Reconstructing the Prior Probabilities of Allelic Phylogenies

G. Brian Golding[1]

Department of Biology, McMaster University, Hamilton, Ontario, L8S 4K1, Canada

Manuscript received April 25, 2001
Accepted for publication March 4, 2002

ABSTRACT

In general when a phylogeny is reconstructed from DNA or protein sequence data, it makes use only of the probabilities of obtaining some phylogeny *given* a collection of data. It is also possible to determine the prior probabilities of different phylogenies. This information can be of use in analyzing the biological causes for the observed divergence of sampled taxa. Unusually "rare" topologies for a given data set may be indicative of different biological forces acting. A recursive algorithm is presented that calculates the prior probabilities of a phylogeny for different allelic samples and for different phylogenies. This method is a straightforward extension of Ewens' sample distribution. The probability of obtaining each possible sample according to Ewens' distribution is further subdivided into each of the possible phylogenetic topologies. These probabilities depend not only on the identity of the alleles and on $4N\mu$ (four times the effective population size times the neutral mutation rate) but also on the phylogenetic relationships among the alleles. Illustrations of the algorithm are given to demonstrate how different phylogenies are favored under different conditions.

HOMOLOGOUS genes that are not identical in sequence are called alleles. These allelic sequences are often used as markers to provide inferences about the biology of the individuals that harbor them. In 1972 Ewens described the probability of obtaining different samples of alleles when each allele is chosen randomly from a large population. This later became known as Ewens' distribution. The probability of different samples is influenced by the population size, by the mutation rates among alleles, and by the history of the population.

In its simplest form this distribution makes the following assumptions. Let N represent the effective number of diploid, randomly mating individuals in a population of constant size. The population has discrete nonoverlapping generations. Consider samples of a locus taken randomly from these individuals and assume that an infinite allele model (Kimura 1983) holds at this locus. Let μ be the number of selectively neutral mutations per generation at this locus. If evolution has proceeded long enough for an equilibrium to be reached (the probabilities at equilibrium are designated by a $\wedge$), then the probability of sampling n_1 alleles of one sort, n_2 of another, . . . up to n_k alleles of another sort is

$$\hat{E}(n_1, n_2, \ldots, n_k) = n!\theta^{k-1}/1^{\alpha_1}\,2^{\alpha_2}\cdots n^{\alpha_n}\,\alpha_1!\alpha_2!\cdots\alpha_n!S_n(\theta),$$

where $\theta = 4N\mu$, k is the number of distinct alleles sampled (the number of different allelic types), $n = \Sigma n_i$ is the total sample size, $S_n(\theta) = (1 + \theta)(2 + \theta)\cdots(n - 1 + \theta)$, and α_i is the number of values in the set $(n_1, n_2, \ldots, n_k)$ that are equal to i (Ewens 1972). For example,

$$\begin{aligned}\hat{E}(5, 2, 2, 1) &= 10!\theta^3/1^1 2^2 3^0 4^0 5^1\ 1!2!0!0!1!\ S_{10}(\theta)\\ &= 90720\ \theta^3/(1 + \theta)(2 + \theta)\cdots(9 + \theta)\end{aligned}$$

is the probability of a sample with 5 alleles of type 1, 2 of type 2, 2 of type 3, and 1 of type 1. The number of ways that a sample can be obtained is a multinomial but the α's arise because each of the allelic types are "unlabeled." Therefore, picking 10 A_1 alleles and 5 A_2 alleles is equivalent to picking 5 A_1 alleles and 10 A_2 alleles since the subscripts are arbitrary.

Here, for simplicity of notation, I calculate the same probability but do not multiply by the number of ways that a sample can be obtained. This probability is denoted by P followed by the vector describing the alleles sampled, measures only the probability of a particular sample, but ignores the number of different ways in which this probability could be obtained. This probability can be written as

$$\hat{P}(n_1, n_2, \ldots, n_k) = \prod_{i=1}^{k}(n_i - 1)!\theta^{k-1}/S_n(\theta).$$

To calculate E it is necessary to simply multiply by the number of ways to obtain the sample, *e.g.*,

$$E(n_1, n_2, \ldots, n_k) = P(n_1, n_2, \ldots, n_k)\binom{n}{n_1\, n_2 \cdots n_k}/\prod \alpha_i!.$$

For samples that consist of just two genes these probabilities are well known and have been shown to be

$$P(2) = \frac{1}{1 + \theta}$$

[1]*Address for correspondence:* Department of Biology, LSB 533, McMaster University, 1280 Main St. W., Hamilton, Ontario, L8S 4K1, Canada. E-mail: golding@mcmaster.ca

and

$$P(1, 1) = \frac{\theta}{1 + \theta}$$

(Malécot 1969). Samples of two genes have only two possible topologies. For the above sample where both genes are identical alleles there is no evidence for any particular branching order among these two genes. The most parsimonious conclusion is that they are samples from a genealogy with a single allelic branch. For the sample with two different alleles, these two genes must be sampled from different allelic branches.

