COMMUN. STATIST.—STOCHASTIC MODELS, 14(1&2), 273–295 (1998)

The age of a mutation in a general coalescent tree

R.C. Griffiths [1] and Simon Tavaré [2]
Monash University and University of Southern California *

Abstract

Kimura and Ohta showed that the expected age of a neutral mutation observed to be of frequency x in a population is $-2x(1-x)^{-1}\log x$. We put this classical result in a general coalescent process context that allows questions to be asked about mutations in a sample, as well as in the population. In the general context the population size may vary back in time. Assuming an infinitely-many-sites model of mutation, we find the distribution of the number of mutant genes at a particular site in a sample; the probability that an allele at that site of a given frequency is ancestral; the distribution of the age of a mutation given its frequency in a sample, or population; and the distribution of the time to the most recent common ancestor, given the frequency of a mutation in a sample, or in the population.

Keywords. Age of a mutation, Coalescent process, Population Genetics, Samples of DNA.

*[1]Department of Mathematics, Monash University, Clayton, VIC 3168, Australia. [2] Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, CA 90089-1113.

1 Introduction

A classical result in population genetics, derived in a diffusion process setting by Kimura and Ohta (1973), is that the expected age of a neutral mutation observed to be of frequency x in a stationary population is

$$\frac{-2x}{1-x}\log x. \tag{1.1}$$

There is an implicit reversibility argument (made explicit by subsequent authors) that the age distribution back in time of a mutant allele known to have frequency x in the population is the same as the distribution of the time to extinction forward in time of the allele conditional on extinction. The history of the problem and relevant references are given in Watterson's (1996) review of Kimura's use of diffusion theory. A related result of Watterson and Guess (1977) shows the probability that an allele of frequency x in a population is the oldest is x, the same as the probability that such an allele A will eventually be fixed in the population forward in time.

Many theoretical results that can be obtained from diffusion process arguments can also be derived by using the underlying coalescent tree (Kingman 1982), and this approach will be used here. The ancestry of a sample of n genes from a classical neutral diffusion process can be described by a coalescent tree. Let $T_n, T_{n-1}, \ldots, T_2$ denote the lengths of time for which the sample of size n has $n, n-1, \ldots, 2$ distinct ancestors back in time to its most recent common ancestor. In the usual coalescent process, corresponding to a constant size population, the number of distinct ancestors $A_n(t)$ time t ago is a time-homogeneous death process with death rate μ_j from state j given by

$$\mu_j = \binom{j}{2}, \quad j = n, n-1, \ldots, 2. \tag{1.2}$$

The times T_j are therefore distributed as independent exponential random variables with rates μ_j.

This paper derives results about ages of mutations and the time to the most recent common ancestor of a sample of genes, and the population, under a general distribution for the ancestral tree. We assume that:

(A1) $T_n, \ldots, T_2$ are continuous random variables.

(A2) The ancestral tree is binary; when there are k ancestral lines, each pair has probability $\binom{k}{2}^{-1}$ of being the next pair to coalesce.

(A3) Mutations occur according to a Poisson process of rate $\theta/2$ along the edges of the tree (conditional on the edge lengths).

In this paper, we use generic time units. In the population genetics setting, these may be converted to generations by letting one time unit correspond to $2N$ generations, appropriate for the coalescent approximation of a population of current size $2N$ genes. The compound parameter θ is given by $\theta = 4Nu$, where u is the mutation rate per sequence per generation.

The motivation for considering a general tree comes from a coalescent model with variable population size, and from other processes such as birth processes generated forward in time. Specific results are derived in this paper for variable population size, particularly with exponential growth, in addition to the classical coalescent process. Reversibility arguments are not used; rather, results follow directly from the tree structure. If the results were placed in the context of a population process described by a diffusion process, it would not be reversible.

In this paper we assume the infinitely-many-sites model of mutation (Watterson, 1975): a new mutation in the population is assumed to occur at a site in an infinitely-long DNA sequence where there has never previously been a mutation. Thus in this model the number of mutations in the ancestral tree of n sample genes is the number of segregating sites in these n sequences. A mutation on an edge of the tree at a site occurs at that site in all leaves subtended by that edge, while other leaves contain the ancestral base. Note that each mutation that has arisen in the history of the sample back to its most recent common ancestor is represented in that sample. The results in this paper may be interpreted in Kimura and Ohta's setting by focusing on the alleles present at a particular segregating site: for example, the result in (1.1) can be interpreted as the mean age of a mutation, at that segregating site, of frequency x in the population.

1.1 Some new results

In this section, we record some of the new results derived later in the paper.

In Section 3, we show that in the infinitely-many-sites model the probability $q_{n,b}$ that a segregating site has b mutant bases is given by

$$q_{n,b} = \frac{(n-b-1)!(b-1)!\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}E(T_k)}{(n-1)!\sum_{k=2}^{n} kE(T_k)}, \quad 0 < b < n. \tag{1.3}$$

In Section 2, we show how the relevant expectations may be evaluated by simulation.

We show in Section 5 that for the constant population size coalescent process, the expected age of an allele observed to have b copies in a sample of n is

$$2\binom{n-1}{b}^{-1}\sum_{j=2}^{n}\binom{n-j}{b-1}\frac{n-j+1}{n(j-1)}. \tag{1.4}$$

Section 5 also provides results for the variable population size case, as well as limiting forms for the whole population. In Section 6, the expected time to the most recent common ancestor of the sample, conditional on this allele having b copies in the sample, is shown to be

$$2\Big(1-\frac{1}{n}\Big)+2\binom{n-1}{b}^{-1}\sum_{j=2}^{n}\binom{n-j}{b-1}\frac{1}{j(j-1)}. \tag{1.5}$$

