)}{P_{n_A}(1|T)}. \quad (22)$$

Thus the variance is

$$V_{n_A,T}(\Lambda|T) = E_{n_A,T}(\Lambda^2|T) - E_{n_A,T}(\Lambda|T)^2. \quad (23)$$

For a given interval λ, the distribution of mutations is Poisson as in (13). However, when λ represents an instance of a random variable, the probability that m mutations occur is a compound distribution,

$$P_{n_A,T}(M = m|T) = \int_0^T P_{n_A,T}(\Lambda = \lambda|T) P(M = m|\lambda)\, d\lambda,$$

which simplifies to

$$\frac{\mu^m \phi(n_A)}{P_{n_A}(1|T)} \sum_{r=2}^{n_A} \frac{e^{-(\mu-\binom{r}{2})T} \left(1 - e^{-(\mu-\binom{r}{2})T} \sum_{k=0}^{m} \left(\mu - \binom{r}{2}\right)^k T^k \Big/ k!\right)}{\psi(n_A,r) \left(\mu - \binom{r}{2}\right)^{m+1}}. \quad (24)$$

In determining the moments of this distribution, the Poisson terms in the numerator of (24) where summed over the range of m ($0 \leq m < \infty$) and it was found that the double summation simplified to a Taylor series of an exponential function. Thus, the first and second moments are

$$E_{n_A,T}(M|T) = \frac{\mu\phi(n_A)}{P_{n_A}(1|T)} \sum_{r=2}^{n_A} \frac{\binom{r}{2}T + e^{-\binom{r}{2}T} - 1}{\psi(n_A,r)\binom{r}{2}^2} \quad (25)$$

and

$$E_{n_A,T}(M^2|T) = \frac{\mu\phi(n_A)}{P_{n_A}(1|T)} \sum_{r=2}^{n_A} \frac{\binom{r}{2}^2 T^2\mu + \binom{r}{2}^2 T - 2\binom{r}{2}T\mu - \binom{r}{2} + 2\mu + \left(\binom{r}{2} - 2\mu\right)e^{-\binom{r}{2}T}}{\psi(n_A,r)\binom{r}{2}^3}, \quad (26)$$

respectively. The variance follows from the usual relation

$$V_{n_A,T}(M|T) = E_{n_A,T}(M^2|T) - E_{n_A,T}(M|T)^2. \quad (27)$$

By employing the relations

$$\sum_{r=2}^{n_A} \frac{1}{\binom{r}{2}\psi(n_A, r)} = \frac{1}{\phi(n_A)},$$

$$\sum_{r=2}^{n_A} \frac{1}{\binom{r}{2}^2 \psi(n_A, r)} = \frac{1}{\phi(n_A)} \sum_{r=2}^{n_A} \frac{1}{\binom{r}{2}},$$

and

$$\sum_{r=2}^{n_A} \frac{1}{\binom{r}{2}^3 \psi(n_A, r)} = \frac{1}{2\phi(n_A)} \left[\left[\sum_{r=2}^{n_A} \frac{1}{\binom{r}{2}} \right]^2 + \sum_{r=2}^{n_A} \frac{1}{\binom{r}{2}^2} \right],$$

it can be shown that

$$E_{n_A,T}(M|T) = \mu E_{n_A,T}(\Lambda|T)$$

and

$$V_{n_A,T}(M|T) = \mu E_{n_A,T}(\Lambda|T) + \mu^2 V_{n_A,T}(\Lambda|T).$$

The distribution of the number of mutations in the interval Λ should come to resemble a Poisson distribution with parameter T for high values of T. This is because λ becomes closer to the fixed quantity T as T increases, and for a fixed time interval the distribution of mutations is Poisson (see 13). One way to check this is to examine the difference between the expectation given in (25) and the expectation of the Poisson distribution, μT, for increasing values of T. We find

$$\lim_{T\to\infty}(\mu T - E_{n_A,T}(M|T)) = \mu \sum_{r=2}^{n_A} \frac{1}{\binom{r}{2}}.$$

Thus we see that at the limit, the expected number of mutations is equal to that expected if Λ where fixed at T less the expected number of mutations that would occur in the coalescent time of a sample of size n_A (see 14).

