Sampling theory for neutral alleles in a varying environment

R. C. GRIFFITHS[1] AND SIMON TAVARÉ[2]

[1] *Department of Mathematics, Monash University, Clayton, Victoria 3168, Australia*
[2] *Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, California 90089–1113, U.S.A.*

SUMMARY

We develop a sampling theory for genes sampled from a population evolving with deterministically varying size. We use a coalescent approach to provide recursions for the probabilities of particular sample configurations, and describe a Monte Carlo method by which the solutions to such recursions can be approximated. We focus on infinite-alleles, infinite-sites and finite-sites models. This approach may be used to find maximum likelihood estimates of parameters of genetic interest, and to test hypotheses about the varying environment. The methods are illustrated with data from the mitochondrial control region sampled from a North American Indian tribe.

1. INTRODUCTION

Much of the inferential machinery for stochastic models in population genetics has been developed under the assumption of approximately constant population size. Perhaps the best known is the Ewens sampling formula (Ewens 1972), which gives the stationary distribution of the allelic partition of a sample evolving under the infinite-alleles assumption. Recent advances in DNA sequencing technology have markedly increased the availability of molecular data, particularly from human populations. In this context, the assumption of constant population size is often unwarranted.

There are several approaches to varying population size in the literature. Models based on branching processes with infinite-alleles mutation structure are discussed by Griffiths & Pakes (1988) and Taib (1992). In the more classical population genetics setting, there is an extensive literature on the varying environments models, those with deterministic variation in population size. The effects of bottlenecks are a popular focus. References include Nei *et al.* (1975), Chakraborty & Nei (1977), and Watterson (1984, 1989). Chakraborty (1977) and, more recently, Slatkin & Hudson (1991), Rogers & Harpending (1992) and Marjoram & Donnelly (1994*a*,*b*) discuss the infinite-sites model. A discussion of random environments models, those in which the variation in population size is random through time, appears in Donnelly (1986). Pollak (1984) studies the infinite-alleles process.

In this article we develop a sampling theory for populations evolving in a varying environment. We use a coalescent approach to provide recursions for the probabilities of particular sample configurations, and describe a Monte Carlo method by which the solutions to such recursions can be approximated. Applications of our method include maximum likelihood estimation of parameters of genetic interest, and a framework within which hypotheses about the varying environment might be tested.

In §2 we discuss the structure of the coalescent (Kingman 1982*a*) in a deterministically varying environment. Section 3 derives the appropriate sampling equations for the infinite-alleles and infinite-sites processes, and §4 outlines our Monte Carlo likelihood approach. In §5, we apply the methods to Ward *et al.*'s (1991) sample of mitochondrial DNA from a group of native North American Indians. In §6 we discuss other application of these methods, and describe computer software that is available.

2. THE COALESCENT IN A VARYING ENVIRONMENT

Imagine a haploid population evolving according to the Wright-Fisher model in a deterministically varying environment. Think of time going back into the past from now, which is labeled generation 0. Suppose there are $M_0 = N$ individuals now, M_1 in generation 1, M_2 in generation 2, and so on. The reproductive mechanism is equivalent to each of the M_r individuals in generation r choosing their parents uniformly and at random from the M_{r+1} individuals in generation $r+1$, independently of the choices in generations $0, 1, \ldots, r-1$. Define the relative size function ν_N by

$$\begin{aligned} \nu_N(x) &= \frac{M_{\lfloor Nx \rfloor}}{N} \\ &= \frac{M_t}{N}, \qquad \frac{t}{N} \leqslant x \frac{t+1}{N}, \qquad t = 0, 1, \ldots . \end{aligned} \tag{1}$$

We are interested in the behaviour of the process when the size of each generation is large, and we shall suppose that

$$\lim_{N\to\infty} \nu_N(x) = \nu(x), \tag{2}$$

exists and is strictly positive for all $x > 0$. By way of example, if $M_r = (1-s)^r M_0$, we have a geometric decrease of population size going back into the past (corresponding of course to geometric growth in the usual direction of time). If $s = \beta/N$, and time is measured in units of N generations then

$$\lim_{N\to\infty} \nu_N(x) = \lim_{N\to\infty}\left(1-\frac{\beta}{N}\right)^{\lfloor Nx\rfloor} = \mathrm{e}^{-\beta x} \equiv \nu(x),\ x > 0.$$

In the case of approximately constant population size, we have $\nu(x) \equiv 1$ for all x, and for a bottleneck occurring at time t_0, we might take $\nu(x) = 1$, $0 \leqslant x < t_0; = \beta, x \geqslant t_0$.

From now on we study the evolution of the genetic structure of a sample of individuals having population size function ν, arising as the limit as $N \to \infty$ of the Wright-Fisher model with time measured in units of N generations. The same approximating process can be used when the offspring distributions in each generation are exchangeable (Kingman 1982*b*). Suppose that in generation r the number of offspring born to a typical individual has variance $\sigma^2(r)$, and define $\tau_N^2(x) = \sigma^2(\lfloor Nx\rfloor)$. If time is measured in units of N generations, and we suppose that $\lim_{N\to\infty}\tau_N^2 (= \tau^2(x)$, then $\nu(x)$ is replaced by $\nu(x)\tau^{-2}(x)$. There are cases in which variable population size processes are better studied in their original, discrete timescale, particularly those which have very small population sizes for many generations. Although we do not explicitly examine such cases in this paper, the methods developed here can be exploited in that setting too.