Section 6 also discusses the variable population size case, and shows in particular that the expected time to the most recent common ancestor of the *population*, conditional on a mutation observed to be of frequency x in the population is

$$2-\frac{2x}{1-x}\Big(1+\frac{2-x}{1-x}\log x\Big). \tag{1.6}$$

The counterpart of Kimura and Ohta's result (1.1) for the mean age of a mutation in a general coalescent tree is

$$\frac{\frac{1}{2}\sum_{k=2}^{\infty}k(k-1)(1-x)^{k-2}E\big(S_k^2-S_{k+1}^2\big)}{\sum_{k=2}^{\infty}k(k-1)(1-x)^{k-2}E(T_k)}, \tag{1.7}$$

where $S_k = \sum_k^{\infty} T_j$. The general counterpart of Watterson and Guess' result is derived in Section 4. In particular, the probability that an allele observed to be of frequency x is the oldest in the population is found to be

$$\frac{\sum_{k=2}^{\infty} k(k-1)x^{k-2}E(T_k)}{\sum_{k=2}^{\infty} k(k-1)(1-x)^{k-2}E(T_k) + \sum_{k=2}^{\infty} k(k-1)x^{k-2}E(T_k)}; \tag{1.8}$$

this evaluates to x when the population size is constant.

We note that here, and in what follows, there is an implicit assumption of convergence being used in deriving population results from those of a sample of size n by letting $n \to \infty$. In particular, in the general setting the waiting times $\{T_j\}$ in (A1) have distributions that depend on n, and they need not be independent. In deriving results about times to common ancestors, we require that the limiting tree height be almost surely finite. For example, in trees generated from homogeneous binary birth processes of rate $\lambda_k, k = 2, 3, \ldots$, formulae for characteristics of finite trees continue to hold, but the series in formulae as $n \to \infty$ may not converge. In a coalescent process the total height of the tree is finite, but this holds in a birth process if and only if $\sum_2^{\infty} \lambda_k^{-1} < \infty$.

1.2 Ancestral classes in the coalescent

We have noted that the waiting times $\{T_j\}$ in (A1) need not be independent, and that their distribution may depend on n. The tree is thought of as being generated in two steps: first generate the ancestral times that determine edge lengths, then form the binary tree by coalescing edges. The construction can be done either backward in time by coalescence, or forward in time where coalescence corresponds to a birth, which is equally likely to occur from each existing edge. In this subsection, we review some results for topological properties of the trees that will be exploited in the sequel.

Kingman (1982) studied the distribution of ancestral classes in the coalescent process. If n sample individuals are labeled from 1 to n, then these individuals are partitioned into k equivalence classes, according to which of the k ancestors they have. There are several different ways to write the distribution of these classes depending on whether the individuals are labeled within classes, and whether classes are ordered or not. Perhaps the easiest description is when the individuals are unlabeled, and classes are ordered. Then the distribution of the partition is the same as the distribution of n balls placed uniformly at random in k cells, with no cell empty. There are

$\binom{n-1}{k-1}$ ordered arrangements with equal probability, corresponding to integer solutions of $r_1 + \cdots + r_k = n$, with $r_i \geq 1$, $i = 1, \ldots, k$. (This is of course a classical combinatorial result; see Feller (1968, Chapter 2) for the probabilistic setting.) The probability that a particular class has size b is

$$p_{n,k}(b) = \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}}, \tag{1.9}$$

since once b is fixed, there are $n - b$ balls left to arrange into $k - 1$ cells. The expected number of classes with b descendants in the sample depends on subtree arrangements in the coalescent tree.

In the limit as $n \to \infty$ the relative proportions $X_1, \ldots, X_k$ of the population that are subtended by k ancestors have a Dirichlet distribution with constant density

$$f(x_1, \ldots, x_k) = 1, 0 < x_i < 1, \; x_1 + \cdots + x_k = 1. \tag{1.10}$$

The relative frequency in a particular cell thus has a Beta density

$$(k-1)(1-x)^{k-2}, 0 < x < 1. \tag{1.11}$$

As $n \to \infty$, with $b/n \to x$,

$$p_{n,k}(b) \sim (k-1)(1-x)^{k-2}\frac{1}{n},$$

consistent with (1.11).

Arguments leading to (1.9), (1.10), and (1.11) are combinatorial and really depend only on exchangeability in the coalescent structure, and not on the distribution of ancestral times. Therefore (1.9), (1.10) and (1.11) are true under assumptions (A1) and (A2).

Another way to derive (1.9) and (1.10) is to relate the coalescent tree to a classical urn model. Identify k ancestors with different colored balls in an urn. In the coalescent tree when an ancestor line branches into two lines in forward time place an extra ball of the same colour as the parent line into the urn. Then when there are n balls in the urn the colours are distributed as individuals subtended by the k ancestors. The distribution of colored balls in the urn and the Dirichlet limit are classical results (see *e.g.* Feller 1971).

2 Variable population size

We review some aspects of the variable population size model described in Griffiths and Tavaré (1994b). In a model of growth forward in time (contraction back in time) denote the ratio of population sizes a time t back from the present by $\lambda(t)$, and let $\nu(t) = 1/\lambda(t)$. Let $\{A_n(t), t \geq 0\}$ be the death process described by (1.2), and let $\{A_n^\nu(t), t \geq 0\}$ denote the corresponding quantity in the variable population size case. Then

$$A_n^\nu(t) = A_n\left(\int_0^t \nu(u)du\right). \tag{2.1}$$

An explicit formula for the distribution of $A_n(t)$ is well known (Tavaré 1984, Griffiths 1980) for the constant population size case, and it follows from (2.1) that

$$P(A_n^\nu(t) = k) = \sum_{j=k}^{n} \rho_j(t)\frac{(-1)^{j-k}(2j-1)k_{(j-1)}n_{[j]}}{k!(j-k)!n_{(j)}}, k = 1, \ldots, n, \tag{2.2}$$

where $\rho_k(t) = \exp\left(-\binom{k}{2}\int_0^t \nu(u)du\right)$, $a_{(j)} = a(a+1)\ldots(a+j-1)$, and $a_{[j]} = a(a-1)\ldots(a-j+1)$. The mean waiting time in state j is then given by

$$E(T_j) = \int_0^\infty P(A_n^\nu(t) = j)dt\,, j = 2, \ldots, n.$$

Assuming that $\int_0^\infty \nu(u)du = \infty$, we can find the distribution of the number of distinct ancestors $A^\nu(t)$ of the whole population at time t by letting $n \to \infty$ in (2.2). We obtain

$$P(A^\nu(t) = k) = \sum_{j=k}^{\infty} \rho_j(t)\frac{(-1)^{j-k}(2j-1)k_{(j-1)}}{k!(j-k)!}, k \geq 1. \tag{2.3}$$

