Genealogical-tree probabilities in the infinitely-many-site model

R.C. Griffiths
Mathematics Department
Monash University
Clayton,Vic. 3168
Australia

February 3, 2009

Abstract

This paper considers the distribution of the genealogical tree of a sample of genes in the infinitely-many-site model where the relative age ordering of the mutations (nodes in the tree) is known. A computer implementation of a recursion for the probability of such trees is discussed when the nodes are age-labeled, or not.

Key words : Genealogical Trees, Infinitely-many-site model, Population Genetics

1 Introduction

Ethier and Griffiths (1987) study a reformulation of Watterson's (1975) infinitely-many-site model in population genetics as a measure-valued diffusion process. A gene is thought of as an infinitely long sequence of completely linked sites where mutations occur at sites never before mutated. Sites within a gene are labeled by elements of [0,1] and genes have a type space $E = [0,1]^{Z_+}$. A gene is of type $\mathbf{x} = (x_0, x_1, \ldots) \in E$ if $x_0, x_1, \ldots$ is the time-ordered sequence of sites at which mutations have occurred in the line of descent of that gene, where x_0 is the most recently mutated site. If the process is considered under a stationary distribution then a collection of n genes always forms a genealogical tree as defined below.

Formally let $n \in N$. Then $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in E^n$, is a *tree* if :

each sequence $\mathbf{x}_i$ has distinct coordinates for fixed $i \in \{1, \ldots, n\}$; (1.1)

if $i, i' \in \{1, \ldots, n\}$, $j, j' \in Z_+$ and $x_{i,j} = x_{i',j'}$, then $x_{i,j+k} = x_{i',j'+k}$
for $k = 0, 1, \ldots$; (1.2)

there exist $j_1, \ldots, j_n \in Z_+$ such that $x_{1,j_1} = \cdots = x_{n,j_n}$. (1.3)

Given $n \in N$, let $\mathcal{T}_n = \{(\mathbf{x}_1, \ldots, \mathbf{x}_n) \text{ is a tree}\}$. Define equivalence relations $\sim$ and $\approx$ by $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \sim (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ if $\exists$ a bijection $\zeta : [0,1] \mapsto [0,1]$ with $y_{i,j} = \zeta(x_{i,j})$ for $i = 1, \ldots, n$ and $j = 0, 1, \ldots$ and by $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \approx (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ if $\exists$ a bijection $\zeta :\mapsto [0,1]$ and a permutation σ of $(1, \ldots, n)$ such that $y_{\sigma(i),j} = \zeta(x_{i,j})$ for $i = 1, \ldots, n$

and $j = 0, 1, \ldots$. Equivalence classes in $\mathcal{T}_n/\sim$ are related to labeled trees and in $\mathcal{T}_n/\approx$ to unlabeled trees. Griffiths (1987) uses this relationship to count genealogical trees. Denote

$$(\mathcal{T}_d/\sim)_\circ = \{T \in \mathcal{T}_d/\sim \ : \ \mathbf{x}_1, \ldots, \mathbf{x}_d \text{ are distinct for all } (\mathbf{x}_1, \ldots, \mathbf{x}_d) \in T\}$$

and similarly for $(\mathcal{T}_d/\approx)_\circ$, where we do not distinguish between an equivalence class and a typical member.

Let $T \in \bigcup_d (\mathcal{T}_d/\sim)_\circ$, then T can be thought of as a labeled graph-theoretic tree. Taking the *root* of the tree as the first common coordinate of representative sequences $(\mathbf{x}_1, \ldots, \mathbf{x}_d) \in T$, then T is a rooted labeled tree whose root has at least two edges attached. The *nodes* of the tree are the distinct coordinates of $(\mathbf{x}_1, \ldots, \mathbf{x}_d)$, up to and including the root. The *leaves* of the tree are the first coordinates of the sequences. In n possibly non-distinct sequences the graph-theoretic tree is thought of as the tree constructed from distinct sequences among the n. (This is different from the usage in Griffiths (1987) where distinct nodes are appended to the beginning of each sequence.)

The basic site-type space [0,1] is not critical in the description of a labeled tree. Indeed, later in this paper, nodes will be labeled by non-negative integers according to their relative ages as mutations back in time.

Let $p(T, \mathbf{n})$ be the probability of obtaining the alleles $T \in (\mathcal{T}_d/\sim)_\circ$ at equilibrium with multiplicities $\mathbf{n} = (n_1, \ldots, n_d)$. Ethier and Griffiths (1987) derive the recursion

$$\begin{aligned} n(n-1+\theta)p(T,\mathbf{n}) = &\sum_{k:n_k\geq 2} n_k(n_k-1)p(T,\mathbf{n}-\mathbf{e}_k) \\ &+ \ \theta \sum_{\substack{k:n_k=1,\\ x_{k,0} \text{ distinct},\\ \mathcal{S}\mathbf{x}_k\neq\mathbf{x}_j \ \forall \ j}} p(\mathcal{S}_kT,\mathbf{n}) \\ &+ \ \theta \sum_{\substack{k:n_k=1,\\ x_{k,0} \text{ distinct.}}} \ \sum_{j:\mathcal{S}\mathbf{x}_k=\mathbf{x}_j} p(\mathcal{R}_kT,\mathcal{R}_k(\mathbf{n}+\mathbf{e}_j)). \end{aligned} \tag{1.4}$$

The boundary condition is $p(T_1,(1)) = 1, \ T_1 \in \mathcal{T}_1/\sim$. $\mathcal{S}$ is a shift operator which deletes the first coordinate (*i.e.*, the last mutation) of a sequence. $\mathcal{S}_kT$ deletes the first coordinate of the *kth* sequence of T. $\mathcal{R}_kT$ removes the *kth* sequence of T. θ is a scaled mutation parameter. The *degree* of a tree $(T, \mathbf{n})$ is defined as $n-1+$ the number of nodes of T and (1.4) is recursive in this degree.