It is convenient to define the population-size intensity function Λ by

$$\Lambda(x) = \int_0^x \frac{\mathrm{d}t}{\nu(t)}, \qquad x > 0. \tag{3}$$

We assume that $\Lambda(\infty) = \infty$, so that each pair of individuals, and the sample, may be traced back to a common ancestor with probability one. The density λ of Λ is given by

$$\lambda(x) = \frac{1}{\nu(x)}, \qquad x > 0. \tag{4}$$

The structure of the coalescent in a varying environment may be described as follows. Consider a sample of n individuals taken at time 0, and let $\{A_n(t), t \geqslant 0\}$ be the stochastic process that gives the number of distinct ancestors of the sample a time t in the past. $A_n(\cdot)$ is a non-homogeneous Markov death process that starts from $A_n(0) = n$, and moves down in steps of 1 until reaching 1, at which point the sample has been traced back to a common ancestor. The transition probabilities of the process are determined by

$$\mathbb{P}(A_n(t+h) = j | A_n(t) = i) = \begin{cases} \binom{i}{2}\lambda(t)h + o(h), & j = i-1 \\ 1 - \binom{i}{2}\lambda(t)h + o(h), & j = i \\ 0, & \text{otherwise.}\end{cases} \tag{5}$$

Let $T_n, T_{n-1}, \ldots, T_2$ be the lengths of time the ancestral process spends in states $n, n-1, \ldots, 2$ respectively. The joint density of $(T_n, \ldots, T_2)$ is

$$g(t_n, \ldots t_2) = \prod_{j=2}^{n} \binom{j}{2}\lambda(s_j)\exp\left\{-\binom{j}{2}(\Lambda(s_j) - \Lambda(s_{j+1}))\right\},$$

for $0 \leqslant t_n, \ldots, t_2 < \infty$, where $s_{n+1} = 0$, $s_n = t_n, s_j = t_j + \ldots + t_n$, $j = 2, \ldots, n-1$.

The coalescent may be thought of as a tree with nodes at the common ancestors of the sample, and j branches of length T_j, $j = 2, 3, \ldots, n$. Mutations are superimposed on the coalescent tree by supposing that they occur independently on each branch according to Poisson processes of rate $\theta/2$, where $\theta = \lim_{N\to\infty} 2Nu$ and u is the probability of a mutation in a given gene in a given generation.

Our model can also be described as one in which the population size remains constant over time and mutations arise according to a non-homogeneous Poisson process. The formulation we have chosen seems more natural, as there is little evidence that substitution rates change dramatically over time.

The effects of mutations may be modeled in many different ways, depending on the type of data at hand. We concern ourselves with the infinite-alleles and infinite-sites models, and mention briefly in the discussion how the methods developed here might be modified to account for more complex DNA sequence data.

3. SAMPLING EQUATIONS

In this section, we derive some integral equations satisfied by the sampling probabilities of the infinite-alleles and infinite-sites models. Viewed as functions of unknown population parameters, these can be thought of as likelihoods. One aim is to develop a method for solving such integral equations, and therefore of approximating likelihoods.

(*a*) *Infinite-alleles models*

Under the infinite-alleles mutation scheme, each mutation produces a type that has never before been seen in the population. This model is often used to describe data in which differences between alleles can be detected, but more specific details of each allele are not available. Allozyme frequency data provide an example. One consequence of this assumption is that a sample of size n may be represented as a configuration $\boldsymbol{a} = (a_1, \ldots, a_n)$, where

a_i = number of alleles represented i times

and $|\boldsymbol{a}| \equiv a_1 + 2a_2 + \ldots + na_n = n$. It is convenient to think of the configuration $\boldsymbol{b}$ of samples of size $j < n$ as being an n-vector with coordinates $(b_1, b_2, \ldots, b_j, 0, \ldots, 0)$, and we assume this in the remainder of this section. We also define $\boldsymbol{e}_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$, the ith unit vector.

We derive an equation satisfied by the sampling probabilities $q(t, \boldsymbol{a}), |\boldsymbol{a}| > 1$ defined by

$$q(t,\boldsymbol{a}) = \mathbb{P}(\text{sample of size } |\boldsymbol{a}| \text{ taken at time } t \text{ has configuration } \boldsymbol{a}), \quad (6)$$

with $q(t,\boldsymbol{e}_1) = 1$ for all $t \geqslant 0$. Suppose then that the configuration at time t is $\boldsymbol{a}$. In common with most coalescent arguments, we look at the configuration of the sample at the first event before time t. Here, the 'first event' is one that changes the configuration of the sample. If the event prior to t was a coalescence, then necessarily the configuration changed. If the event prior to t was a mutation, the configuration changes only if the mutation does not occur to one of the a_1 individuals who are singleton types (for then $a_1 \to a_1$). Using independence of the events involved, we see that the time S_t of the first event prior to t satisfies, for $s > t$,