The distribution of coalescence times can be found easily by simulation, as follows. For $2 \leq j \leq n$, let

$$S_j = T_n + \cdots + T_j$$

denote the time taken for $A_n^\nu(\cdot)$ to reach state $j-1$, and define $S_{n+1} = 0$. The waiting times $T_n, \ldots, T_2$ form a stochastic process such that the distribution of S_j, conditional on $S_{j+1} = s\,, j = n, \ldots, 2$ has a probability density function given by

$$f_j(t,s) = \binom{j}{2}\nu(t)\exp\Big(-\binom{j}{2}\int_s^t \nu(u)du\Big),\ t > s. \tag{2.4}$$

$\{S_\ell\}$ is a (backward) Markov process starting from $S_{n+1} = 0$.

Let $\{U_\ell\}$ be a sequence of mutually independent uniform random variables and $F_j(t;s)$ be the distribution function corresponding to (2.4). Simulated times can be found by successively solving

$$1 - F_j(T_j + S_{j+1}; S_{j+1}) = U_j \tag{2.5}$$

for $j = n, n-1, \ldots, 2$. Note that

$$1 - F_j(t;s) = \exp\Big(-\binom{j}{2}\int_s^t \nu(u)du\Big),\ t > s,$$

and

$$E(S_j \mid S_{j+1} = s) = \int_s^\infty (1 - F_j(t;s))dt.$$

Exponential growth

For exponential growth, $\lambda(t) = \exp(-\beta t)$, $\beta > 0$ and (2.5) reduces to

$$T_j + S_{j+1} = \beta^{-1}\log\Big(\exp(\beta S_{j+1}) - 2\frac{\beta}{j(j-1)}\log(U_j)\Big),\ j = n, \ldots, 2. \tag{2.6}$$

Plots of simulated mean coalescence times $E(S_j), j = 2, \ldots, 25$ in a sample of $n = 100$ genes in an exponential growth model for $\beta = 0.0, 0.1, 0.5, 1.0, 2.0, 5.0$ are shown in Figure 1. (The mean coalescence times for larger values of j are essentially constant in β, and equal to the value for $\beta = 0$; these are not shown in Figure 1.) Each plot is based on 50,000 replications. Note that coalescence times decrease with increasing β, because time scaling is relative to the population size at time 0.

It is well known (cf. Slatkin and Hudson 1991) that if β is large then the phylogeny produced by an exponential growth model is *star shaped.* This can be seen mathematically as follows.

Let $\{S_j^\beta\}$ be the coalescence times in the exponential growth model, and $\{S_j\}$ in the constant size population model. From the representation (2.1) it follows that

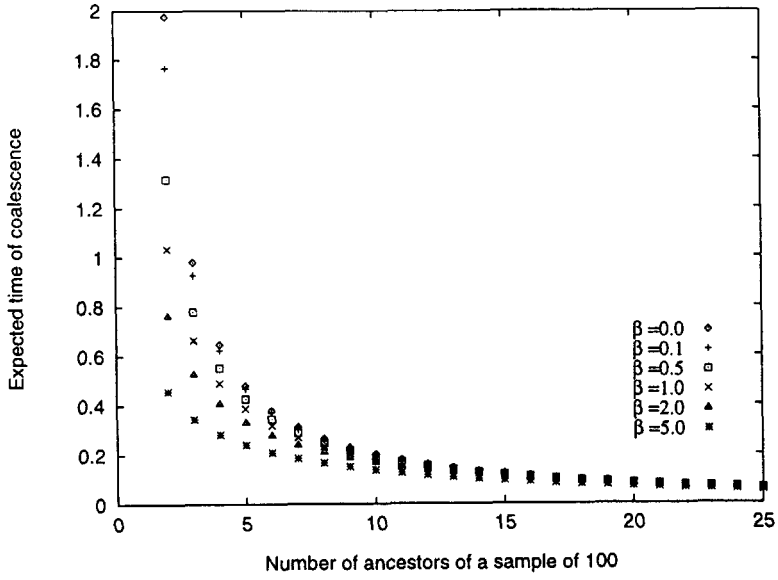

Figure 1: Expected coalescence times in a sample of 100

$$\int_0^{S_j^\beta} \nu(u)du = S_j,\ j = 2, \ldots, n,$$

and hence

$$S_j^\beta = \beta^{-1}\log(1 + \beta S_j). \tag{2.7}$$

Our interest here is in results as $\beta \to \infty$. It follows immediately from (2.7) that

$$\begin{aligned} T_n^\beta &= \beta^{-1}\log(1 + \beta T_n), \\ T_j^\beta &= \beta^{-1}\log\Big\{(1 + \beta S_j)/(1 + \beta S_{j+1})\Big\},\ j < n. \end{aligned}$$

As $\beta \to \infty$, we see that

$$\begin{aligned} &S_j^\beta \to 0 \text{ with probability } 1, j = 2, \ldots, n \\ &T_n^\beta \sim \beta^{-1}\log(\beta) + \beta^{-1}\log(T_n), \\ &T_j^\beta \sim \beta^{-1}\log\Big\{S_j/S_{j+1}\Big\},\ j < n. \end{aligned}$$

To explain the star-shaped coalescent tree that is produced as $\beta \to \infty$, we note that

$$T_j^\beta / T_n^\beta \to 0,\ j < n, \tag{2.8}$$

so that the first coalescence times in the sample dominate the tree. Note however that later coalescences will have some effect unless $\log(\beta)$ is large as well.