The distribution of fixed differences between samples from isolated populations: Expressions developed in the previous section, together with expressions (13) through (18), can be used to describe the expectation and the variance of the distribution of fixed differences between populations isolated for some time T. The expectation,

$$E_{n_A,n_B}(M|T) = \sum_{n_{A_T}=1}^{n_A} P_{n_A}(n_{A_T}|T) \sum_{n_{B_T}=1}^{n_B} P_{n_B}(n_{B_T}|T) E(M|n_{A_T}, n_{B_T}), \quad (28)$$

requires calculation of $E(M|n_{A_T}, n_{B_T})$. When $n_{A_T} > 1$ and $n_{B_T} > 1$ the situation is similar to that in (14), so that

$$E(M|n_{A_T}, n_{B_T}) = \sum_{i=3}^{\max(n_{A_T}+1, n_{B_T}+1)} P(i) \sum_{j=1}^{i-1} P(j|i) E(M|i,j). \quad (29A)$$

When either $n_{A_T} = 1$ or $n_{B_T} = 1$, then, in the case of $n_{A_T} = 1$,

$$E(M|1, n_{B_T}) = \sum_{j=1}^{n_{B_T}} (j|n_{B_T}+1)(E(M|n_{B_T}+1, j) + E_{n_A,T}(M|T)). \quad (29B)$$

When both n_{A_T} and n_{B_T} are equal to 1 then

$$E(M|1,1) = E_{n_A,T}(M|T) + E_{n_B,T}(M|T) + 2\mu. \quad (29C)$$

The expression for the variance,

$$V_{n_A,n_B}(M|T) = \sum_{n_{A_T}=1}^{n_A} P_{n_A}(n_{A_T}|T) \sum_{n_{B_T}=1}^{n_B} P_{n_B}(n_{B_T}|T) V(M|n_{A_T}, n_{B_T}) + \sum_{n_{A_T}=1}^{n_A} P_{n_A}(n_{A_T}|T) \sum_{n_{B_T}=1}^{n_B} P_{n_B}(n_{B_T}|T) E(M|n_{A_T}, n_{B_T})^2 - \left(\sum_{n_{A_T}=1}^{n_A} P_{n_A}(n_{A_T}|T) \sum_{n_{B_T}=1}^{n_B} P_{n_B}(n_{B_T}|T) E(M|n_{A_T}, n_{B_T}) \right)^2, \quad (30)$$

follows a similar development. When $n_{A_T} > 1$ and $n_{B_T} > 1$,

$$V(M|n_{A_T}, n_{B_T}) = \sum_{i=3}^{\max(n_{A_T}+1, n_{B_T}+1)} P(i) \sum_{j=1}^{i-1} P(j|i)V(m|i,j) + \sum_{i=3}^{\max(n_{A_T}+1, n_{B_T}+1)} P(i) \sum_{j=1}^{i-1} P(j|i)E(M|i,j)^2 - \left(\sum_{i=3}^{\max(n_{A_T}+1, n_{B_T}+1)} P(i) \sum_{j=1}^{i-1} P(j|i)E(M|i,j) \right)^2. \quad (31A)$$

When either $n_{A_T} = 1$ or $n_{B_T} = 1$ then, in the case of $n_{A_T} = 1$,

$$V(M \mid 1, n_{B_T}) = \sum_{j=1}^{n_{B_T}} P(j|n_{B_T}+1)(V(M|n_{B_T}+1, j) + V_{n_A,T}(M|T)) + \sum_{j=1}^{n_{B_T}} P(j|n_{B_T}+1)(E(M|n_{B_T}+1, j) + E_{n_A,T}(M|T))^2 - \left(\sum_{j=1}^{n_{B_T}} P(j|n_{B_T}+1)(E(M|n_{B_T}+1, j) + E_{n_A,T}(M|T)) \right)^2. \quad (31B)$$

When both samples are completely coalesced,

$$V(M|1, 1) = V_{n_A,T}(M|T) + V_{n_B,T}(M|T) + 4\mu^2 + 2\mu. \quad (31C)$$

An expression for the probability density of the number of fixed differences can be constructed in a manner analogous to (16):

$$\begin{aligned} P_{n_A,n_B}(M = m|T) &= \int_0^\infty P(M = m|l) \sum_{n_{A_T}=1}^{n_A} P_{n_A}(n_{A_T}|T) \\ &\quad \sum_{n_{B_T}=1}^{n_B} P_{n_B}(n_{B_T}|T)P(L = l|n_{A_T}, n_{B_T})\, dl \\ &= \sum_{n_{A_T}=1}^{n_A} P_{n_A}(n_{A_T}|T) \sum_{n_{B_T}=1}^{n_B} P_{n_B}(n_{B_T}|T) \\ &\quad \int_0^\infty P(M = m|l)P(L = l|n_{A_T}, n_{B_T})\, dl. \end{aligned} \quad (32)$$