$$\begin{aligned}
&\mathbb{P}(S_t > s)\\
&= \mathbb{P}(\text{no event that changes the configuration } \boldsymbol{a} \text{ in time } (t,s)),\\
&= \exp\left(-\int_t^s \binom{n}{2}\lambda(u)\mathrm{d}u\right)\exp\left(-\frac{n\theta}{2}\left(1-\frac{a_1}{n}\right)(s-t)\right),\\
&\equiv \exp\left(-\int_t^s \gamma(u,\boldsymbol{a})\mathrm{d}u\right),
\end{aligned}$$

where

$$\gamma(u,\boldsymbol{a}) = \binom{n}{2}\lambda(u) + \frac{n\theta}{2}\left(1-\frac{a_1}{n}\right). \quad (7)$$

Given that the first event occurs at time s, the probability that the event is a coalescence is

$$\binom{n}{2}\lambda(s)/\gamma(s,\boldsymbol{a}),$$

and the probability of a mutation is

$$\frac{n\theta}{2}\left(1-\frac{a_1}{n}\right)/\gamma(s,\boldsymbol{a}).$$

It remains to record the effects of each of these possibilities. If the coalescence occurred, then at time s the sample must have had configuration $\boldsymbol{a} + \boldsymbol{e}_j - \boldsymbol{e}_{j+1}$, and an individual in one of $a_j + 1$ allelic classes of size j had an offspring, reducing the number of j classes to a_j, and increasing the number of $(j+1)$ classes to a_{j+1}. This event has probability $j(a_j+1)/(n-1)$, $j = 1, \ldots, n-1$. If a mutation occurred, the configuration could have been $\boldsymbol{a} - 2\boldsymbol{e}_1 + \boldsymbol{e}_2$, and a mutation occurred to an individual in the 2 class (probability $2(a_2+1)/(n-a_1)$), or $\boldsymbol{a} - \boldsymbol{e}_1 - \boldsymbol{e}_{j-1} + \boldsymbol{e}_j$ and the mutation occurred to an individual in a j class, producing a singleton mutant and a new $(j-1)$ class (probability $j(a_j+1)/(n-a_1)$). Combining these possibilities, we define

$$\begin{aligned}
\mathcal{L}q(s,\boldsymbol{a}) = {}& \frac{n\theta}{2\gamma(s,\boldsymbol{a})}\Bigg[\frac{2(a_2+1)}{n}q(s,\boldsymbol{a}-2\boldsymbol{e}_1+\boldsymbol{e}_2)\\
&+\sum_{j=3}^{n}\frac{j(a_j+1)}{n}q(s,\boldsymbol{a}-\boldsymbol{e}_1-\boldsymbol{e}_{j-1}+\boldsymbol{e}_j)\Bigg]\\
&+\frac{\binom{n}{2}\lambda(s)}{\gamma(s,\boldsymbol{a})}\left[\sum_{j=1}^{n-1}\frac{j(a_j+1}{n-1}q(s,\boldsymbol{a}+\boldsymbol{e}_j-\boldsymbol{e}_{j+1})\right], \quad (8)
\end{aligned}$$

with the convention that $q(s,\boldsymbol{a}) = 0$ if any $a_i < 0$. In particular,

$$\mathcal{L}q(s,\boldsymbol{e}_n) = \frac{\binom{n}{2}\lambda(s)}{\gamma(s,\boldsymbol{e}_n)}q(s,\boldsymbol{e}_{n-1}). \quad (9)$$

Averaging over the time to the first event before t, we see that the sampling probabilities satisfy the integro-recurrence equation

$$q(t,\boldsymbol{a}) = \int_t^\infty \mathcal{L}q(s,\boldsymbol{a})\gamma(s,\boldsymbol{a})\exp\left(-\int_t^s \gamma(u,\boldsymbol{a})\mathrm{d}u\right)\mathrm{d}s. \quad (10)$$

When the population has constant size, so that $\lambda(s) \equiv 1$, the integral equation in (10) reduces to a recursion given by Karlin & McGregor (1972); see also Pollak (1984) and Tavaré (1994). In that case, the sampling probability $q(t,\boldsymbol{a})$ is independent of t, and is given by the Ewens sampling formula (Ewens 1972). Of particular interest is the sampling probability $q(\boldsymbol{a}) \equiv q(0,\boldsymbol{a})$ for a sample of size $|\boldsymbol{a}|$ taken at time 0.

(*b*) *Infinite-sites models*

We turn now to the analogous development for the infinite-sites model. In this case, each mutation occurs at a DNA site that has not mutated previously, and so introduces a new segregating site into the sample. If distinct sequences are labelled as alleles, then the allele frequencies behave just like the infinite-alleles model. Our presentation leans heavily on the results described in Griffiths & Tavaré (1994*b*), to which the reader is encouraged to turn for much greater detail than space permits us here.