Let $p^*(T, \mathbf{n})$ be the probability of a corresponding unlabeled tree in $(\mathcal{T}_d/\approx)_\circ$ with multiplicity of the sequences given by $\mathbf{n}$. p^* is related to p by a combinatorial factor. Let P_d denote the set of permutations of $(1, \ldots, d)$. Given $T \in \mathcal{T}_d/\sim$ and $\sigma \in P_d$, define $T_\sigma = \{(\mathbf{x}_{\sigma(1)}, \ldots, \mathbf{x}_{\sigma(d)}) : (\mathbf{x}_1, \ldots, \mathbf{x}_d) \in T\}$ and $\mathbf{n}_\sigma = (n_{\sigma(1)}, \ldots, n_{\sigma(d)})$. Let $a(T,\mathbf{n}) = |\{\sigma \in P_d : T_\sigma = T, \mathbf{n}_\sigma = \mathbf{n}\}|$, then

$$p^*(T,\mathbf{n}) = \frac{n!}{n_1! \ldots n_d! a(T,\mathbf{n})} p(T,\mathbf{n}).$$

In population genetics an important death process is the coalescent process (Kingman (1982)). This describes the lines of descent of a sample of n genes in the diffusion time scale. A review article is Tavaré (1984).

Let $T_n, \ldots, T_2$ be the times spent when there are $n, n-1, \ldots, 2$ ancestors of a sample of n genes. These are mutually independent, exponential random variables with rate parameters $\frac{1}{2}n(n-1),\ldots,\frac{1}{2}.2.1$. Mutations occur in a Poisson process of rate $\frac{1}{2}\theta$ along the edges of the binary splitting tree formed by the coalescent process, conditional on $T_n,\ldots,T_2$. Label these mutations by mutually independent, identically distributed uniform random variables in [0,1]. Progressing backwards in time along the coalescent tree from the genes in the sample to the common ancestor the sequences of mutations along the paths, $\mathbf{x}_1,\ldots,\mathbf{x}_n$, form a tree satisfying (1.1) – (1.3) which has a probability distribution satisfying (1.4). The recursion (1.4) is derived by using the generator in the measure-valued diffusion process. The equivalence of this and the coalescent approach is shown in Theorem 5.6 of Ethier and Griffiths (1987).

An alternative way of realizing the coalescent process with mutation is to consider both the death process and the mutation process simultaneously. Then if there are r ancestors of a sample, the waiting time until the next event backwards in time has an exponential distribution with rate parameter $\frac{1}{2}r(r+\theta-1)$. The event is a mutation, with probability $\theta/(r+\theta-1)$, or a coalescence, with probability $(r-1)/(r+\theta-1)$.

In this paper the stochastic mechanism generating trees will be supposed to be the coalescent process, and results derived from this framework.

In a sample of n genes the distribution of the configuration of the first coordinates of the sequence labels is the Ewens' sampling distribution (Ewens (1972)). Upon mutation a gene is given a first coordinate label completely new to the population, so marginally the first coordinates behave as labels of genes in the infinitely-many-alleles model. Let $\alpha(i), i=1,\ldots,n$ be the number of alleles in a sample of n genes with i representative genes and $d=\sum\alpha(i)$. Then the probability of this configuration is Ewens' sampling formula

$$\frac{n!\theta^{d-1}}{\alpha(1)!\ldots\alpha(n)!1^{\alpha(1)}\ldots n^{\alpha(n)}(1+\theta)\cdots(n-1+\theta)}\ .$$

Donnelly and Tavaré (1987) prove that if the alleles in the sample are age-ordered from oldest to youngest and η_i is the number of genes of the *ith* oldest allele in the sample, $i=1,\ldots,d$, then the probability of such a configuration is

$$\frac{n!\theta^{d-1}}{\eta_d(\eta_d+\eta_{d-1})\cdots(\eta_d+\ldots+\eta_1)(1+\theta)\cdots(n-1+\theta)}\ . \tag{1.5}$$

The number of segregating sites in a tree T is the number of nodes $-\ 1$. Watterson (1975) derives the probability generating function of the number of segregating sites in a sample of n genes as

$$H_n(z)=\prod_{j=2}^{n}\frac{j-1}{j-1+\theta-\theta z}. \tag{1.6}$$

This paper extends (1.1)–(1.3) to trees whose nodes are labeled according to their relative ages, and gives an analogous recursion to (1.4). The relative ages of mutations in a line of descent is determined by the structure of $x\ \in\ E$ even if they occur together in the same coalescent branch. Thus age-ordering mainly fixes the relative ages between divergent lines of descent.

A method of simulating an age-labeled tree and the relationship to a birth process with immigration is discussed.

A computer implementation of (1.4) and the age version (2.1) is discussed. Table 1 shows probabilities of all trees with sample sizes 2,3 and 1,2,3 nodes.

A sample can be partitioned into the ancestral allele types in its line of descent from a common ancestor. This is explored in Theorem 3 and the following material. A formula for the expected partition into these allele types is (3.5). A limit as the sample size tends to infinity provides a partition of the population into ancestral allele types. A tabulation of the expected partition of the population is shown in Table 2.