We follow standard practice and designate these genealogies using the *Newick* format of nested parentheses. In our case, however, we designate only the allelic genealogy for which we have evidence. This does not consider that genes that are the same allele might have diverged and differ elsewhere in their DNA. For the purposes of the allelic phylogeny, they are identical. Hence, in the case of a sample of two genes both with identical alleles, there is no structured genealogy, only the trivial genealogy that both genes stem from a common ancestor. If the two genes carry different alleles, then the allelic genealogy states only that they are different. Similarly for samples of three genes, the possible topologies are (3), (2, 1) and (1, (1, 1)). The first topology is a trivial one, just stating the genes are allelically identical. Note also, that with these samples, the samples correspond to just one genealogy [the sample (1, (1, 1)) represents just one topology and not the three possible rooted topologies because the alleles are unlabeled].

With samples of two or three genes no partitions within a single sample are required. However, with a sample of four genes there is a partitioning imposed by the genealogy. For samples of four genes ($n = 4$), there may be anywhere from one to four allelic types present ($k = 1, \ldots, 4$). If $k = 1$, again, there is evidence for only one possible topology, *e.g.*, (4). If $k = 2$ then the topologies are (3, 1) or (2, 2). If $k = 3$, however, there are two distinct topologies for the same sample—(2, (1, 1)) and (1, (2, 1)); the former indicates that the two unique alleles are more closely related, while the latter indicates a closer relationship between the allele sampled twice and one of the unique alleles. Thus, more than one allelic genealogy is possible for a single sample. When $k = 4$ there are again two distinct topologies possible—(1, (1, (1, 1))) and ((1, 1), (1, 1)) both from the same sample of four unique alleles. Ewens' distribution provides us the ability to determine the probability of obtaining the sample (1, 1, 1, 1) but not how it is further subdivided into these two possibilities.

A recursive algorithm to subdivide Ewens' sample probabilities into the different topologies possible and to calculate how the probability of a single arbitrary sample is distributed among these different topologies is developed below. This information can be of use in analyzing the biological causes for the observed divergence of sampled taxa. Unusually "rare" topologies for a given data set may be indicative of different biological forces acting.

RESULTS

Counts: To determine the probabilities of different samples it is necessary first to consider the number of ways that branching can occur, the number of tree topologies possible, and the number of samples possible within a tree topology.

Not all topologies are equally likely to occur. This was demonstrated by Tajima (1983) in a calculation based on the assumption that any one branch is as likely to split into two new branches as another. Given this assumption, it is possible to determine the *a priori* probabilities of different topologies. A recursive algorithm is used starting at the root of the tree. Beginning at the root, the initial branch point in the tree will create two branches subtending m_1 taxa along one lineage and m_2 taxa along another lineage. Tajima calculated the probability of obtaining a particular topology for a phylogeny as

$$P(m_1, m_2) = 2/(m - 1) \quad \text{if } m_1 \neq m_2$$
$$P(m_1, m_2) = 1/(m - 1) \quad \text{if } m_1 = m_2,$$

where $m = m_1 + m_2$ is the total number of branches in the tree. This calculation is applied repeatedly from the root of the tree until all branch points have been considered and is calculated as the product sum down to the leaves of the tree. For example, Tajima's number for three topologies is

$$(((1, 1), 1), (1, 1)), \quad T = \frac{1}{2}$$

$$((((1, 1), 1), 1), 1), \quad T = \frac{1}{3}$$

$$(((1, 1), (1, 1)), 1), \quad T = \frac{1}{6}.$$

This number is appropriate for different taxa but not for samples from within a population as is shown below.

In addition to these numbers, it is necessary to ensure that all possible topologies are considered and that all numbers of possibilities within these topologies are considered. The number of tree topologies was determined by Cavalli-Sforza and Edwards (1967). If only the topologies are considered without regard to the order of taxa at the terminal branches, then the number of possible tree topologies is

$$a_n = a_1 a_{n-1} + a_2 a_{n-2} + \cdots + a_{n-1/2} a_{n+1/2} \quad \text{if } n \text{ is odd,}$$

$$a_n = a_1 a_{n-1} + a_2 a_{n-2} + \cdots + \frac{1}{2} a_{n/2}(a_{n/2} + 1) \quad \text{if } n \text{ is even.}$$

For example, when $n = 5$, then $a_5 = 3$,

(1, (1, (1, (1, 1)))) (1, ((1, 1), (1, 1))) (((1, 1), 1), (1, 1));

and when $n = 6$, then $a_6 = 6$,

(1, (1, (1, (1, (1, 1))))) (1, (1, (1, 1), (1, 1))) (1, ((1, 1), (1, (1, 1))))

((1, 1), (1, (1, (1, 1)))) ((1, (1, 1)), (1, (1, 1))) ((1, 1), ((1, 1), (1, 1))).