The mutational structure in the infinite-sites process is such that each gene in the sample can be thought of as an infinitely long sequence of completely linked sites, each of which is a 0 or a 1. A 0 denotes the ancestral type, a 1 the mutant type. Each mutation results in a new segregating site in the sample, and changes the 0 to a 1 at that site. Ethier & Griffiths (1987) and Griffiths (1987) show how each set of sequences corresponds to a rooted genealogical tree, and show how this tree is represented by writing each sequence in the sample as a vector $\boldsymbol{y} = (y_{i0}, y_{i1}, \ldots)$ of integers. For example, the sequences

gene 1 . . . 1 0 1 0 0 0 1 0 1 . . .
gene 2 . . . 1 0 1 0 0 0 0 0 0 . . .
gene 3 . . . 1 0 0 1 0 1 0 0 0 . . .
gene 4 . . . 1 0 0 1 0 1 0 1 0 . . .
gene 5 . . . 1 0 0 1 0 1 0 1 0 . . .
gene 6 . . . 0 1 0 0 1 0 0 0 0 . . .

may be represented as

gene 1 (9,7,3,1,0)
gene 2 (3,1,0)
gene 3 (6,4,1,0)
gene 4 (8,6,4,1,0)
gene 5 (8,6,4,1,0)
gene 6 (5,2,0)

The rooted genealogical tree is given in figure 1.

Griffiths (1989) shows how to compute the probability of the tree in the constant population-size model. When the ancestral labelling is unknown,

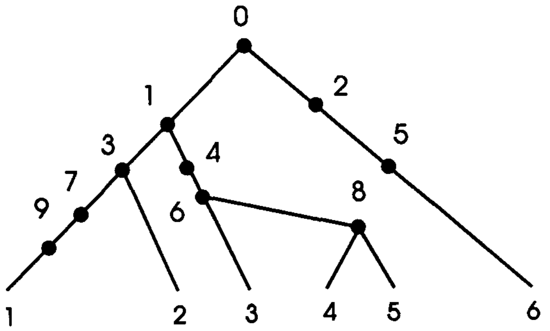

Figure 1. Rooted genealogical tree.

the sequences may be represented as an unrooted tree in which the vertices represent genes and the number of mutations between sequences are represented as numbers along the edges. The unrooted tree corresponding to the example sequences is given in figure 2, with the vertices labelled as to which gene they represent. The unlabelled vertex corresponds to an inferred sequence in the history of the sample. The relationship between rooted and unrooted trees is explained in more detail in Griffiths & Tavaré (1994*b*).

They also show how the probability of unrooted trees may be computed in the constant population-size setting. The aim here is to extend this to the variable population size case.

To this end, suppose that in a sample of n genes there are d distinct sequences $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d$, grouped according to their multiplicities $\boldsymbol{n} = (n_1, \ldots, n_d)$. Write the data in the form $(T, \boldsymbol{n}) \equiv (\boldsymbol{x}, \ldots, \boldsymbol{x}_d, \boldsymbol{n})$, where T denotes a rooted tree and $\boldsymbol{n}$ the multiplicities of the leaves. Denote the size of the sample by $|\boldsymbol{n}| = n_1 + \ldots + n_d$, and let $p(t, (T, \boldsymbol{n}))$ be the probability that a particular ordered sample taken at time t has configuration $(T, \boldsymbol{n})$. Let

$$p^0(t, (T, \boldsymbol{n})) = \frac{n!}{n_1! \ldots n_d!} p(t, (T, \boldsymbol{n}))$$

be the probability of the corresponding unordered sample. In the spirit of (7) and (8), we derive a recurrence satisfied by these sampling probabilities.

First note that if the sample is $(T, \boldsymbol{n})$ at time t, with $|\boldsymbol{n}| = n$, the time S_t of the first event that changes the structure of the sample has distribution determined by

$$\mathbb{P}(S_t > s) = \exp\left(-\int_t^s \gamma(u, n) \mathrm{d}u\right), \qquad (11)$$

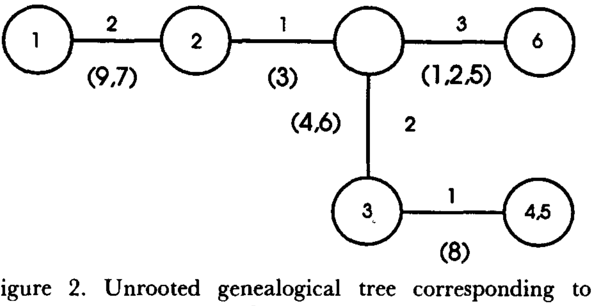

Figure 2. Unrooted genealogical tree corresponding to figure 1.