It is possible to show that

$$E(\log(T_n)) = \log\Big(\frac{2}{n(n-1)}\Big) - \gamma,$$

where Euler's $\gamma = 0.577216\ldots$, which gives the asymptotic formula

$$E(T_n^\beta) \sim \beta^{-1}\Big\{\log(\beta) + \log\Big(\frac{2}{n(n-1)}\Big) - \gamma\Big\}.$$

3 The number of mutant genes in a sample

The distribution of the number of mutant genes arising from a single mutation in an ancestor is of interest. We assume that the mutation is segregating in the sample, so that it arose before the most recent common ancestor of the sample. In this section we derive this distribution under the general conditions (A1), (A2) and (A3).

The simplest way to do this is via a marked Poisson process argument. Following Ethier and Griffiths (1987), we think of the DNA sequences as being represented by unit intervals. We label the locations of new mutations that arise in the sample using a sequence of independent and identically distributed random variables having the uniform distribution on (0,1). Thus for any set $M \subset (0,1)$, mutations with locations in M arise in a branch of the coalescent at rate $\theta|M|/2$, where $|M|$ is the probability of M under the uniform distribution. We note that our results apply without change when the locations of the mutations have any continuous density over (0,1); we focus on the uniform case for simplicity. Now let $\mathbf{T}$ denote the sequence of waiting times $T_2, \ldots, T_n$ in the coalescent tree of the sample. Let $C_h = C(x, b, h)$ denote the event that there is a mutation with label U in the interval $(x, x+h) \subseteq (0,1)$ that subtends b copies in the sample, and let I_k denote the event that this mutation arises when the sample has k ancestors. We have, to order $o(h)$,

$$\begin{aligned} P(C|\mathbf{T}) &= \sum_k P(C_h, I_k|\mathbf{T}) \\ &= \sum_k p_{n,k}(b) P(I_k, U \in (x, x+h)|\mathbf{T}) \\ &= \sum_k p_{n,k}(b) \left(kT_k \frac{\theta}{2} h + o(h) \right). \end{aligned}$$

Averaging over the distribution of $\mathbf{T}$ gives

$$P(C_h) \sim \frac{\theta h}{2} \sum_k k p_{n,k}(b) E(T_k), h \downarrow 0. \tag{3.1}$$

Summing (3.1) over b gives

$$P(\text{There is a mutation with label in } (x, x+h)) \sim \frac{\theta h}{2} \sum_k k E(T_k). \tag{3.2}$$

Dividing (3.1) by (3.2) and letting $h \to 0$ shows that the probability that a particular segregating site has b copies of the mutant type in the sample of n is

$$q_{n,b} = \frac{\sum_{k=2}^n k p_{n,k}(b) E(T_k)}{\sum_{k=2}^n k E(T_k)}, 0 < b < n. \tag{3.3}$$

Substituting for $p_{n,k}(b)$ from (1.9) and simplifying gives (1.3).

If $B_{n,k}$ is a random variable having the probability distribution (1.9), it is routine to show that

$$E(B_{n,k}) = \frac{n}{k}, \text{ var}(B_{n,k}) = \frac{n(k-1)(n-k)}{k^2(k+1)}. \tag{3.4}$$

The mean number μ of genes with this mutation, from (3.3) and (3.4), is therefore

$$\mu = n \frac{\sum_{k=2}^n E(T_k)}{\sum_{k=2}^n k E(T_k)}.$$

Let $W_n = \sum_{k=2}^n T_k$, be the time to the most recent common ancestor (TMRCA) of the sample, and let $G_n = \sum_{k=2}^n k T_k$ be the total edge length in the coalescent tree. The mean number of mutant genes μ can be therefore be expressed as

$$\mu = n \frac{E(W_n)}{E(G_n)}. \tag{3.5}$$

The variance, from (3.3) and (3.4), is

$$\sigma^2 = E(\text{var}(B_{n,K})) + \text{var}(n/K), \tag{3.6}$$

where the random variable K has the distribution of the number of ancestors when the mutation occurred; that is,

$$P(K = k) = kE(T_k)/E(G_n),\ k = n, \ldots, 2.$$

Constant population size

Here $E(T_k) = 2/(k(k-1))$, $k = n, \ldots, 2$, and so

$$\sum_{k=2}^{n} kp_{n,k}(b)E(T_k) = \frac{2}{b}, \tag{3.7}$$

and

$$q_{n,b} = b^{-1} / \sum_{j=1}^{n-1} j^{-1},\ b = 1, \ldots, n-1. \tag{3.8}$$

The mean and variance of the distribution in (3.8) are

$$\begin{aligned} \mu &= (n-1)/\textstyle\sum_{j=1}^{n-1} j^{-1}, \\ \sigma^2 &= n(n-1)/(2\textstyle\sum_{j=1}^{n-1} j^{-1}) - \left((n-1)/\textstyle\sum_{j=1}^{n-1} j^{-1}\right)^2. \end{aligned} \tag{3.9}$$

As $n \to \infty$,

$$\mu \sim \frac{n}{\log n} \text{ and } \sigma^2 \sim \frac{n^2}{2 \log n}.$$

Note that the quantity $\frac{\theta}{2} \sum kp_{nk}(b)E(T_k)$ is the expected number of segregating sites (and therefore mutations) that have b copies of the mutant in the sample of n. In the constant population size case, this has value θ/b, in agreement with a result of Fu (1995).

Variable population size.