2 Age-labeled trees

An age-labeled tree is constructed from the coalescent process by labeling the mutations occurring on the edges of the binary ancestral tree by consecutive integers representing relative age ordering of the mutations. A convention will be that the root type has label 0 , and the youngest mutation backward in time has the largest label. In a sample size of n form the sequences $\mathbf{y}_1, \dots, \mathbf{y}_n$ of mutation labels from the leaves to the root. If there are d distinct sequences $\mathbf{x}_1, \dots, \mathbf{x}_d$ among the n, arrange them by the magnitude of their first coordinates, so that $\mathbf{x}_1$ represents the oldest allele, and $\mathbf{x}_d$ the youngest. The earlier tree notation will be abused slightly, by letting T denote a particular age-labeled tree, rather than an equivalence class of sequences. The ages of the nodes of trees are not included in Ethier and Griffiths' measure-valued diffusion. Ethier (1989) includes information about ages in a diffusion process, though genealogical trees are not constructed in his process.

Theorem 1. Let T be a tree represented by distinct age-labeled, ordered sequences $\mathbf{x}_1, \dots, \mathbf{x}_d$ of multiplicity $\mathbf{n} = (n_1, \dots, n_d)$ in a sample of n genes. The probability of obtaining a particular ordered sample of these sequences $p_a(T, \mathbf{n})$ satisfies the recursion

$$n(n-1+\theta)p_a(T,\mathbf{n}) = \sum_{k:n_k \geq 2} n_k(n_k-1)p_a(T,\mathbf{n}-\mathbf{e}_k) + \theta\delta_{n_d,1}p_a(T',\mathbf{n}')\ , \tag{2.1}$$

where T' is the tree formed by reordering the distinct sequences of $\mathbf{x}_1, \dots, \mathbf{x}_{d-1}, \mathcal{S}\mathbf{x}_d$ according to the age of their first coordinates and $\mathbf{n}'$ the corresponding multiplicities. The boundary condition is $p_a(T_1, (1)) = 1$. Recursion is on the degree of the trees.

The probability of obtaining an unordered sample of age-labeled sequences $\mathbf{x}_1, \dots, \mathbf{x}_d$ is

$$p_a^*(T,\mathbf{n}) = \frac{n!}{n_1! \dots n_d!} p_a(T,\mathbf{n}). \tag{2.2}$$

Proof. Consider the probability of obtaining sequences $\mathbf{x}_1, \dots, \mathbf{x}_d$ such that the first n_1 are of type $\mathbf{x}_1, \dots$, and the last n_d are of type $\mathbf{x}_d$.

Conditional on the last event being a coalescence, the probability of obtaining such a sample is

$$\sum_{k:n_k \geq 2} \frac{n_1! \dots n_d!}{n!} \frac{n_k-1}{n-1} \frac{(n-1)!}{n_1! \dots (n_k-1)! \dots n_d!} p_a(T,\mathbf{n}-\mathbf{e}_k). \tag{2.3}$$

At the instant before coalescence the $n-1$ genes are in an unordered arrangement, and the probability of an ordered arrangement is required. The factor $(n_k-1)/(n-1)$ is the probability that the parent is from the $n_k - 1$ genes with label $\mathbf{x}_k$ before coalescence.

The last event can only be a mutation if $n_d = 1$. There are two cases to consider conditional on mutation. If $\mathcal{S}\mathbf{x}_d$ is still a singleton sequence the conditional probability of obtaining the ordered arrangement is

$$\frac{1}{n} p_a(T',\mathbf{n}'), \tag{2.4a}$$

since mutation must occur on a particular gene. If $\mathcal{S}\mathbf{x}_d = \mathbf{x}_k$ the probability is

$$\frac{n_1! \dots n_d!}{n!} \frac{n_k + 1}{n} \frac{n!}{n_1! \dots (n_k + 1)! \dots n_{d-1}!} p_a(T', \mathbf{n}'). \tag{2.4b}$$

Multiplying (2.2) by $(n-1)/(n-1+\theta)$ and (2.4a) or (2.4b) by $\theta/(n-1+\theta)$ gives (2.1).

A simple case is of an age-ordered tree in a sample of 2, with paths to the root

$$\mathbf{x}_1 = (i_0, i_1, \dots, i_{k-1}, 0), \quad \mathbf{x}_2 = (j_0, j_1, \dots, j_{\ell-1}, 0).$$

The coordinates are disjoint, apart from 0, and decreasing within sequences.
Let $\mathbf{n} = (1, 1)$; then

$$p_a^*(T, \mathbf{n}) = \left(\frac{1}{2}\right)^{k+\ell-1} \left(\frac{\theta}{1+\theta}\right)^{k+\ell} \frac{1}{1+\theta}.$$

The recursion (1.4) of Griffiths and Ethier (1987) has an alternative proof as a corollary of Theorem 1 (with sequences of integers instead of elements of [0,1]).

Corollary. The recursion (1.4) for $p(T, \mathbf{n})$ holds.

Proof. Let the youngest node have a label s and the sequences be of lengths $m_1, \dots, m_d$. Denote by $\mathcal{T}$ the set of age-labeled trees

$$\{(\mathbf{y}_1, \dots, \mathbf{y}_d) : y_{i,j} = x_{i,\sigma(j)}, j = 1, \dots, m_i, \ \sigma \ \in \ P_s, \ y_{i,0} > y_{i,1} > \dots > y_{i,m_i}, i = 1, \dots, d\}.$$

Let Y be a tree constructed from $(\mathbf{y}_1, \dots, \mathbf{y}_d)$ in $\mathcal{T}$ and $p'_a(Y, \mathbf{n})$ its probability. Then $p'_a(Y, \mathbf{n})$ satisfies a similar equation to (2.1), but in the last term on the right the youngest sequence is not necessarily $\mathbf{y}_d$. Then

$$p(T, \mathbf{n}) = \sum_{(\mathbf{y}_1, \dots, \mathbf{y}_d) \ \in \ \mathcal{T}} p'_a(Y, \mathbf{n}) \ .$$

Summation in the similar equation gives (1.4). Note that in the youngest sequence $\mathbf{y}_k$, (say) $y_{k,0}$ is always a leaf node of the tree. The terms in the last two summations on the right of (1.4) are formed when the singleton sequence $\mathbf{y}_k$ is the youngest.