where

$$\gamma(u, n) = \binom{n}{2}\lambda(u) + \frac{n\theta}{2}. \qquad (12)$$

Looking back at the effects of either coalescences or mutations shows that the analogue of (8) is

$$\mathcal{L}p^0(s, (T, \boldsymbol{n})) =$$

$$\sum_{k:n_k \geqslant 2} \frac{n(n_k - 1)\lambda(s)}{2\gamma(s, n)} p^0(s, (T, \boldsymbol{n} - \boldsymbol{e}_k)) +$$

$$\sum_{\substack{k:n_k=1, x_{k0}\text{ distinct}\\ \mathcal{S}\boldsymbol{x}_k \neq \boldsymbol{x}_j \text{ for all } j}} \frac{\theta}{2\gamma(s, n)} p^0(s, (\mathcal{S}_k T, \boldsymbol{n})) + \qquad (13)$$

$$\sum_{\substack{k:n_k=1\\ x_{k0}\text{ distinct}}} \sum_{j:\mathcal{S}\boldsymbol{x}_k = \boldsymbol{x}_j} \frac{\theta(n_j + 1)}{2\gamma(s, n)} p^0(s, (\mathcal{R}_k T, \mathcal{R}_k(\boldsymbol{n} + \boldsymbol{e}_j))),$$

where $\mathcal{S}$ is the shift operator that deletes the first coordinate of a sequence, $\mathcal{S}_k T$ deletes the first coordinate of the kth sequence of T, and 'x_{k0} distinct' means that $x_{k0} \neq x_{ij}$ for all $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$ and $(i,j) \neq (k, 0)$. The boundary condition is $p^0(t, (T, \boldsymbol{e}_1)) = 1$. Analogous to (10), we now have

$$p^0(t, (T, \boldsymbol{n})) =$$

$$\int_t^\infty \mathcal{L}p^0(s, (T, \boldsymbol{n}))\gamma(s, n) \exp\left(-\int_t^s \gamma(u, n)\mathrm{d}u\right)\mathrm{d}s. \qquad (14)$$

This recursion can be used to find the corresponding probability $q^0(t, (Q, \boldsymbol{n}))$ that the sample taken at time t has labelled, unrooted genealogical tree Q with multiplicities $\boldsymbol{n}$. From Griffiths & Tavaré (1994*b*),

$$q^0(t, (Q, \boldsymbol{n})) = \sum_{T \in C(Q)} p^0(t, (T, \boldsymbol{n})), \qquad (15)$$

where $C(Q)$ is the class of distinct labelled rooted trees constructed from Q. Finally, the probability $q^*(t, (Q, \boldsymbol{n}))$ of the unlabelled, unrooted tree may be found by dividing the probability in (15) by the combinatorial quanitity $a(Q, \boldsymbol{n})$ defined in Griffiths & Tavaré (1994*b*). This quantity is not a function of the unknown parameters of the model, and as such plays no role in the use of the likelihood to find the maximum likelihood estimates of parameters. We can therefore base our estimation on $q^0(0, (Q, \boldsymbol{n}))$.

4. MONTE CARLO LIKELIHOODS

Except in the case of constant population size, the recursions in (10) and (14) are difficult to solve, either explicitly or numerically, when the sample size is at all large. With this difficulty in mind, we describe a Markov chain Monte Carlo method that proves useful in approximating the solutions we are looking for. The methods are based on a generalization of a technique exploited for constant population size models in Griffiths & Tavaré (1994*a*,*b*).

The recursions in (10) and (14) have a common form, and they may be written

$$q(t, x) = \int_t^\infty \sum_y r(s; x, y) q(s, y) g(t, x; s) \mathrm{d}s, \qquad (16)$$

where $r(s;x,y) \geqslant 0$ and $g(t,x;s)$ is the density of the time of the first interesting event after time t;

$$g(t,x;s) = \gamma(s,x)\exp\left(-\int_t^s \gamma(u,x)\mathrm{d}u\right). \qquad (17)$$

Define

$$f(s;x) = \sum_y r(s;x,y),$$

$$P(s;x,y) = \frac{r(s;x,y)}{f(s;x)}, \qquad (18)$$

and rewrite (16) as

$$q(t,x) = \int_t^\infty f(s;x)\sum_y P(s;x,y)q(s,y)g(t,x;s)\mathrm{d}s. \qquad (19)$$

We associate a non-homogeneous Markov chain $\{X(t), t \geqslant 0\}$ with (19) as follows: Given that $X(t) = x$, the time of the next transition has density $g(t,x;s)$, and given that a change of state occurs at time s, the probability that the next state is y is $P(s;x,y)$. This construction can be continued, resulting in a Markov process with a set of absorbing states. These absorbing states correspond to x for which $q(\cdot,x)$ can be computed easily, either explicitly, or by conventional numerical methods. We use the process $X(\cdot)$ to give a probabilistic representation of $q(t,x)$. Let $\tau_1 < \tau_2 \ldots < \tau_k = \tau$ be the jump times of $X(\cdot)$, satisfying $\tau_0 \equiv t < \tau_1$, where τ is the time to hit the absorbing states. Then

$$q(t,x) = \mathbb{E}_{(t,x)}q(\tau, X(\tau))\prod_{j=1}^{k} f(\tau_j; X(\tau_{j-1})), \qquad (20)$$

where $\mathbb{E}_{(t,x)}$ denotes expectation with respect to $X(t) = x$. A useful modification of (20) is obtained by choosing τ to be any finite stopping time for $\{X(t), t \geqslant 0\}$.