These numbers grow very quickly. They are required to ensure that all possible tree topologies are considered and are used to help interpret the probabilities for each topology.

We also require the number of ways that alleles can be sampled within each branching pattern. The number of combinations of alleles in this case is given by the appropriate multinomial divided by 2^i, where i is the number of nodal points in the tree that subtend symmetric branches. For example,

$$(3, (1, 1)) \text{ has } C(3, (1, 1)) = \binom{5}{3\ 1\ 1}/2^1 = 10 \text{ possible combinations,}$$

$$(1, (3, 1)) \text{ has } C(1, (3, 1)) = \binom{5}{3\ 1\ 1}/2^0 = 20 \text{ possible combinations,}$$

$$(1, (1, (1, (1, 1)))) \text{ has } C(1, (1, (1, (1, 1)))) = \binom{5}{1\ 1\ 1\ 1\ 1}/2^1 = 60 \text{ possible combinations,}$$

$$(1, ((1, 1), (1, 1))) \text{ has } C(1, ((1, 1), (1, 1))) = \binom{5}{1\ 1\ 1\ 1\ 1}/2^3 = 15 \text{ possible combinations.}$$

Recursion: The probability of each sample, with consideration for its topological structure, can be determined by evaluating all possible origins of a sample. This methodology follows from Golding (1984). It is assumed that a new population of N diploid individuals is produced each generation by a random sampling (with replacement) of the gametes of the previous generation. Generations are distinct and nonoverlapping. The probability of obtaining particular samples of gametes is used to describe the genetic structure of such a population over time. As an example, consider the probability, $P(2)$, of obtaining a sample of two identical genes (sampled without replacement). It was shown by Malécot (1969) that a recursion relationship can be built for this probability that relates the value in generation $t + 1$ to the value in generation t. This is done by determining how the mating scheme, mutation, and other factors influence the probability. Along with this information, the probability in generation t is sufficient to describe enough of the structure of the population in generation $t + 1$ to calculate the new probability. Hence, designating probabilities in the next generation with a prime,

$$P(2)' = \frac{1}{2N} + (1 - \frac{1}{2N} - 2\mu)P(2),$$

if terms $< \mu^2$, $(1/2N)^2$, and $\mu(1/2N)$ are ignored. This is a standard formula that describes the expected homozygosity within a population and how it changes over time.

For larger samples of n genes, the multinomial term is $\frac{1}{2}n(n-1)1/2N$ (again ignoring small terms). This is a description of the fraction of the number of ways to sample n genes from a population that involve two genes having a common ancestor from among the $2N$ gametes in the entire population. For the sample (2, 1), it is possible that genes 1 and 2, 1 and 3, or 2 and 3 were samples of the same gene from the previous generation chosen twice (probability $3/2N$). But only one of these events (1 and 2) would contribute to the probability of the sample (2, 1) in the next generation and only if the sample otherwise had the pattern (1, 1). Thus,

$$P(2, 1)' = \frac{1}{2N}P(1, 1) + (1 - \frac{3}{2N} - 2\mu)P(2, 1) + \mu P(3).$$

The last term contributes to the probability of $P(2, 1)$ in the next generation via a mutation of one of three identical genes sampled [which would occur with probability $P(3)$]. This mutation then creates the appropriate sample (2, 1). There are $C(2, 1) = 3$ ways that these samples can be obtained and hence three equations like this. For a fuller explanation of these types of equations see Golding (1984).

It is quite possible to extend these results to add allelic phylogenetic structure to the samples. To illustrate this, consider a sample of five genes with allelic topology (3, (1, 1)). These five genes might have their origins as duplicates of four or fewer preexisting alleles in the sample, they might be new mutations, or they might be neither. These possibilities reflect the actions of random drift, mutation, or of neither event occurring.

For the sample (3, (1, 1)) only the three identical alleles can be duplicates (via random drift) of other alleles in the sample from the previous generation as the remaining two alleles are distinct. Because the other alleles are distinct, they cannot be an identical copy of another allele in the sample from the previous generation. The probability that all three alleles are copies of a single allele in the sample is $(1/2N)^2$ [corresponding sample probability $P(1, (1, 1))$], the probability that two of the three are copies of a single allele in the sample is approximately $\frac{1}{2}(3 \times 2)\ 1/2N$ [corresponding sample probability $P(2, (1, 1))$], and the probability that two of the five are copies of a single allele in the sample is approximately $\frac{1}{2}(5 \times 4)\ 1/2N$ (Malécot 1969). Again, many other samples might arise but they do not necessarily contribute to the probability of this sample.