The graph in Figure 2 was produced by evaluating the relative frequency $E(W_n)/E(G_n)$ for a sample of $n = 100$ by simulation, using 100,000 runs on each of 400 β grid points. The expected frequency of mutant genes decreases as β increases, and converges to n^{-1} as $\beta \to \infty$.

4 Which is the mutant gene?

Suppose that at a particular segregating site an allele is observed to have frequency a in a sample of n genes. A result of Watterson and Guess (1977)

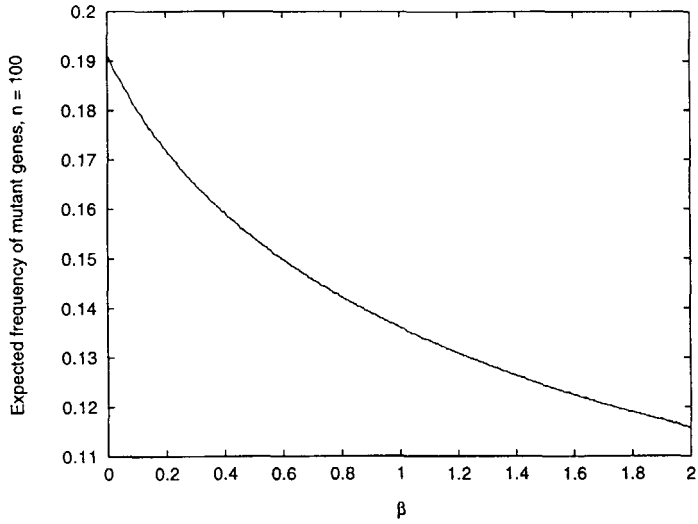

Figure 2: Expected frequency of mutant genes

can be used to show that in the constant population size model the probability that this type is ancestral is its relative frequency a/n. In other ancestral trees where the ancestral times have a different distribution this result is not necessarily true. In this section we derive a formula for the probability that an allele of frequency a is ancestral, under the conditions (A1), (A2) and (A3).

An extended version of this problem in the context of mutations occurring at sites in DNA sequences is to determine the joint distribution of ancestor bases in a sample of sequences. In the infinitely-many-sites model this is equivalent to placing a root in an unrooted tree, and determining the probability of each possible rooted tree conditional on the observed unrooted tree. This is a much harder problem than determining which is a single mutant gene, and explicit results are difficult to obtain. Griffiths and Tavaré (1994a, 1995) study unrooted and rooted genealogical trees and show how to compute the probability numerically.

Given a configuration of a copies of an allele A and $b = n - a$ copies of another allele, the conditional probability that A is the ancestral type is

$$P(a,b) = \frac{\frac{\theta}{2}\sum_{k=2}^{n} kp_{n,k}(b)E(T_k)}{\frac{\theta}{2}\sum_{k=2}^{n} kp_{n,k}(b)E(T_k) + \frac{\theta}{2}\sum_{k=2}^{n} kp_{n,k}(a)E(T_k)} = \frac{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}E(T_k)}{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}E(T_k) + \sum_{k=2}^{n} k(k-1)\binom{n-k}{a-1}E(T_k)}. \quad (4.1)$$

To derive the limiting form of (4.1), we note that as $n \to \infty$, with $x = a/n, 0 < x < 1$ fixed, $np_{n,k}(b) \to (k-1)x^{k-2}$, and $np_{n,k}(a) \to (k-1)(1-x)^{k-2}$. It follows that the limiting version of (4.1) is

$$P_\infty(x) = \frac{\sum_{k=2}^{\infty} k(k-1)x^{k-2}E(T_k)}{\sum_{k=2}^{\infty} k(k-1)(1-x)^{k-2}E(T_k) + \sum_{k=2}^{\infty} k(k-1)x^{k-2}E(T_k)}. \quad (4.2)$$

(As noted in Section 1.1, in taking the limit there is an implicit assumption of convergence of (4.1) to (4.2).) $P_\infty(x)$ is interpreted as the probability that an allele in the population with relative frequency x is the oldest. It is easy to see that $\lim_{x\to 0} P_\infty(x) = 0$, $\lim_{x\to 1} P_\infty(x) = 1$, $P_\infty(\frac{1}{2}+x) = 1 - P_\infty(\frac{1}{2}-x), 0 < x < \frac{1}{2}$.

Constant population size

In the constant population size model $P(a,b) = a/(a+b)$, from the identity (3.7), and $P_\infty(x) = x$.

Variable population size

For the model of Section 2, it is enough to assume that $\lambda(t)$ is non-increasing and continuous at $t = 0$ to ensure convergence of (4.1) to (4.2).

By way of example, consider the case of exponential population growth, where $\lambda(t) = \exp(-\beta t)$. For each fixed $x \in (1/2, 1)$, it is intuitively clear that $P_\infty(x)$ should increase as β increases. Graphs of $P_\infty(x)$ for illustrative values of β are shown in Figure 3. As β increases there is indeed an increased probability of the most frequent allele being ancestral.

5 Distribution of the age of a mutation

Let $\xi_{n,b}$ denote the age of a mutant having b copies in a sample of n genes, for $0 < b < n$. If the mutation occurred while there were k ancestors of