One application of the representation in (20) provides a means to approximate $q(x) \equiv q(0,x)$. Simulate many independent copies of the process $\{X(t), t \geqslant 0\}$ starting from $X(0) = x$, and compute the observed value of the functional under the expectation sign in (20) for each of them. The average of these numbers is an unbiased estimate of $q(x)$, and we may use conventional theory to see how accurately $q(x)$ has been estimated. If it is known how to calculate $q(\tau, X(\tau))$ for a suitable stopping rule τ, then this can be used to reduce simulation time and the variance of the estimate of $q(x)$.

The probability $q(t,x)$ is usually a function of some unknown parameters, which we denote here by Γ; we write $q_\Gamma(t,x)$ to emphasize the dependence on Γ. A version of importance sampling may be used to construct a single process $X(\cdot)$ with parameters Γ_0, from which estimates of $q_\Gamma(t,x)$ may be found for other values of Γ. We have

$$q_\Gamma(t,x) = \int_t^\infty \sum_y f_{\Gamma,\Gamma_0}(t,x;x,y)P_{\Gamma_0}(s;x,y)q_\Gamma(s,y)g_{\Gamma_0}(t,x;s)\mathrm{d}s, \qquad (21)$$

where

$$f_{\Gamma,\Gamma_0}(t,x;s,y) = \frac{f_\Gamma(s;x)g_\Gamma(t,x;s)P_\Gamma(s;x,y)}{g_{\Gamma_0}(t,x;s)P_{\Gamma_0}(s;x,y)},$$

and $f_\Gamma(s;x)$ and $\mathrm{P}_\Gamma(s;x,y)$ and defined in (18). The representation analogous to (20) is

$$q_\Gamma(t,x) = \mathbb{E}_{(t,x)}q(\tau, X(\tau))\prod_{j=1}^{k} f_{\Gamma,\Gamma_0}(\tau_{j-1}, X(\tau_{j-1}); \tau_j, X(\tau_j)), \qquad (22)$$

from which estimates of $q_\Gamma(t,x)$ may be simulated as described above. In practice, several different values of the generating parameters Γ_0 are used, and the results combined to form a single estimate of $q_\Gamma(t,x)$ for several different values of Γ. In the next two sections, we specialize this description to the infinite-alleles and infinite-sites models.

(*a*) *Infinite-alleles models*

In this case, the state space consists of points x of the form $\boldsymbol{a}$, and the Markov chain $X(\cdot)$ is denoted by $\boldsymbol{a}(\cdot)$. For $|\boldsymbol{a}| = n > 1$, define

$$h(s,\boldsymbol{a}) = \theta\sum_{j=2}^{n}\frac{j(a_j+1)}{n} + \lambda(s)\sum_{j=1}^{n-2} j(a_j+1), \qquad \boldsymbol{a} \neq \boldsymbol{e}_n,$$
$$= (n-1)\lambda(s), \qquad \boldsymbol{a} = \boldsymbol{e}_n.$$

The functions f and P of (18) are given by

$$f(s;\boldsymbol{a}) = \frac{nh(s,\boldsymbol{a})}{2\gamma(s,\boldsymbol{a})},$$

and

$$P(s;\boldsymbol{a},\boldsymbol{b}) = \begin{cases} \dfrac{2\theta(a_2+1)}{nh(s,\boldsymbol{a})}, & \boldsymbol{b} = \boldsymbol{a} - 2\boldsymbol{e}_1 + \boldsymbol{e}_2, \\ \dfrac{\theta j(a_j+1)}{nh(s,\boldsymbol{a})}, & \boldsymbol{b} = \boldsymbol{a} - \boldsymbol{e}_1 - \boldsymbol{e}_{j-1} + \boldsymbol{e}_j, \quad 3 \leqslant j \leqslant n, \\ \dfrac{\lambda(s)j(a_j+1)}{h(s,\boldsymbol{a})}, & \boldsymbol{b} = \boldsymbol{a} + \boldsymbol{e}_j - \boldsymbol{e}_{j+1}, \quad 1 \leqslant j \leqslant n-1, \end{cases}$$

with the convention that $P(s;\boldsymbol{a},\boldsymbol{b}) = 0$ for any state $\boldsymbol{b}$ which has any $b_j < 0$. In particular,

$$P(s;\boldsymbol{e}_n,\boldsymbol{e}_{n-1}) = 1, \qquad n \geqslant 2.$$

The initial conditions are $q(t,\boldsymbol{e}_1) = 1$ for all $t \geqslant 0$.