The probability that either of the two unique alleles

arose via mutation from the previous generation is $2\mu(1-\mu)$ [arising from the sample (3, 2) to yield (3, (1, 1))]. The probability that both of the unique alleles arose via mutation is μ^2.

Again, for simplicity all terms with second order probabilities are ignored. These events and their probabilities do not significantly alter the first order terms (GOLDING 1984). Therefore, probabilities multiplied by terms such as μ^2 and $(1/2N)^2$ can be ignored and the above terms can be simplified to $3/2NP(2, (1, 1))$ and $2\mu P(3, 2)$.

The probability that none of these events occurred and that the sample has originated as an identical sample from the previous generation is $\sim(1 - 10/2N - 3\mu)$ (MALÉCOT 1969) and hence the complete recursion equation for $P(3, (1, 1))$ is

$$P(3, (1, 1))' = \frac{3}{2N}P(2, (1, 1)) + (1 - \frac{10}{2N} - 3\mu)P(3, (1, 1)) + 2\mu P(3, 2).$$

Similarly

$$P(1, (3, 1))' = \frac{3}{2N}P(1, (2, 1)) + (1 - \frac{10}{2N} - 3\mu)P(1, (3, 1)) + \mu P(4, 1).$$

At this point, four new probabilities have been referenced [$P(2, (1, 1))$, $P(1, (2, 1))$, $P(3, 2)$, and $P(4, 1)$] but not determined. Recursion equations for these probabilities are constructed in the same way. These, in turn, will reference further probabilities. Continuing in this way leads to a small, linear system of equations.

These equations can then be solved to show that at equilibrium

$$\hat{P}(3, (1, 1)) = \left(\frac{6 + \theta}{20 + 3\theta}\right)\left[\frac{2\theta^2}{(1 + \theta)(2 + \theta)(3 + \theta)(4 + \theta)}\right]$$

$$\hat{P}(1, (3, 1)) = \left(\frac{7 + \theta}{20 + 3\theta}\right)\left[\frac{2\theta^2}{(1 + \theta)(2 + \theta)(3 + \theta)(4 + \theta)}\right].$$

As before, to get the probabilities corresponding to Ewens' sample probabilities, the values should be multiplied by the number of possible combinations:

$$\hat{E}(3, (1, 1)) = C(3, (1, 1)) \times \hat{P}(3, (1, 1)) = \left(\frac{6 + \theta}{20 + 3\theta}\right)\left[\frac{20\theta^2}{(1 + \theta)(2 + \theta)(3 + \theta)(4 + \theta)}\right]$$

$$\hat{E}(1, (3, 1)) = C(1, (3, 1)) \times \hat{P}(1, (3, 1)) = \left(\frac{14 + 2\theta}{20 + 3\theta}\right)\left[\frac{20\theta^2}{(1 + \theta)(2 + \theta)(3 + \theta)(4 + \theta)}\right].$$

The portion in square brackets is Ewens' probability for $E(3, 1, 1)$. As required, $E(3, (1, 1)) + E(1, (3, 1)) = E(3, 1, 1)$. This also demonstrates the dependence of these particular topologies on the value of θ. As θ increases from zero, the probability of $\hat{E}(3, (1, 1))$ increases from 30 to 33%. For any value of θ this topology is nearly one-half as frequent as that of the sample $\hat{E}(1, (3, 1))$.

Slightly more complicated examples of these recursion equations are

$$P(2, (3, 1))' = \frac{3}{2N}P(2, (2, 1)) + \frac{1}{2N}P(1, (3, 1)) + (1 - \frac{15}{2N} - 5\mu)P(2, (3, 1)) + \mu P(4, 2)$$

$$P(1, ((2, 1), (1, 1)))' = \frac{1}{2N}P(1, ((1, 1), (1, 1))) + (1 - \frac{15}{2N} - 2\mu)P(1, ((2, 1), (1, 1))) + \mu P(1, (3, (1, 1))) + 2\mu P(1, ((2, 1), 2)).$$

Again here, the last term is multiplied by a factor of two since both alleles in the pattern (1, 1) on the left could have originated from a mutation. The term (2, 1) has only a single allele that could have arisen by mutation. A complete set of equations for any sample of four genes is given in the APPENDIX.