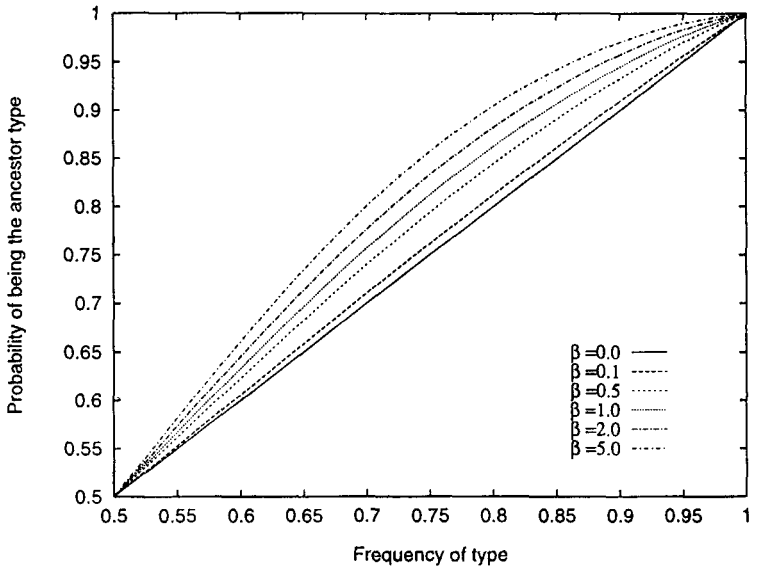

Figure 3: Probability of an allele being ancestral

the sample, then the conditional distribution of $\xi_{n,b}$ would be distributed as $UT_k + S_{k+1}$, where U is a uniform random variable on $[0,1]$, independent of $\{T_j\}$. Arguing as in the derivation of (3.1) and (3.2), the Laplace transform of $\xi_{n,b}$ is

$$\begin{aligned} E(e^{-\phi\xi_{n,b}}) &= \frac{\sum_{k=2}^{n} kp_{n,k}(b)E\big(T_k \exp(-\phi UT_k - \phi S_{k+1})\big)}{\sum_{k=2}^{n} kp_{n,k}(b)E(T_k)} \\ &= \frac{\sum_{k=2}^{n} kp_{n,k}(b)\phi^{-1}E\big((1-\exp(-\phi T_k))\exp(-\phi S_{k+1})\big)}{\sum_{k=2}^{n} kp_{n,k}(b)E(T_k)}, \end{aligned} \tag{5.1}$$

where $S_{k+1} = \sum_{j=k+1}^{n} T_j$. Let $A_n(t)$ denote the number of ancestors of the sample of n a time t ago. Then

$$P(A_n(t) = k) = P(S_{k+1} \le t) - P(S_k \le t). \tag{5.2}$$

Inverting (5.1) and using (5.2), the density of $\xi_{n,b}$ is

$$g_{n,b}(t) = \frac{\sum_{k=2}^{n} kp_{n,k}(b)P(A_n(t) = k)}{\sum_{k=2}^{n} kp_{n,k}(b)E(T_k)}$$

$$= \frac{E\left(A_n(t)(A_n(t)-1)\binom{n-A_n(t)}{b-1}\right)}{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}E(T_k)}, \; t > 0 \qquad (5.3)$$

Moments of $\xi_{n,b}$, from (5.1), (5.2), and (5.3) are

$$E(\xi_{n,b}^j) = \frac{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}\frac{1}{j+1}E\left(S_k^{j+1} - S_{k+1}^{j+1}\right)}{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}E(T_k)}, j = 1, 2 \ldots, \qquad (5.4)$$

from which the mean and variance of $\xi_{n,b}$ can be obtained.

To derive the population version of (5.4), we assume that $\{A_n(t), t \geq 0\}$ converges in distribution to a process $\{A(t), t \geq 0\}$ as $n \to \infty$, and that the time taken for $A(\cdot)$ to reach 1 is finite with probability 1. Then as $n \to \infty$, and $b/n \to x$, $0 < x < 1$, we see that

$$E(\xi_x^j) = \frac{\sum_{k=2}^{\infty} k(k-1)(1-x)^{k-2}\frac{1}{j+1}E\left(S_k^{j+1} - S_{k+1}^{j+1}\right)}{\sum_{k=2}^{\infty} k(k-1)(1-x)^{k-2}E(T_k)}, j = 1, 2 \ldots . \qquad (5.5)$$

In this population limit the density of the age of a mutant gene that has a relative frequency x is, from (5.3),

$$\begin{aligned} g_x(t) &= \frac{\sum_{k=2}^{\infty} k(k-1)(1-x)^{k-2}P(A(t)=k)}{\sum_{k=2}^{\infty} k(k-1)(1-x)^{k-2}E(T_k)} \\ &= \frac{E\left(A(t)(A(t)-1)(1-x)^{A(t)-2}\right)}{\sum_{k=2}^{\infty} k(k-1)(1-x)^{k-2}E(T_k)}. \end{aligned} \qquad (5.6)$$

The numerator in (5.6) is the second derivative of the probability generating function of $A(t)$ with argument $1-x$.

It is possible to find an empirical simulated density in (5.3) or (5.6) by estimating the expectation in the numerator by the mean of $A_n(t)(A_n(t)-1)\binom{n-A_n(t)}{b-1}$ or $A(t)(A(t)-1)(1-x)^{A(t)-2}$, with a large number of runs on a grid of time points.

Constant population size

From (5.4), we see that

$$E(\xi_{n,b}) = 2\binom{n-1}{b}^{-1}\sum_{j=2}^{n}\binom{n-j}{b-1}\frac{n-j+1}{n(j-1)}, \qquad (5.7)$$

with an asymptotic form as $n \to \infty$, $b/n \to x$ of

$$E(\xi_x) = \frac{-2x}{1-x} \log x. \tag{5.8}$$

Equation (5.8) is the well known formula (1.1) derived by Kimura and Ohta (1973). The density (5.6) is also known in various forms (Watterson, 1977; Tavaré, 1984). In this case the distribution function is identical to

$$P(\xi_x \leq t) = E(1-x)^{A(t)-1}, \; t > 0. \tag{5.9}$$

Variable population size

In the variable population size case, define Z_x by

$$\xi_x = \int_0^{Z_x} \nu(u)du.$$

Analogous to (5.9), we then have

$$P(Z_x \leq z) = E(1-x)^{A(z)-1}, \; z > 0. \tag{5.10}$$

Figure 4 graphs $E(\xi_x)$ as a function of x in the exponential growth model for illustrative values of β. The expected age of a mutation decreases in β for fixed x. This is consistent with the time scale of the tree being shortened as β increases.