Alternatively suppose τ is chosen to be the stopping time

$$\tau = \inf\{t : |\boldsymbol{a}(t)| \leqslant 2\}.$$

Defining

$$p(\tau) = \int_0^\infty \mathrm{e}^{-\theta u}\lambda(\tau+u)\exp(-(\Lambda(\tau+u) - \Lambda(\tau)))\mathrm{d}u,$$

we see that

$$q(\tau,\boldsymbol{a}(\tau)) = \begin{cases} p(\tau), & \boldsymbol{a}(\tau) = \boldsymbol{e}_2, \\ 1 - p(\tau), & \boldsymbol{a}(\tau) = 2\boldsymbol{e}_1. \end{cases}$$

(*b*) *Infinite-sites models*

From (12) and (14), we may define

$$h(s;(T,\boldsymbol{n})) = \sum_{k:\, n_k \geqslant 2} n(n_k - 1)\lambda(s) + \theta m, \tag{23}$$

where $n = |\boldsymbol{n}|$ and m is given by

$$m = \sum_{\substack{k:n_k=1,\, x_{k,0}\text{ distinct}\\ S\boldsymbol{x}_k\neq\boldsymbol{x}_j\text{ for all } j}} 1 + \sum_{\substack{k:n_k=1\\ x_{k,0}\text{ distinct}}} \sum_{j:S\boldsymbol{x}_k=\boldsymbol{x}_j} (n_j+1).$$

The functions f and P of (18) are therefore

$$f(s;(T,\boldsymbol{n})) = \frac{h(s;(T,\boldsymbol{n}))}{2\gamma(s,n)},$$

and

$$P(s;(T,\boldsymbol{n}),(T',\boldsymbol{n}')) = \begin{cases} \dfrac{n(n_k-1)\lambda(s)}{h(s;(T,\boldsymbol{n}))}, & (T',\boldsymbol{n}') = (T,\boldsymbol{n}-\boldsymbol{e}_k), \qquad n_k \geqslant 2, \\ \dfrac{\theta}{h(s;(T,\boldsymbol{n}))}, & (T',\boldsymbol{n}') = (\mathcal{S}_k T,\boldsymbol{n}), \\ \dfrac{\theta(n_j+1)}{h(s;(T,\boldsymbol{n}))}, & (T',\boldsymbol{n}') = (\mathcal{R}_k T,\mathcal{R}_k(\boldsymbol{n}+\boldsymbol{e}_j)). \end{cases}$$

The second type of move is possible for those k satisfying $n_k = 1$, $x_{k,0}$ distinct, and $S\boldsymbol{x}_k \neq \boldsymbol{x}_j$ for all j. The third type of move can occur for those j and k for which $n_k = 1$, $x_{k,0}$ distinct and $S\boldsymbol{x}_k = \boldsymbol{x}_j$. The process is absorbed into $(T_0, \boldsymbol{e}_1)$, where T_0 is a tree with a singleton vertex.

If τ is the stopping time when first $|\boldsymbol{n}(t)| = 2$, and $\boldsymbol{n}(\tau) = (1,1)$, then

$$p^0(\tau,(T(\tau),\boldsymbol{n}(\tau)))) =$$
$$2\int_0^\infty \frac{(\theta u)^{i+j}\mathrm{e}^{-\theta u}}{2^{i+j}i!\,j!}\lambda(\tau+u)\exp(-(\Lambda(\tau+u)-\Lambda(\tau)))\mathrm{d}u,$$

where i and j are the numbers of mutant sites on the two edges of $T(\tau)$. If $\boldsymbol{n}(\tau) = (2)$, then $p^0(\tau,(T(\tau),\boldsymbol{n}(\tau))) = p(\tau)$.

5. MITOCHONDRIAL DATA

Ward *et al.* (1991) sequenced the first 360 base pairs of the mitochondrial control region (D-loop) of 63 North American native Indians from the Nuu-Chah-Nulth of Vancouver Island. The region has no transversions, and so each site in the sequences is binary: either purine (A,G) or pyrimidine (C, T). The 201 pyrimidine sites have been analysed extensively (Lundstrom *et al.* 1992); we focus on the 159 purine sites. The data exhibit five segregating sites, defining seven different alleles. The data are presented in table 1.

There is a simple diagnostic for assessing whether a collection of sites is consistent with the infinite-sites model. In the present setting, the condition is that in any pair of sites not all of the patterns GG, AA, AG, and GA are observed. Clearly, sites 2 and 4 violate this condition. Inconsistencies are usually due to back-substitutions, and suggest that a more detailed model of sequence evolution should be used. One approach is mentioned in the discussion.

There are several ways in which subsets of the data that are consistent may be chosen. One is to select a maximally consistent subset of the sites, for example {1,3,4,5} or {1,2,3,5}. An alternative is to choose a subset of genes (or individuals) that form a consistent set. For these data, removing the five individuals with allele C produces a consistent subset of individuals. From now on, we use the resulting sample of 58 genes, which are linked by the unrooted tree shown in figure 3.

Under the assumption of constant population size, the maximum likelihood estimator of θ under the infinite-alleles model is $\hat{\theta} = 1.48$, with an estimated variance of 0.55. The corresponding estimator of θ under the infinite-many-sites model may be found using the method described above with $\nu(x) \equiv 1$, giving $\hat{\theta} = 1.19$ with an estimated variance of 0.35. As is typical for such data, the precision of the estimate is not high.