The general format of these equations is

$$P(\cdots)' = \sum_i \frac{n_i(n_i - 1)}{4N} P(\cdots, n_i - 1, \cdots) + \left(1 - \frac{n(n-1)}{4N} - \sum_i \psi(n_i)\, n_i \mu\right)P(\cdots) + \tilde{\sum}\mu P(\cdots, n_i + 1, \cdots),$$

where

$$\psi(n_i) = \begin{cases} 1 & \text{if } n_i \neq 1 \\ 0 & \text{if } n_i = 1 \end{cases}$$

and where the sum $\tilde{\Sigma}$ extends only over elements that have the pattern $(n_i, 1)$ in the tree on the left-hand side and this portion of the topology becomes $(n_i + 1)$ on the right-hand side.

At equilibrium the general solution is given by the recursive formula

$$\hat{P}(\cdots) = \frac{1}{n(n-1) + \sum_i \psi(n_i) n_i \theta} \times \left[\sum_i n_i(n_i - 1)\hat{P}(\cdots, n_i - 1, \cdots) + \theta\tilde{\sum}\hat{P}(\cdots, n_i + 1, \cdots)\right].$$

These equations cannot be generally solved in closed form. As an example of their intractability, consider an example with just six alleles sampled. The equilibrium solution is

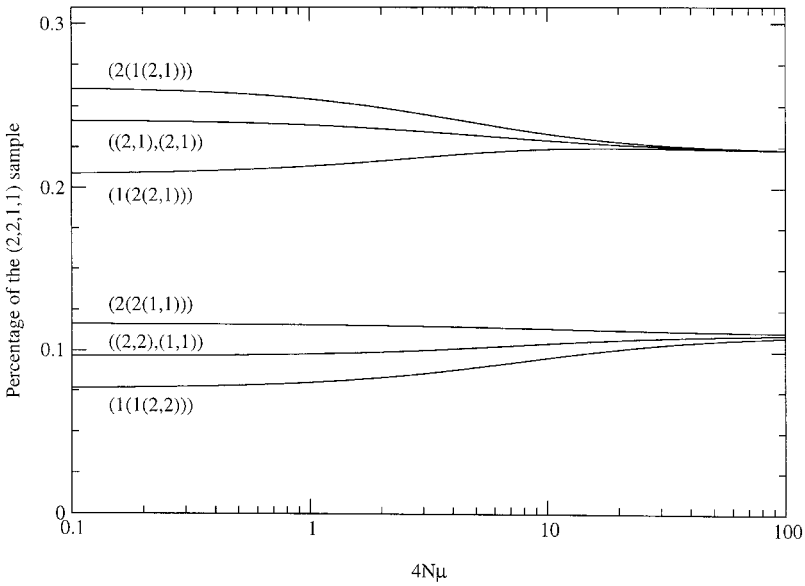

FIGURE 1.—The relative proportions for each of the possible topologies for a sample of (2, 2, 1, 1).

$$\hat{P}((1,(1,1)),(1,(1,1))) = \left(\frac{18\theta^6 + 993\theta^5 + 22{,}347\theta^4 + 263{,}462\theta^3 + 1{,}720{,}410\theta^2 + 5{,}913{,}100\theta + 8{,}376{,}000}{(15+\theta)(10+\theta)^2(20+3\theta)(6+\theta)(15+2\theta)}\right) \times \left(\frac{\theta^5}{(1+\theta)(2+\theta)(3+\theta)(4+\theta)(5+\theta)}\right).$$

Although these probabilities cannot be solved in complete generality, the form of the recursion equation above is very straightforward. It describes a simple (if large) linear matrix of equations that forms an extremely sparse system. They can be solved easily either numerically or in closed form for reasonably small sample sizes. They can also be easily calculated in nonequilibrium situations to investigate the dynamic behavior of allelic genealogies.

As $\theta \to \infty$ these probabilities are

$$\hat{P}(n_1, (n_2 \cdots, n_k)) = x\prod_{i=1}^{k}(n_i - 1)!\theta^{k-1}/S_n(\theta),$$

where

$$x = 2^i T/k!,$$

where i is the number of symmetrical nodes in the tree (ignoring the number of alleles at any one leaf or branch of the tree), and T is Tajima's number for the particular tree topology.

The relationship of these samples as $\theta \to \infty$ can be used as a rapid estimate of the probability for simulation and likelihood methods. These results show, however, that while Tajima's number is quite adequate for samples where all genes are unique (*i.e.*, $k = n$), it is not