Evaluation of $P(L = l|n_{A_T}, n_{B_T})$ takes one of three routes depending on the values of n_{A_T} and n_{B_T}.

1. When both samples have multiple lineages at T (*i.e.*, for $n_{A_T} > 1$ and $n_{B_T} > 1$)

$$P(L = l|n_{A_T}, n_{B_T}) = \sum_{i=3}^{\max(n_{A_T}+1, n_{B_T}+1)} P(i) \sum_{j=1}^{i-1} P(j|i)P(L = l|i,j). \quad (33)$$

2. When one sample has coalesced prior to T and the other has not, the total length includes two independent random variables. For example, when sample A has completely coalesced prior to T, then

$$P(L = l|1, n_{B_T}) = \sum_{j=1}^{n_{B_T}} P(j|n_{B_T}+1) \int_0^\xi P_{n_A,T}(\Lambda = \lambda|T)P(L = l - \lambda|n_{B_T}+1, j)\, d\lambda, \quad (34)$$

where $\xi = T$ when $l \geq T$, and $\xi = l$ when $l < T$.

3. When both samples have coalesced prior to T then the total length is the sum of three independent random variables. Let the length of the tree between T and the time of node *1* be described by an exponential distribution with parameter *1/2* [see (11)]. Let Ω represent the sum of the length between the time of sample A coalescence and T and the length between the time of sample B coalescence and T. Then

$$P(\Omega = \omega) = \int_\gamma^\delta P_{n_A,T}(\Lambda = \lambda|T)P_{n_B,T}(\Lambda = \omega - \lambda|T)\, d\lambda. \quad (35)$$

When $0 \leq \omega < T$, $\gamma = 0$ and $\delta = \omega$. When $T \leq \omega \leq 2T$, $\gamma = \omega$ and $\delta = T$.

The distribution of the length of the sum of all three random variables is

$$P(L = l|1, 1) = \frac{1}{2} \int_0^\sigma P(\Omega = \omega)e^{-(l-\omega)/2}\, d\omega, \quad (36)$$

where $\sigma = l$ when $l \leq 2T$, and $\sigma = 2T$ when $l > 2T$.

RESULTS

Samples from a single population: Table 1 shows, for a variety of sample sizes, the proportion of all genealogies that include an exclusive clade. It is apparent that this proportion is less when the sample sizes are similar and decreases as total sample size increases. It is also clear that a small minority of genealogies include exclusive clades, even for small samples. This last finding leads to a simple statistical statement that may be appropriate for some samples. Under the assumptions of the model, any time that two samples are drawn from a natural population and observed to have one or more fixed differences between them, then the implied genealogical structure of the entire sample is so unlikely as to suggest a failure of the model. Put more briefly, anytime two random samples, one of size 3 or more and the other of size 4 or more, are found to have one or more fixed differences, then that number of differences is statistically significant.

It should be stressed that these results are only applicable when sample designations are applied prior to evaluation of fixed differences. Thus, for instance,

TABLE 1

The probability that the genealogies of the two samples are exclusive of each other

n_A	n_B	n	$\sum P(i)$
2	4	6	0.147
3	3	6	0.080
3	4	7	0.044
2	8	10	0.075
3	7	10	0.015
4	6	10	0.006
5	5	10	0.004
2	18	20	0.035
5	15	20	$9.0*10^{-5}$
10	10	20	$3.4*10^{-6}$
2	48	50	0.014
10	40	50	$9.0*10^{-11}$
25	25	50	$2.1*10^{-15}$

fixed differences observed between two samples collected from different localities could be used to reject a null hypothesis of panmixia. These results are not directly applicable to the case when some pattern of fixed differences is observed within a sample and one might wish to subdivide the sample on the basis of observed differences.