All possible trees T with $n = 2, 3$ and the number of nodes $1, 2, 3, 4$ are enumerated and $p^*(T, \mathbf{n})$ is calculated from (1.4) for illustrative values of θ in Table 1.

Table 1

Tree probabilities with sample sizes two and three.

				θ		
Sequences,multiplicities	0.1	0.5	1.0	1.5	2.0	2.5
(0),2	0.909091	0.666667	0.500000	0.400000	0.333333	0.285714
(0),1;(1,0),1	0.082645	0.222222	0.250000	0.240000	0.222222	0.204082
(0),1;(2,1,0),1	0.003757	0.037037	0.062500	0.072000	0.074074	0.072886
(1,0),1;(2,0),1	0.003757	0.037037	0.062500	0.072000	0.074074	0.072886
(0),1;(3,2,1,0),1	0.000171	0.006173	0.015625	0.021600	0.024691	0.026031
(1,0),1;(3,2,0),1	0.000512	0.018519	0.046875	0.064800	0.074074	0.078092
(0),3	0.865801	0.533333	0.333333	0.228571	0.166667	0.126984
(0),2;(1,0),1	0.080583	0.195556	0.194444	0.166531	0.138889	0.115898
(0),1;(1,0),2	0.039355	0.088889	0.083333	0.068571	0.055556	0.045351
(0),2;(2,1,0),1	0.003068	0.027852	0.042438	0.044362	0.041667	0.037660
(0),1;(2,1,0),2	0.001789	0.014815	0.020833	0.020571	0.018519	0.016197
(1,0),1;(2,0),2	0.004202	0.035556	0.050926	0.050939	0.046296	0.040792
(0),1;(1,0),1;(2,0);1	0.002558	0.026074	0.043210	0.047580	0.046296	0.042925
(0),1;(1,0),1;(2,1,0),1	0.001249	0.011852	0.018519	0.019592	0.018519	0.016797
(0),2;(3,2,1,0),1	0.000130	0.004326	0.009924	0.012509	0.013117	0.012759
(0),1;(3,2,1,0),2	0.000081	0.002469	0.005208	0.006171	0.006173	0.005785
(1,0),1;(3,2,0),2	0.000272	0.008395	0.017940	0.021453	0.021605	0.020353
(1,0),2;(3,2,0),1	0.000311	0.009778	0.021283	0.025791	0.026235	0.024908
(1,0),1;(2,0),1;(3,0),1	0.000041	0.001738	0.004801	0.006797	0.007716	0.007949
(0),1;(2,1,0),1;(3,0),1	0.000179	0.007190	0.019033	0.026269	0.029321	0.029846
(0),1;(2,1,0),1;(3,2,1,0),1	0.000057	0.001975	0.004630	0.005878	0.006173	0.005999
(1,0),1;(2,0),1;(3,2,0),1	0.000153	0.005531	0.013374	0.017353	0.018519	0.018219
(0),1;(1,0),1;(3,2,1,0),1	0.000020	0.000790	0.002058	0.002799	0.003086	0.003111
(0),1;(3,1,0),1;(2,1,0),1	0.000020	0.000790	0.002058	0.002799	0.003086	0.003111

As an example of an age-labeled tree T consider two ordered sequences, (y,z) of multiplicity 2 and (w,x,z) of multiplicity 1. The probabilities $p_a^*(T,\mathbf{n})$ of all the different possible age-labeled sequences

$$(1,0),(3,2,0)\ ;\ (2,0),(3,1,0)\ ;\ (3,0),(2,1,0)$$

are respectively

$$\frac{\theta^3(19\theta^2+62\theta+52)}{36(1+\theta)^4(2+\theta)^3}\ ;\frac{\theta^3(5\theta+8)}{12(1+\theta)^4(2+\theta)^2}\ ;\frac{\theta^3}{4(1+\theta)^4(2+\theta)}\ .$$

If the age ordering of the nodes is unknown, adding the above probabilities,

$$p^*(T,\mathbf{n})=\frac{\theta^3(43\theta^2+152\theta+136)}{36(1+\theta)^4(2+\theta)^3}\ .$$

It is possible to estimate θ numerically by maximum likelihood. Respective estimates, standard deviations and probabilities $\left(\hat{\theta},\mathrm{sd}(\hat{\theta}),p_a^*(T,\mathbf{n};\hat{\theta})\right)$ in this example for the three labeled trees are (1.88,1.75,0.01083), (1.82,1.69,0.00931), (1.73,1.59,0.00625) and for the combined estimate from $p^*(T,\mathbf{n})$, (1.82,1.69,0.02637).

In Table 1 most of the trees have a maximum probability with respect to θ for $\theta \in (0.1, 2.5)$.

A tree-growing simulation from Ethier and Griffiths (1987) can be adapted to age-labeled trees.