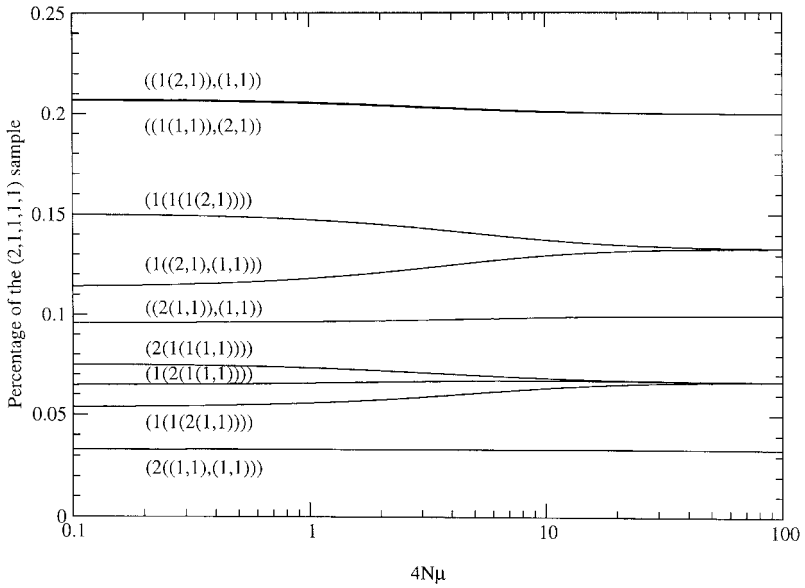

FIGURE 2.—The relative proportions for each of the possible topologies for a sample of (2, 1, 1, 1, 1).

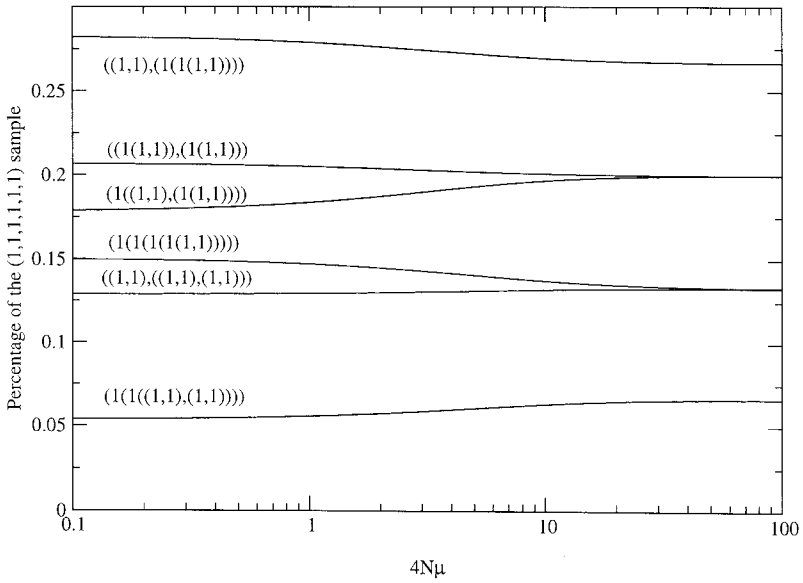

FIGURE 3.—The relative proportions for each of the possible topologies for a sample of (1, 1, 1, 1, 1, 1).

when some alleles are sampled more than once ($k < n$). In the latter case, the differences in combinatorics again become prevalent.

As $\theta \rightarrow 0$ the probabilities are also simplified and more easily calculated. For example, in the case of any genealogy with a sample of three alleles, the probabilities are

$$\hat{P}(n_1, (n_2, n_3)) = x\prod_{i=1}^{k}(n_i - 1)!\theta^{k-1}/S_n(\theta),$$

where

$$x = \frac{n_1}{n}\left(\frac{n_2}{n - n_2} + \frac{n_3}{n - n_3}\right).$$

DISCUSSION

The above methodology provides a rapid way to calculate prior probabilities of different allelic samples including differences in their genealogical relationships. This does not consider the effects of different branch lengths within a genealogy. Since the branch lengths are continuous, there are an infinite number of possible genealogies all within the same sample structure and topology. All of these possibilities are summed within the probabilities as given here and only the probability of the overall topological structure is considered. This methodology is intimately connected with the coalescent process. Individual coalescent paths are summed within the probabilities presented here to yield only those coalescents that are distinguishable via allelic differences.

One of the biggest influences on the relative probabilities of different allelic samples is simply the combinatorics of the number of ways to obtain a particular sample. The sample (1, (3, 1)) is twice as likely as (3, (1, 1)) due simply to this combinatorial difference. There is still an influence of mutation rate on the relative likelihoods of these samples; the frequency of sample (3, (1, 1)) is higher if θ is larger, but this effect is comparatively small.

The genealogical relationships among samples of (2, 2, 1, 1), (2, 1, 1, 1, 1), and (1, 1, 1, 1, 1, 1) are shown in Figures 1, 2, and 3, respectively. These are plotted from the exact numerical solution of the equilibrium recursion equations. The figures all have $n = 6$ and illustrate several features of these probabilities. First, the dependence of the probabilities on the value of θ can be almost nonexistent or, alternatively, can become very noticeable for the same overall allelic sample. In Figure 1, (2, (1, (2, 1))) changes from 0.2614 to 0.2234% of the total sample (2, 2, 1, 1) as θ changes from 0.1 to 100. But for (2, (2, (1, 1))) the relative probability changes from 0.1163 to 0.1116%. The same is true in Figure 2 for sample (1, ((2, 1), (1, 1))) (0.1137–0.1332%) *vs.* sample (2, ((1, 1), (1, 1))) (0.0330–0.0333%).