In some cases, prior knowledge about the genealogy of two samples can be useful. For example, the genealogy of two samples that are known to have a fixed difference between them must include an exclusive clade of one or both samples. Expression (16) can be easily modified to a conditional probability density,

$$P(M = m|M > 0) = \frac{P(M = m)}{1 - P(M = 0)}, \qquad (37)$$

the probability that two samples have m fixed differences given the presence of at least one fixed difference.

This distribution is especially appropriate for samples of genes that are known to differ in the electrophoretic mobility of their corresponding proteins. For example KREITMAN (1983) sequenced eleven copies of the *Drosophila melanogaster* Alcohol Dehydrogenase (Adh) gene including five copies associated with a fast (AdhF) electrophoretic phenotype and six copies associated with a slow (AdhS) electrophoretic phenotype. Assuming that the electrophoretic difference was caused by a single fixed difference, we can use expression (37) to ask whether the observed number of fixed differences are more (or less) than we expect by chance. More explicitly, if x is the observed number of fixed differences, then the probability of observing x or more fixed differences is

$$P(M \geq x|M > 0) = \frac{\sum_{m=x}^{\infty} P(M = m)}{1 - P(M = 0)} = 1 - \frac{\sum_{m=1}^{x-1} P(M = m)}{1 - P(M = 0)}. \qquad (38)$$

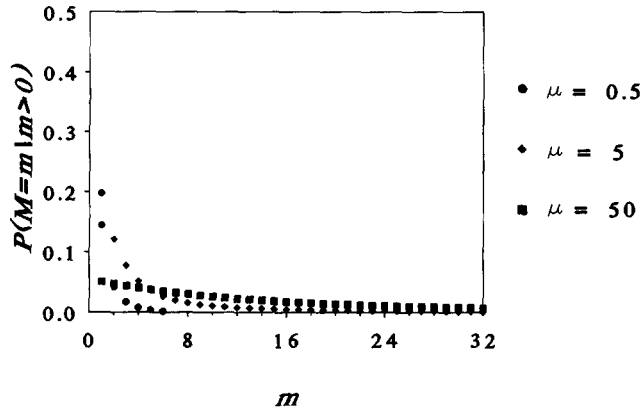

FIGURE 2.—The probability of m mutations given that m is greater than zero. Results for three different mutation rates are shown. In all cases $n_A = n_B = 5$.

A similar expression can be developed for the probability of observing x or fewer fixed differences.

In the case of the Kreitman data, the actual number of fixed differences is 3. The mutation rate per site per $2N$ generations was estimated (from the observed average heterozygosity per site) to be 0.003, which corresponds to a mutation rate per $2N$ generations for the entire length of 2721 base pairs of approximately 8. Using expression (38) we find that the probability of observing 3 or more fixed differences, given at least one fixed difference, is 0.603. One difficulty with this example is that two of the sequences (F1-2S and F1-F) appear to reflect recombination events involving sequences that otherwise include fixed differences between the two allele classes. Excluding these two sequences, we observe 7 fixed differences between a sample of 5 AdhS alleles and 4 AdhF alleles. The probability of 7 or more fixed differences is 0.3054.

This example illustrates a curious property of the probability distribution of fixed differences between samples from the same population. Regardless of the sample sizes and mutation rates, the mode of the distribution is always zero and the distribution decreases monotonically as m increases. Even for very large mutation rates, m fixed differences is always more likely than $m + 1$ fixed differences. For high mutation rates, the distribution becomes very flat so that a wide range of outcomes is likely. Figure 2 illustrates this property for two samples of 5 gene copies each.

In the case of the Kreitman data, we see that prior knowledge of at least one fixed difference was followed by the observation of 7 fixed differences (excluding the possible recombinants). Yet because the mutation rate is high (for 2721 sites) this is a likely outcome. Even an observation of 30 fixed differences would not be inconsistent with the model, since the probability of 30 or more fixed differences is 0.055.