Consider a discrete-time Markov process whose state space is collections of unordered sequences of increasing integers representing age-labeled trees.
1. Begin at $\tau = 2$ with two identical sequences $\{\mathbf{x}_1 = (0),\ \mathbf{x}_2 = (0)\}$.
2. Suppose at a particular time τ the state of the process is $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ and the largest first coordinate is s. At time $\tau + 1$ make a transition by either
(i) duplicating one of the sequences

$$\{\mathbf{x}_1, \dots, \mathbf{x}_m\} \mapsto \{\mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{x}_k, \mathbf{x}_{k+1}, \dots, \mathbf{x}_m\}$$

with probability $(m-1)/m(m+\theta-1)\ \ k = 1, \dots, m$; or
(ii) adding a term to one of the sequences

$$\{\mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_m\} \mapsto \{\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, (s+1, \mathbf{x}_j), \mathbf{x}_{j+1}, \dots, \mathbf{x}_m\}$$

with probability $\theta/m(m+\theta-1)$, $j = 1, \dots, m$;
3. Stop when the number of sequences is $n+1$ for the first time and discard the last duplicated sequence. The n sequences left form an age-labeled tree. At time τ the tree grown is of degree τ.

The proof that (1) – (3) lead to the recursion (2.1) is similar to that in Ethier and Griffiths (1987), p542, but is simpler, since $a(T,\mathbf{n})$ there is replaced by unity. It is also possible to show that an event of the type 2(ii) in the simulation corresponds to an ancestral mutation while there are m ancestors of the sample of n.

Hoppe (1984,1987) considers an urn scheme for simulating an age-ordered sample in the infinitely-many-alleles model which has a similarity to (1) – (3).

Tavaré (1987) relates a linear birth process $\{Z(t),\ t \geq 0,\ Z(0) = 0\}$ with birth rate $\lambda_n = n$, and immigration rate θ , to (1.5) . If $\{N_i(t), i = 1, 2, \dots\}$ is a partition of $Z(t)$ into family sizes of immigrants arriving at times $0 < t_1 < t_2 < \dots$ then

$$P(\{N_i(t) = \eta_i\ ;\ i = 1, 2, \dots, k\} |\ Z(t) = \sum_1^k \eta_i)$$

is identical to (1.5).

The next theorem uses Tavaré's idea to give a similar interpretation to the simulation (1) – (3). Note, however, that the process is fundamentally different from Tavaré's. Immigrant alleles in the process of Theorem 2 may be destroyed, and the birth rate when there are n individuals is $\lambda_n = n - 1$, not $\lambda_n = n$. Perhaps *immigrant* is a misnomer here.

Theorem 2. Let $\{Z(t),\ t \geq 0, Z(0) = 1\}$ be the population size in a continuous time Markov birth process with birth rates $\lambda_1 = 1$ and $\lambda_n = n - 1$, $n = 2, \dots$. Immigrants arrive into the population independently as a Poisson process of rate θ but (contrary to usual) do not increase the population size.

Construct a tree from the process by assigning a root node 0 to the last immigrant before $Z(t) = 2$, then nodes $1, 2, \ldots$ to successive immigrants. A new node and edge is joined to one of the existing end edges of the tree chosen at random. A new edge is appended to its parent node when a birth occurs. A formal construction of sequences representing a tree is analogous to $2(i)$ and $2(ii)$ corresponding to birth or immigration. Let $(T(t), \mathbf{n}(t))$ be the tree constructed at time t, containing only the labeled nodes. Denote the random time $t_n = \sup\{t\ ; Z(t) \leq n\}$. Then

$$p_a^*(T, \mathbf{n}) = P(T(t_n) = T, \mathbf{n}(t_n) = \mathbf{n})\ .$$

Proof. The proof is almost immediate by noting that the transitions $2(i)$ and $2(ii)$ agree with the imbedded Markov chain from the continuous process where births and immigration occur, and that step 3 corresponds to stopping the n sequence tree at t_n .

3 Allele frequencies

Of interest is the number of genes in a sample of n of the jth oldest allele in the history of the sample. Such an allele may, or may not, be represented in the sample. Theorem 3 gives a recursion for the distribution of the number of genes of this type in a sample. A corollary considers the number of genes of the common ancestor's type in a sample. Beder (1988) and Griffiths (1986) consider similar problems.

Theorem 3. Let T be a tree represented by n age-labeled sequences and
$q_n(r, s; j) = P(\{T$ has $s + 1$ nodes and the allele of the jth node has multiplicity r in the sample.$\})$
Then

$$n(n - 1 + \theta)q_n(r, s; j) = n(r - 1)q_{n-1}(r - 1, s; j)\ +\ n(n - 1 - r)q_{n-1}(r, s; j)$$

$$+\ (n - r)\theta q_n(r, s - 1; j)\ +\ (r + 1)\theta q_n(r + 1, s - 1; j)\ +\ \delta_{j,s}\delta_{r,1}n\theta Q_n(s - 1)\ , \quad (3.1)$$

$n = 2, 3, \ldots\ ,\ s = 0, 1 \ldots\ ,\ j = 0, \ldots, s\ ,\ r = 0, \ldots, n$ and where $Q_n(s)$ is the probability that T has $s+1$ nodes, with probability generating function (1.6). The boundary condition is $q_1(r, s; j) = 1$ if $j = s = 0\ , r = 1$ and zero otherwise. Interpret $q_n(r, s; j)$ as 0 if $j > s$ or $r > n$ or $r, s < 0$. A particular case is

$$q_n(r, 0; 0) = \delta_{r,n} \prod_{j=2}^{n} \frac{j - 1}{j - 1 + \theta}\ .$$

Proof. Consider the coalescent process and whether the last event is a coalescence with probability $(n - 1)/(n - 1 + \theta)$ or a mutation with probability $\theta/(n - 1 + \theta)$.