Some of the numerical relationships among genealogies are initially surprising. In Figure 2, the probabilities of samples ((1, (2, 1)), (1, 1)) and ((1, (1, 1), (2, 1))) are numerically nearly identical. For these samples, moving the "2" to the inner or outer cluster of individuals does not matter to the probability. However, moving the 2 for any of the other samples can make a large difference. The samples (1, (1, (1, (2, 1)))) and (1, (1, (2, (1, 1)))) have the same overall branching pattern (third from the top and second from the bottom, respectively). The 2 has simply been moved inside or outside of the innermost cluster of samples. But the former sample is almost three times more likely.

In Figure 3, the samples ((1, 1), ((1, 1), (1, 1))) and (1, (1, ((1, 1), (1, 1)))) (second from the bottom and bottom, respectively) have the sample allelic sample

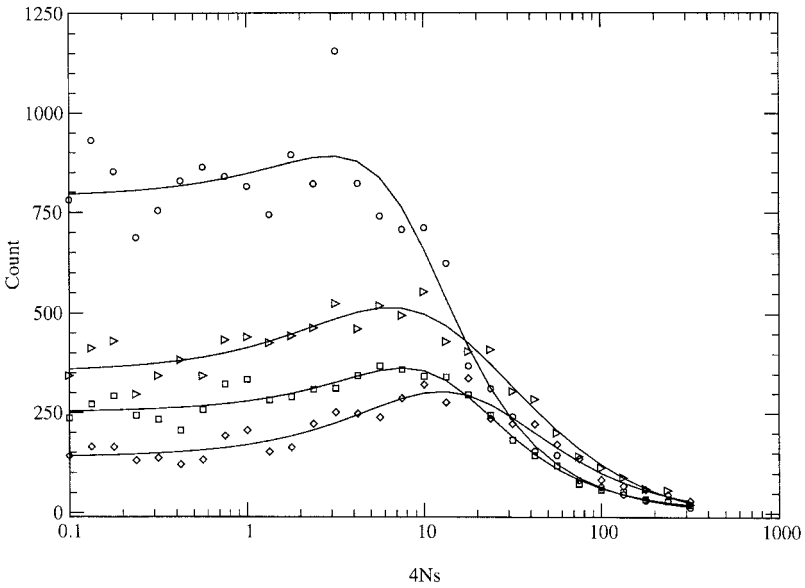

FIGURE 4.—The frequency of different genealogies changes with differing amounts of selection. Each of the four phylogenies with a sample of (*A*, *A*, *A*, *a*) is shown (where *a* is a selectively deleterious allele). Counts are based on 10,000 samples of four genes.

with the same topological structure. They differ only in the placement of the root from which these samples originated and yet this difference is sufficient to cause the former sample to be twice as likely as the latter.

Differences among genealogical structures are at the heart of many modern genetic methods. This work demonstrates that these genealogies can have very different probabilities even in the absence of external forces. This makes a knowledge of their prior probabilities a necessary precondition to the interpretation of genealogies. External forces such as migration will greatly alter these probabilities (results not shown) but without a knowledge of their values in the absence of migration this change cannot be utilized.

Changes to the probabilities of different samples can also be done via the effects of natural selection. Selection studies have shown that unlabeled genealogies do not change in length or shape as selection is present (GOLDING 1996). This unusual feature was later confirmed by NEUHAUSER and KRONE (1997) and PRZEWORSKI *et al.* (1999). However, this feature is also present only when the genealogy is unlabeled. Figure 4 shows changes in the frequency of the four genealogies that involve three distinct selectively neutral alleles and one selectively deleterious allele. These genealogies are (*a*, (*A*, (*A*, *A*))), ((*A*, *a*), (*A*, *A*)), (*A*, (*a*, (*A*, *A*))), and (*A*, (*A*, (*a*, *A*))) (top to bottom, respectively, on the left of the graph), where *a* is a selectively deleterious allele. All three samples with the same topology are now present since the alleles have been effectively "labeled" by the distinction of their selective states.

Note that when selection is weak, the three selectively neutral alleles are more likely to cluster together and form a high frequency class for this sample but when selection is strong this is the least likely topology for the same sample. This appears to be a reflection of the phenomena that selectively deleterious alleles are less likely to leave selectively neutral descendants than are neutral alleles to leave deleterious descendants (GOLDING *et al.* 1986). This graph is based on 10,000 samples of four alleles from a Monte Carlo simulation. The difference in the relative orders of these genealogies occurs only when the count of these samples is very small. Hence the utility of this difference to detect selection is probably limited.