Samples from isolated population: Examples of the expectation and variance of the number of fixed

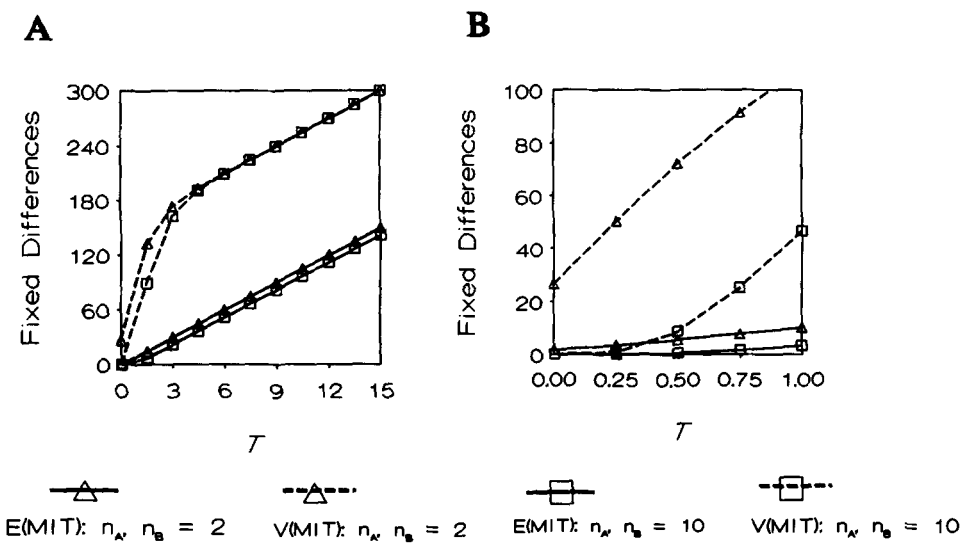

FIGURE 3.—The expectation and variance of the distribution of fixed differences between samples from isolated populations. In all cases, $\mu = 5$. A, provides results for a wide range of T values. B, provides results for low values of T.

differences, as functions of T, are given in Figure 3. From Figure 3A it is clear that both the expectation and variance become essentially linear functions of T with slopes of 2μ, for high values of T. This is expected because the major component is increasingly Poisson and the components due to sample coalescence become relatively less for increasing T. The vertical distance between the lines for the variance and those for the expectation is largely accounted for by the quantity $4\mu^2 + 2\mu$ [see (31C)]. For low values of T (Figure 3B), the distribution is dominated by the coalescent process. Though monotonically increasing, the expectation and variance have different slopes.

For the examples in Figure 3 and for a wide range of sample sizes and mutation rates, for which results are not shown, the variance as a function of time takes on a slope of approximately 2μ at or below $T = 3$. This suggests a generalization: for $T > 3$, the distribution of fixed differences between samples from isolated populations resembles a Poisson distribution having parameter $2T\mu$. In other words, when the time since isolation is greater than $6N$ generations, then the distribution of fixed differences fits the standard neutral model of molecular evolution.

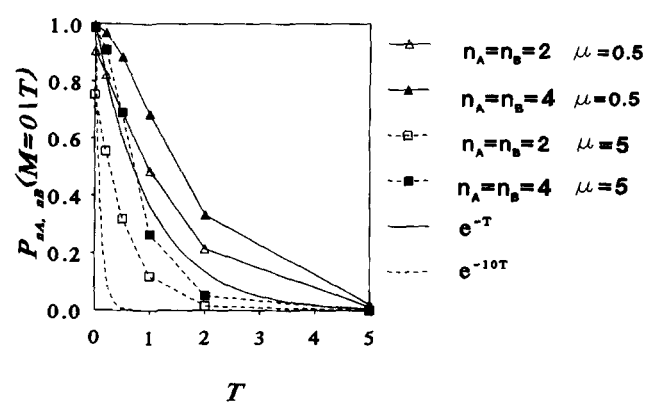

FIGURE 4.—The probability of zero fixed differences between samples from populations isolated for time T. Examples are given for two different sample sizes and for two different mutation rates. Also provided is the zero term from a Poisson distribution having parameter $2T\mu$.

The probability of zero fixed differences is plotted as a function of T in Figure 4. Also shown is the probability of zero events under a Poisson distribution of parameter $2T\mu$. Except for very low values of T, when all values are near zero, and for high values of T, the Poisson distribution considerably underestimates the probability of zero differences. This discrepancy is greater with higher mutation rates and larger sample sizes.