6 TMRCA in a sample

Let $\eta_{n,b}$ denote a random variable that is distributed as the conditional distribution of W_n given that a sample of n genes contains b genes of a mutant type and $n-b$ ancestral genes. Clearly the density of $\eta_{n,b}$ is

$$g_{n,b}(t) = f_n(t) \frac{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1} E(T_k \mid W_n = t)}{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1} E(T_k)}, t > 0, \tag{6.1}$$

where $f_n(t)$ is the density of W_n. The mean is

$$E(\eta_{n,b}) = \frac{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1} E(T_k W_n)}{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1} E(T_k)}. \tag{6.2}$$

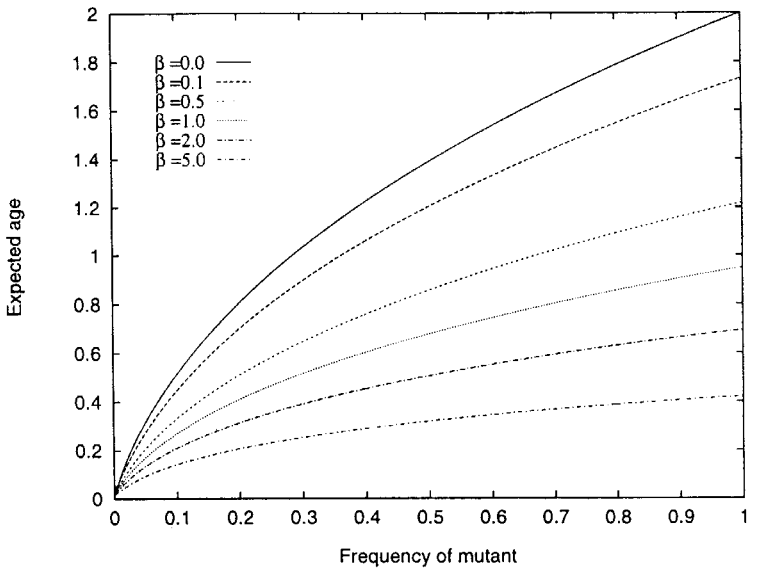

Figure 4: Expected age of a mutant gene

Constant population size

The density in equation (6.1) can be expressed as

$$g_{n,b} = f_n * z_{n,b}\,, \tag{6.3}$$

where $*$ denotes convolution, and

$$z_{n,b}(t) = \frac{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}\exp\left(-\binom{k}{2}t\right)}{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}E(T_k)}, t > 0. \tag{6.4}$$

The representation (6.3) follows from the fact that

$$\begin{aligned}
&\int_0^\infty e^{-\phi t} f_n(t) E(T_k \mid W_n = t)dt \\
&= E(T_k e^{-\phi W_n}) \\
&= \frac{2}{k(k-1)}\Big(1 + \frac{2\phi}{k(k-1)}\Big)^{-2} \prod_{i \neq k}\Big(1 + \frac{2\phi}{i(i-1)}\Big)^{-1} \\
&= \frac{2}{k(k-1)}\Big(1 + \frac{2\phi}{k(k-1)}\Big)^{-1} E(e^{-\phi W_n}),
\end{aligned}$$

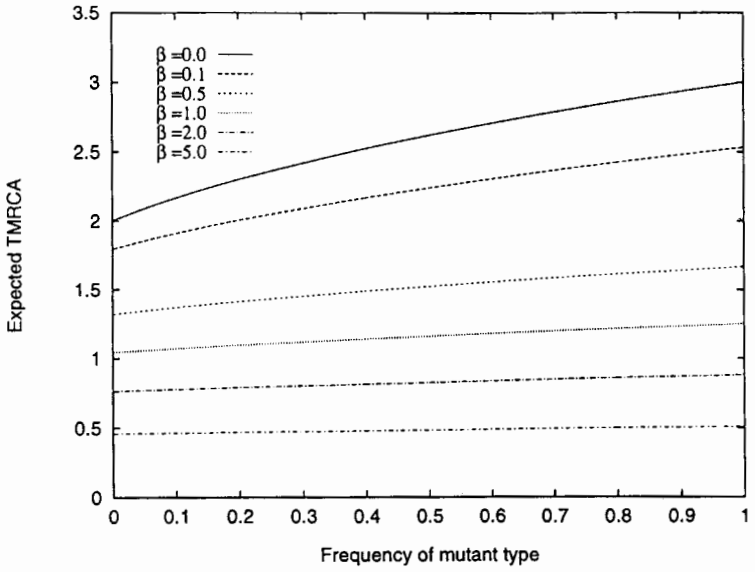

Figure 5: Expected TMRCA

and substituting in the numerator of (6.1).

The mean is

$$E(\eta_{n,b}) = 2\Big(1 - \frac{1}{n}\Big) + 2\binom{n-1}{b}^{-1} \sum_{j=2}^{n} \binom{n-j}{b-1} \frac{1}{j(j-1)}. \tag{6.5}$$

As $n \to \infty$, $b/n \to x$ the density has a similar form to (6.3), (6.4), with $\binom{n-k}{b-1}$ replaced by $(1-x)^{k-2}$ and the mean simplifies to

$$E(\eta_{\infty,x}) = 2 - \frac{2x}{1-x}\Big(1 + \frac{2-x}{1-x}\log x\Big). \tag{6.6}$$

$E(\eta_{\infty,x})$ is monotonic increasing in x and $E(\eta_{\infty,0}) = 2$, $E(\eta_{\infty,1}) = 3$.

Variable population size

The expected TMRCA decreases with β, and for large β the frequency of the mutant type has little influence on the TMRCA, as shown in Figure 5.