The author thanks Dr. R. A. Morton for his helpful comments on this work. This work was supported by a Natural Sciences and Engineering Research Council of Canada grant to G.B.G.

LITERATURE CITED

CAVALLI-SFORZA, L. L., and A. W. EDWARDS, 1967 Phylogenetic analysis. Models and estimation procedures. Am. J. Hum. Genet. **19:** 233.

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

GOLDING, G. B., 1984 The sampling distribution of linkage disequilibrium. Genetics **108:** 257–274.

GOLDING, G. B., 1996 The effect of purifying selection on genealogies, pp. 271–286 in *Progress in Population Genetics and Human Evolution*, edited by P. DONNELLY and S. TAVARE, Institute for Advanced Mathematics and its Application. Springer Verlag, New York.

GOLDING, G. B., C. F. AQUADRO and C. H. LANGLEY, 1986 Sequence evolution within populations under multiple types of mutation. Proc. Natl. Acad. Sci. USA **83:** 427–433.

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

MALÉCOT, G., 1969 *The Mathematics of Heredity.* W. H. Freeman & Sons, San Francisco.

NEUHAUSER, C., and S. M. KRONE, 1997 The genealogy of samples in models with selection. Genetics **145:** 519–534.

PRZEWORSKI, M., B. CHARLESWORTH and J. WALL, 1999 Genealogies and weak purifying selection. Mol. Biol. Evol. **16:** 246–252.

TAJIMA, F., 1983 Mathematical studies on the evolutionary change of DNA sequences. Ph.D. Thesis, University of Texas, Houston.

Communicating editor: Y.-X. FU

APPENDIX

The complete set of equations that describe all possible allelic histories of samples of four genes

Recursion	$C(\cdots)$
$P(4)' = \frac{6}{2N}P(3) + (1 - \frac{6}{2N} - 4\mu)P(4)$	$1\times$
$P(3, 1)' = \frac{3}{2N}P(2, 1) + (1 - \frac{6}{2N} - 3\mu)P(3, 1) + \mu P(4)$	$4\times$
$P(2, 2)' = \frac{2}{2N}P(2, 1) + (1 - \frac{6}{2N} - 4\mu)P(2, 2)$	$3\times$
$P(2, (1, 1))' = \frac{1}{2N}P(1, (1, 1)) + (1 - \frac{6}{2N} - 2\mu)P(2, (1, 1)) + 2\mu P(2, 2)$	$6\times$
$P(1, (2, 1))' = \frac{1}{2N}P(1, (1, 1)) + (1 - \frac{6}{2N} - 2\mu)P(1, (2, 1)) + \mu P(3, 1)$	$12\times$
$P(1, (1, (1, 1)))' = (1 - \frac{6}{2N})P(1, (1, (1, 1))) + 2\mu P(1, (2, 1))$	$12\times$
$P((1, 1), (1, 1))' = (1 - \frac{6}{2N})P((1, 1), (1, 1)) + 4\mu P(2, (1, 1))$	$3\times$

Equilibrium value of $P(\cdots)$	Ewens' value: $E(\cdots) = \Sigma C(\cdots) \times P(\cdots)$
$\hat{P}(4) = 6/S_4$	$\hat{E}(4) = \hat{P}(4) = 6/S_4$
$\hat{P}(3, 1) = 2\theta/S_4$	$\hat{E}(3, 1) = 4 \times \hat{P}(3, 1) = 8\theta/S_4$
$\hat{P}(2, 2) = \theta/S_4$	$\hat{E}(2, 2) = 3 \times \hat{P}(2, 2) = 3\theta/S_4$
$\hat{P}(2, (1, 1)) = \frac{1}{3}\theta^2/S_4$	$\hat{E}(2, 1, 1) = 6 \times \hat{P}(2, (1, 1)) + 12 \times \hat{P}(1, (2, 1)) = 6\theta^2/S_4$
$\hat{P}(1, (2, 1)) = \frac{1}{3}\theta^2/S_4$	
$\hat{P}(1, (1, (1, 1))) = \frac{1}{18}\theta^3/S_4$	$\hat{E}(1, 1, 1, 1) = 12 \times \hat{P}(1, (1, (1, 1))) + 3 \times \hat{P}((1, 1), (1, 1)) = \theta^3/S_4,$
$\hat{P}((1, 1), (1, 1)) = \frac{1}{9}\theta^3/S_4$	
where $S_4 = (1 + \theta)(2 + \theta)(3 + \theta)$.	