The respective terms in (3.1) are derived by considering whether :
(i) the coalescent parent has an allele type of the jth node;
(ii) the coalescent parent does not have an allele type of the jth node;
(iii) the last mutation occurs on a gene with an allele type of the jth node;
(iv) the last mutation occurs on a gene with an allele type not of the jth node.

Let $\mu_n(s;j)$ be the probability that a sample of n genes has a tree T with $s+1$ nodes and that a randomly chosen gene is the same allele type as the jth node. Then

$$n(n-1+\theta)\mu_n(s;j) = n(n-1)\mu_{n-1}(s;j) \ + \ \theta(n-1)\mu_n(s-1;j) \ + \theta\delta_{j,s}Q_n(s-1) \ , \quad (3.2)$$

$n = 2,3,\dots \ , s = 0,1,\dots \ , j = 0,\dots,s$. The boundary condition is $\mu_1(s;j) = \delta_{j,0}\delta_{s,0}$. The expected proportion of genes which are the same allele type as the jth node, given the tree has $s+1$ nodes is $\mu_n(s;j)/Q_n(s)$. This has a similar proof to (3.1). A recursion from (1.6) (or argued directly) is

$$(n-1+\theta)Q_n(s) = \theta Q_n(s-1) + (n-1)Q_{n-1}(s), \tag{3.3}$$

$s = 0,1,\dots$ and $Q_1(s) = \delta_{s,0}$.
Together (3.2) and (3.3) allow straightforward computation of $\mu_n(s;j)$. Let

$$G_n(u,v) = \sum_{s=0}^{\infty}\sum_{j=0}^{s} u^j v^s \mu_n(s;j) \ ,$$

then from (3.2), for $H_n(u)$ defined in (1.6),

$$(n(n-1+\theta) - (n-1)\theta v)G_n(u,v) = n(n-1)G_{n-1}(u,v) + \theta uvH_n(vu) \ .$$

The solution to this is

$$G_n(u,v) = \prod_{k=2}^{n} \frac{k(k-1)}{k(k-1+\theta)-(k-1)\theta v} + \ \theta vu \sum_{r=2}^{n} \frac{H_r(vu)}{r(r-1)} \prod_{k=r}^{n} \frac{k(k-1)}{k(k-1+\theta)-(k-1)\theta v} \ . \tag{3.4}$$

A marginal generating function is $G_n(1,v) = H_n(v)$.

Let $\mu_n(j)$ be the probability that a randomly chosen gene from a sample of n genes has the same allele type as the jth node of the sample's tree T, with the interpretation that $\mu_n(j) = 0$ if T has less than j nodes. Summing over $s \geq j$ in (3.2),

$$(n(n-1)+\theta)\mu_n(j) = n(n-1)\mu_{n-1}(j) + \theta Q_n(j-1) \ ,$$

$n = 1,2\dots$ and $\mu_1(j) = \delta_{j,0}$.

Placing $v = 1$ in (3.4) provides a generating function for $\mu_n(j)$ and an explicit solution is

$$\mu_n(j) = \delta_{j,0}\prod_{k=2}^{n}\frac{k(k-1)}{k(k-1)+\theta} + (1-\delta_{j,0})\sum_{r=2}^{n}\frac{Q_r(j-1)}{r(r-1)}\prod_{k=r}^{n}\frac{k(k-1)}{k(k-1)+\theta}. \tag{3.5}$$

Of course $\sum_{j=0}^{\infty}\mu_n(j) = 1$. Let $\mu(j) = \lim_{n\to\infty}\mu_n(j)$. Then $\mu(j)$ has a similar form to (3.5) with n replaced by ∞. (The products in (3.5) are easily seen to converge.) A particular result is

$$\mu(0) = \prod_{k=2}^{\infty}\frac{k(k-1)}{k(k-1)+\theta} = \pi\theta sec\left(\frac{\pi}{2}\sqrt{1-4\theta}\right) \ ,$$

derived by Beder (1988). One might intrepret $\{\mu(j); j = 0,1,\dots\}$ as an expected partition of the entire population into ancestral types. Another interpretation is a limiting partition of the relative frequencies of types when a tree is grown forever using the simulation (1),(2) not stopping as in (3).

Since all ancestral types may not be represented in the population the expected partition is different from an expected partition of those *in* the population according to relative age, where the expected frequency of the *jth* oldest is

$$\frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^{j-1}\ ,\ j=1,2,\dots\ .$$

This expected frequency is easily obtained from the age-ordered representation of the allele frequencies in the infinitely-many-alleles model. Let $\{Z_i, i=1,2,\ldots\}$ be an *i.i.d.* sequence of random variables with density

$$\theta(1-z)^{\theta-1}\ ,\ 0<z<1.$$

Then the age-ordered frequencies are distributed as $Z_1, Z_2(1-Z_1), Z_3(1-Z_1)(1-Z_2), \ldots$ (Donnelly and Tavaré (1987)).

Table 2 shows $\{\mu(j)\ ,\ j=0,1,\ldots,10\}$ for illustrative values of θ. For θ not too large, alleles from early mutations take up most of the population frequency.

Table 2

Relative frequencies of node types.