DISCUSSION

The probability distribution of fixed differences between samples was developed with two general goals. The first of these, which is shared with most quantitative evolutionary theory, was to promote an intuition on the part of investigators of how evolution might proceed under simplified conditions. In this light, one of the more interesting findings is the very low probability that the genealogy of samples from the same population includes an exclusive clade of either sample. Also of interest, and again in the case of samples from the same population, is that the probability of m fixed differences, $P(M = m)$, decreases monotonically with m, regardless of the sample size and mutation rate.

The second goal was to provide statistical tests with which to contrast observations with the assumptions of the model. Some examples of tests, which apply to the case of two samples from the same population, are described in RESULTS. For two reasons, that fixed differences are not expected and that if they do occur it is with a high variance, these tests will probably not prove very powerful for most questions.

The distribution of fixed differences between samples from isolated populations does not lead directly

to statistical tests, unless an estimate of T is provided. In addition, the unlikely assumption was made that both populations and the ancestral population are all of the same size. This assumption need not be made, as it would be a simple modification to include scalers, perhaps ν_A and ν_B, which would be the ratios of the ancestral population size to the size of population A and B, respectively. These scalers would then be multiplied times the exponential parameters when considering events within either population A or B. For example, the time between sampling and node $n_A - 1$ could be described by an exponential distribution having parameter $\nu_A\binom{n_A}{2}$. Thus the results for samples from isolate populations can lead to statistical tests in cases where there exist prior estimates of time of isolation and relative population sizes, or when these quantities can be estimated from the data. An example of a test that takes the latter approach is that of HUDSON, KREITMAN and AGUADE (1987), hereafter referred to as HKA.

The HKA test requires data on DNA sequence variation both from within and between two species, for each of two or more genetic loci. The population genetic model is the same as in this report, with the additional assumption that locus specific neutral mutation rates per chronological time are constant across species. The test proceeds by finding the values for locus specific mutation rates, time since divergence, and relative population sizes that are most consistent with the data and the assumptions of the model. These then lead to expected levels of sequence variation within and between species for each locus. With one degree of freedom remaining, in the case of two loci, the contrast between the observations and the expected values enables a test.

The HKA test is useful as a test of recent natural selection having acted within one of the species at one of the loci. If the form of selection was balancing, such that two or more functional alleles had persisted in a species for a long period of time, then sequence variation within that species at that locus is expected to be elevated relative to expectations (HUDSON and KAPLAN 1988; STROBECK 1983). If the form of selection was directional such that a recently rare functional allele increased in frequency and became fixed within one of the species, then the hitchhiking effect will cause sequence variation within that species and locus to be reduced relative to expectations (KAPLAN, HUDSON and LANGLEY 1989; MAYNARD SMITH and HAIGH 1974).

HUDSON, KREITMAN and AGUADE (1987) used as a measure of divergence between species, the divergence observed between single randomly picked sequences from each species. This quantity is very tractable in that the expectation and variance under the model is easily calculated. The quantity is not sensitive to natural selection, however, because the divergence of single sequences from isolated species is simply a function of time of divergence and not of genealogical processes within the species. An alternative descriptor of species divergence is the number of fixed differences as described in this report. Recent hitchhiking at a locus in one species is synonymous with a short genealogy for a sample of sequences from that species and locus. This means that the particular sequence associated with the functional allele that is favored by natural selection will become fixed in the population. Thus hitchhiking is expected to increase fixed differences between populations at the expense of sequence variation within populations. Similarly, balancing selection will lengthen the genealogy of samples that include sequences representing the different functional alleles. In this case, sequence variation within species will be increased at the expense of fixed differences between species. In summary, the expectation and variance of the number of fixed differences between species could be used in a modified HKA test, and these modifications are expected to increase the power of the test to reveal natural selection.

This work was supported by National Science Foundation grant BSR-8918164 to the author.

LITERATURE CITED

FELSENSTEIN, J., 1978 The number of evolutionary trees. Syst. Zool. **27:** 27–33.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7.**

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

KAPLAN, N., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to a steady flux of mutations. Genetics **61:** 893–903.

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

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23–35.

NEI, M., and W. H. LI, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA, **76:** 5269–5273.

PAMILO, P., and M. NEI, 1988 Relationships between gene trees and species trees. Mol. Biol. Evol. **5:** 568–583.

STROBECK, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics **13:** 545–555.

TAKAHATA, N., and M. NEI, 1985 Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110:** 325–344.

Communicating editor: R. R. HUDSON