The joint density of W_n and $\xi_{n,b}$ is

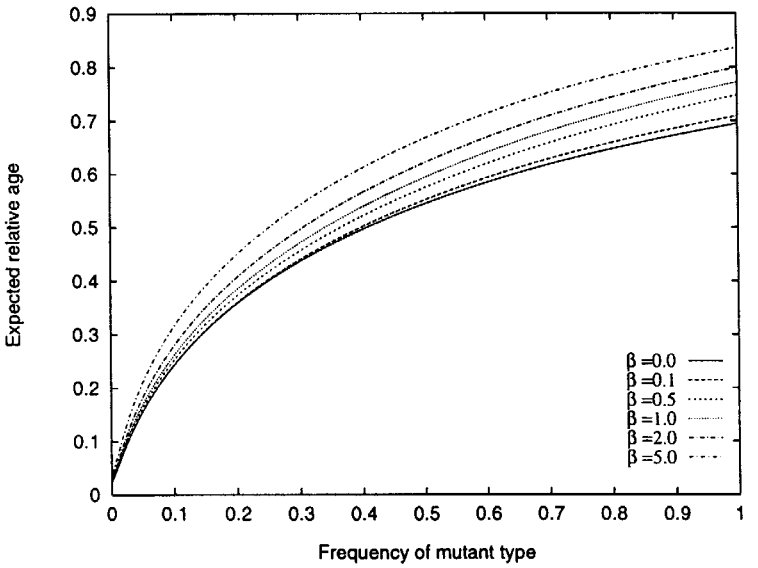

Figure 6: Relative age of a mutation

$$h_{n,b}(s,t) = f_n(s)\frac{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}P(A_n(t)=k \mid W_n=s)}{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}E(T_k)}, s > t > 0. \tag{6.7}$$

This is of interest because, for example, it allows a comparison of the age of a mutation with W_n by calculating

$$E(\xi_{n,b}/W_n) = \frac{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}\frac{1}{2}E\left((S_k^2 - S_{k+1}^2)/W_n\right)}{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}E(T_k)}. \tag{6.8}$$

Limit results for (6.1), (6.2), (6.7) and (6.8) as $n \to \infty$, while $b/n \to x$ are obtained by replacing $\binom{n-k}{b-1}$ by $(1-x)^{k-2}$.

Variable population size

Comparing the plots of $E(\xi_x/W_\infty)$ in Figure 6 with Figure 4, we see that the relative age of a mutation increases as a function of β, rather than decreases, as in Figure 4.

7 TMRCA in the population

Let $\zeta_{n,b}$ denote a random variable that is distributed as the conditional distribution of W_∞, the TMRCA in the population, given that a sample of n genes contains b genes of a mutant type and $n-b$ wild type genes.

Let $\{\alpha_{n,k}(\ell)\}$ be the probability distribution of the number of lines ℓ subtended in a sample of n, from a time back where there are k lines in the population. The argument to find the distribution of the TMRCA of the population $\zeta_{n,b}$, given a base frequency of b in the sample, is similar to that used to find the age distribution at a site. The Laplace transform is

$$E(e^{-\phi\zeta_{n,b}}) = \frac{\sum_{k=2}^{\infty} E\big(T_k \exp(-\phi W_\infty)\big) \sum_{\ell=1}^{k} \alpha_{n,k}(\ell)\ell p_{n,\ell}(b)}{\sum_{k=2}^{n} k p_{n,k}(b) E(T_k)}, \tag{7.1}$$

derived by considering that a mutation occurs on an edge of the sample coalescent tree while k ancestors of the population are in common with ℓ ancestors of the sample. It is possible to show using the methods in Griffiths (1980), or Saunders, Tavaré and Watterson (1984), that

$$\alpha_{n,k}(\ell) = \frac{k!(k-1)!n!(n-1)!}{(n-\ell)!\ell!(\ell-1)!(k-\ell)!(k+n-1)!}.$$

Constant population size

The mean time in the constant population size case is

$$2b \sum_{k=2}^{\infty} (k^{-1}(k-1)^{-1} + k^{-2}(k-1)^{-2}) \sum_{\ell=1}^{k} \alpha_{n,k}(\ell)\ell p_{n,\ell}(b)\,. \tag{7.2}$$

8 Acknowledgements

This research was supported by an Australian Research Council Grant (RCG) and National Science Foundation grant BIR 95-04393 (ST). We thank the editors and referees for comments that improved the presentation of the results.

9 References

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

Feller, W. (1968) *An introduction to probability theory and its applications.* Vol 1, 3rd ed. Wiley, New York.

Feller, W. (1971) *An introduction to probability theory and its applications.* Vol 2, 2nd ed. Wiley, New York.

Fu, Y.-X. (1995) Statistical properties of segregating sites. *Theoret. Popul. Biol.* 48, 172–197.

Griffiths, R. C. (1980) Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoret. Popul. Biol.* 17, 37–50.

Griffiths, R. C. and Tavaré, S. (1994a) Ancestral inference in population genetics. *Statistical Science* 9, 307–319.

Griffiths, R. C. and Tavaré, S. (1994b) Sampling theory for neutral alleles in a varying environment. *Proc. R. Soc. Lond. B* 344, 403–410.

Griffiths, R. C. and Tavaré, S. (1995) Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosci.* 127, 77–98.

Kimura, M., and Ohta, T. (1973) The age of a neutral mutant persisting in a finite population, *Genetics* 75, 199–212.

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

Saunders, I. W., Tavaré, S. and Watterson G. A. (1984) On the genealogy of nested subsamples from a haploid population. *Adv. Appl. Prob.* 16, 471–491.

Slatkin, M. W. and Hudson, R. R. (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129, 555–562.

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

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

Watterson, G. A. (1977) Reversibility and the age of an allele II. Two-allele models, with selection and mutation. *Theoret. Popul. Biol.* 12, 179-196.

Watterson, G. A. (1996) Motoo Kimura's use of diffusion theory in population genetics. *Theoret. Popul. Biol.* 49, 154–188.

Watterson, G. A. and Guess H. A. (1977) Is the most frequent allele the oldest? *Theoret. Popul. Biol.* 11, 141–160.