	θ					
Node	0.1	0.5	1.0	1.5	2.0	5.0
0	0.9061	0.6261	0.4119	0.2809	0.1970	0.0334
1	0.0802	0.1886	0.1764	0.1396	0.1063	0.0213
2	0.0119	0.0996	0.1364	0.1274	0.1065	0.0257
3	0.0016	0.0481	0.0981	0.1092	0.1009	0.0299
4	0.0002	0.0218	0.0667	0.0891	0.0914	0.0337
5		0.0094	0.0434	0.0699	0.0799	0.0370
6		0.0039	0.0273	0.0531	0.0678	0.0397
7		0.0016	0.0166	0.0393	0.0561	0.0418
8		0.0006	0.0099	0.0284	0.0454	0.0433
9		0.0002	0.0058	0.0201	0.0361	0.0441
10		0.0001	0.0033	0.0140	0.0282	0.0444
>10			0.0042	0.0290	0.0844	0.6057

Corollary. Let T be a tree represented by n age-labeled sequences and
$q_n(r;j) = P(\{$The allele of the *jth* node of T has multiplicity r in the sample.$\})$
Then

$$(n(n-1)+r\theta)q_n(r;j) = n(r-1)q_{n-1}(r-1;j)\ +\ n(n-1-r)q_{n-1}(r;j)$$

$$+\ (r+1)\theta q_n(r+1;j)\ +\ \delta_{r,1}n\theta Q_n(j-1)\ , \qquad (3.6)$$

$n = 2, 3, \dots\ ,\ j = 0, 1, \dots\ ,\ r = n, \dots, 0$. The boundary condition is $q_1(r; j) = \delta_{r,1}\delta_{j,0}$. Interpret $q_n(r; j)$ as 0 $r > n$ or $r < 0$. A particular case is

$$q_n(n; j) = \delta_{j,0} \prod_{k=2}^{n} \frac{k-1}{k-1+\theta} \ .$$

Proof. Sum (3.1) over $s \geq j$.

Corollary. Let T be a tree represented by n sequences and let $q_n(r) = q_n(r; 0)$;

i.e. $q_n(r) = P$(The allele of the common ancestor has multiplicity r in the sample).

Then

$$(n(n-1)+r\theta)q_n(r) = n(r-1)q_{n-1}(r-1)+n(n-1-r)q_{n-1}(r) \ + \ (r+1)\theta q_n(r+1) \quad (3.7)$$

$n = 2, 3, \dots\ ,\ r = n, \dots, 0$. The boundary condition is $q_1(r) = \delta_{r,1}$. A particular case is

$$q_n(n) = \prod_{k=2}^{n} \frac{k-1}{k-1+\theta} \ .$$

Proof. Let $j = 0$ in (3.6). Note that the last term vanishes.

The distribution $\{q_n(r)\}$ is of particular interest. It is easy to calculate on a computer from (3.7). From Griffiths (1986) the mean is

$$n \prod_{j=2}^{n} \frac{j(j-1)}{j(j-1)+\theta} \ .$$

4 Computer implementation

The recursive equations (1.4) and (2.2) were implemented in the programming language C on 8086 and Vax computers, allowing numerical calculation of $p^*(T, \mathbf{n})$ and $p_a^*(T, \mathbf{n})$.

Suppose we wish to find the probability of a tree T. Label the nodes of T by integers, the root being 0. All trees appearing in the recursion for the probability are subtrees of T obtained by travelling up the paths from the leaves to the root. Suppose the nodes are structures containing enough information so that specifying the leaf nodes of a subtree determines that subtree. In the author's implementation, data structures are

```
typedef struct {
    int ancestor;
    int *sibs;            /* pointer to siblist */
    int sibnumber;
    int scale;
} NODE;
```

```
typedef struct {
    int *leaves;           /* pointer to leaflist */
    int *multiplicity;     /* pointer to list of multiplicities */
} TREE;
```

The function implementing the recursive equations (1.4) and (2.1) is double P(TREE *tree,double theta,int ageflag) returning $p_a(T, \mathbf{n})$ or $p(T, \mathbf{n})$ according to whether ageflag is 1 or 0. P() is written transparently to the detail of the node data structure, assuming only that the tree nodes are int labels, with the root 0.

It is possible to have either a completely recursive implementation or use a table lookup scheme by defining STORE 0 or 1 for conditional compilation. A completely recursive scheme is sufficient for a small sample size, but is generally inefficient, as the probability of the same subtree may be evaluated many times. The lookup scheme is to check whether the probability of a subtree has been calculated; if so get it from a store, if not calculate the probability and put it in the store. An index scheme is needed for the subtrees of the original tree which appear in the recursion. The author settled for a quick table lookup scheme which is reasonably memory efficient, rather than a complex index scheme using minimal memory. Each subtree has a unique index into an array of double pointers pstore[]. If there are q leaves of the subtree, with sibnumbers $m_1, \ldots, m_q$, in the original tree, then pstore[index] points to a linear ordering of a q-dimensional array of these dimensions. If the multiplicities of the sequences of the required tree are $n_1, \ldots, n_q$ then the probability for the subtree is stored in position pstore[index][entry], where entry is the offset position of $(n_1, \ldots, n_q)$. For convenience the original leaf nodes are considered as having sibnumbers equal to the multiplicity of the sequences.

Calculation of the index for a subtree is done in the following way. Let $\mathbf{x}_1, \ldots, \mathbf{x}_d$ be paths from the leaves to the root of the original tree with coordinates decreasing within sequences. Arrange the sequences so that if for any $1 \leq i, j \leq d$, $\mathbf{x}_i \subset \mathbf{x}_j$, then $i < j$. Partition the nodes, apart from the root, into disjoint sequences

$$\mathbf{y}_j = \{(x_{j,0}, \ldots, x_{j,\ell});\ x_{j,\ell} \neq 0,\ \ x_{j,\ell} \notin \bigcup_{1 \leq i < j} \mathbf{x}_i,\ x_{j,l+1} \in \bigcup_{1 \leq i < j} \mathbf{x}_i\},\ j = 1, \ldots, d\ .$$

That is, $\mathbf{y}_j$ is the portion of the path $\mathbf{x}_j$ from the leaf to the root which does not include nodes in the paths $\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}$. Let $N_0 = 1$, $N_j = \prod_{1 \leq i < j}(|\mathbf{y}_i| + 1), j = 1, \ldots d - 1$, then assign to a node $y_{j,k}$ an integer scale value, NODE.scale of

$$(|\mathbf{y}_j| - 1 - k)N_{j-1},\ k = 0, \ldots, |\mathbf{y}_j| - 1,\ j = 1, \ldots, d,$$

where $|\mathbf{y}|$ denotes the number of elements of $\mathbf{y}$. The root has scale 0. The index of a tree is calculated by adding the scale values of the maximal leafnodes within the partition blocks $\mathbf{y}_1, \ldots, \mathbf{y}_d$. With this scheme all the pointers in pstore[] may not be used, but usually its dimension $\prod_{j=1}^{d}(|\mathbf{y}_j|+1)+1$ is not large. All of pstore[] is needed if the sequences $\mathbf{x}_1, \ldots, \mathbf{x}_d$ are disjoint apart from the root. Memory is dynamically allocated to pstore[index] when a skeleton subtree with index is encountered.

To calculate the combinatorial coefficient $a(T, \mathbf{n})$ first arrange the sequences $\mathbf{x}_1, \ldots, \mathbf{x}_d$ so the multiplicities $n_1, \ldots, n_d$ are non-decreasing. Denote $\alpha(j) = |\{i; n_i = j, i = 1, \ldots, d\}|$. Let $u_1, \ldots, u_s$ be the distinct elements of $\mathbf{x}_1, \ldots, \mathbf{x}_d$ arranged in an arbitary order and define the $s \times d$ incidence matrix S by $S_{i,j} = I_{\mathbf{x}_{j(u_i)}}$, where $I_{\mathbf{x}}$ is the indicator function of $\mathbf{x}$. Let $Q_1, \ldots, Q_N,\ N = \alpha(1)! \ldots \alpha(n)!$ be the $d \times d$ permutation matrices formed by permuting $1, \ldots, d$ within consecutive blocks of length $\alpha(1), \ldots, \alpha(n)$. Then $a(T, \mathbf{n}) = |\{i\ ;\ \exists$ a $s \times s$ permutation matrix P such that $PSQ_i = S, i = 1, \ldots, N\}|$. The author used an algorithm from Neijenhuis and Wilf (1978) to run through permutations.

An example is a tree constructed from sequences and respective multiplicities (5,0),2; (6,1,0),1; (4,1,0),1; (3,2,1,0),3. (The reader is urged to sketch the tree!) The partition to calculate the scale values is $\mathbf{y}_1 = (5)$, $\mathbf{y}_2 = (6, 1)$, $\mathbf{y}_3 = (4)$, $\mathbf{y}_4 = (3, 2)$, and $\mathbf{N} = (1, 2, 6, 12)$. This produces respective scale values (0,2,12,24,6,1,4) for the seven nodes. The multinomial coefficient in $p^*(T, \mathbf{n})$ and $p^*_a(T, \mathbf{n})$ is 420 and $a(T, \mathbf{n}) = 2$, because of the similar sequences $\mathbf{x}_3$ and $\mathbf{x}_4$. If a store is used in the example P() is called 212 times, compared to 19,086 times for a completely recursive scheme.

Software with convenient input and output facility, error checking, debugging facilities, automatic calculation of NODE.scale values from an input tree, and an option of calculating $p^*(T, \mathbf{n})$, $p^*_a(T, \mathbf{n})$ is available on request from the author in source code and executable form on a 8086/8087 family computer.

5 References

Beder, B. Allelic frequencies given the sample's common ancestral type. *Theor. Pop. Biol.* **33**, 126–137. (1988).

Donnelly, P. and Tavaré, S. The ages of alleles and a coalescent. *Adv. Appl. Prob.* **18**, 1–19. (1986).

Donnelly, P. and Tavaré, S. The population genealogy of the infinitely-many neutral alleles model. *J. Math. Biol.* **25**, 381–391. (1987).

Ethier, S. N. The infinitely-many-neutral-alleles diffusion model with ages. To appear in *J. Appl. Prob.* (1989).

Ethier, S.N. and Griffiths, R. C. The infinitely-many-sites model as a measure-valued diffusion. *Ann. Prob.* **15**, 515–545. (1987).

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

Griffiths, R. C. Family trees and DNA sequences. *Proceedings of the Pacific Statistical Congress*, Francis, Manly, Lam (Editors). *Elsevier Science Publishers.* (1986).

Griffiths, R. C. Counting genealogical trees. *J. Math. Biol.* **25**, 423–431. (1987).

Hoppe, F. M. Polya-like urns and the Ewens sampling formula. *J. Math. Biol.* **20**, 91–94. (1984).

Hoppe, F. M. The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biol.* **25**, 123–157. (1987).

Kingman, J. F. C. The coalescent. *Stoch. Proc. Appl.* **13**, 235–248. (1982).

Nijenhuis, A., and Wilf, H. S. Combinatorial Algorithms, 2nd edn. *New York, Academic Press.* (1978).

Tavaré, S. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Pop. Biol.* **26**, 119–164. (1984).

Tavaré, S. The birth process with immigration, and the genealogical structure of large populations. *J. Math. Biol.***25**, 161-171. (1987).

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