To look for the effects of expanding populations, we fitted the varying environments model with $\nu(x) = \mathrm{e}^{-\beta x}$, and estimated the expansion parameter β first using just the allelic configuration of the data, $a_1 = 2, a_3 = 1, a_7 = 1, a_{19} = 1, a_{27} = 1$. An estimate of the likelihood surface of this configuration with respect to β for $\theta = 1.5$ using 60000 runs gave a monotonic decreasing likelihood. The maximum likelihood estimate of β is zero, or close to zero. Fixing β for small values and simulating likelihood surfaces for θ gave maximum likelihood estimates of approximately $\theta = 1.5$. The data in table 1 provide little evidence that $\beta > 0$, possibly because of two alleles having a high frequency.

We analysed the more detailed sequence structure using the likelihood q^0 given in (15). For fixed θ, the Monte Carlo approximant to the likelihood surface as a function of β was found using the surface simulation method described in §4, equations (21) and (22). For $\theta = 1.5$, we generated three such curves using generating parameters $\beta_0 = 0.1$, 1.0, 2.0. These curves are combined into a single approximant by weighting the estimated values inversely proportional to their variances. The estimated log-likelihood curve as a function of the expansion parameter β is given in figure 4.

The maximum likelihood estimate of β is $\hat{\beta} = 1.54$, corresponding to a log-likelihood of $\log q^0 = -13.70$. When $\beta = 0$, the corresponding log-likelihood is

Table 1. *Variable purine sites in the control region*

allele	sequence	allele frequency
A	GGGGA	27
B	GGAGA	3
C	GAGGA	5
D	GAGAA	1
E	GGGAA	19
F	GGGAG	1
G	AGGAA	7

$\log q^0 = -14.21$. This gives a value of 1.02 for the likelihood ratio statistic, and suggests that there is little evidence in these data for population expansion of the exponential type. To take account of the variability in our estimate of θ, we performed similar analyses for $\theta = 0.5$ and $\theta = 1.0$. For $\theta = 1.0$, the maximum likelihood estimate of β is $\hat{\beta} = 0.50$ with a log-likelihood of -14.04, in comparison to the log-likelihood when $\beta = 0$ of -14.14. When $\theta = 0.5$, the maximum likelihood estimate of β is $\hat{\beta} = 0.07$ with a log-likelihood of -15.31, in comparison to the log-likelihood when $\beta = 0$ of -15.32. These results support the observation of no exponential population expansion.

We also investigated the sensitivity of the surface simulation method in estimating θ for fixed values of β. For each value of β, we generated three likelihood curves in θ corresponding to generating parameters $\theta_0 = 0.5, 1.0, 1.5$ and combined them as above. The maximum likelihood estimates of θ are $\hat{\theta} = 1.67$ $(\beta = 1.5)$, $\hat{\theta} = 1.53$ $(\beta = 1.0)$ and $\hat{\theta} = 1.37 (\underset{\wedge}{\beta} = 0.5)$. These values are consistent with the value $\hat{\theta} = 1.19$ found when $\beta = 0.0$.

To understand something about our ability to detect expansion based on allele counts alone, genealogical trees were simulated for a sample size of 50 for $\theta = 1.5$, and $\beta = 0.0, 0.5, 1.0, 2.0, 5.0$, with 20 trees for each value of β. As β increases with θ fixed, the time to the most recent common ancestor decreases (in the timescale used here), the number of alleles decreases, the most recent common ancestor is in the sample with a high frequency, and the shape of the tree changes to a star phylogeny. In the simulated trees the star effect is evident for some trees when $\beta = 1.0$, and is clearly shown for trees with $\beta = 2.0$, 5.0. A typical tree, generated with $\beta = 5.0$, shows this shape:

44:(0)
 2:(1,0)
 2:(2,0)
 1:(3,0)
 1:(4,0).

This particular tree has a simulated likelihood at $\theta = 1.5$, $\beta = 0.0$ of 0.000523 (standard error 0.000004), and at $\theta = 1.5$, $\beta = 5.0$ of 0.0026 (standard error 0.00003). The likelihood ratio statistic is 7.8, suggesting strongly that $\beta > 0$. The likelihood function is quite flat around $\beta = 5.0$. The allelic configuration of these data is $a_1 = 2, a_2 = 2, a_{44} = 1$. The probability of this configuration with $\theta = 1.5$, $\beta = 0.0$ is 0.0013 (exact) and with $\beta = 5.0$ is 0.0039 (simulated with standard error 0.00005). The likelihood ratio statistic is 2.19, hinting weakly that $\beta > 0$. The ability to detect $\beta > 0$ from just the allele configuration does not seem high.

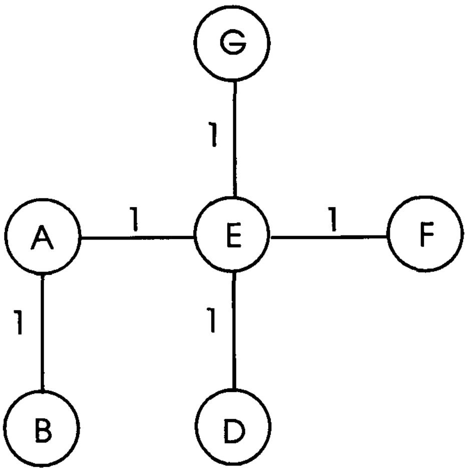

Figure 3. Unrooted tree from the purine sites.

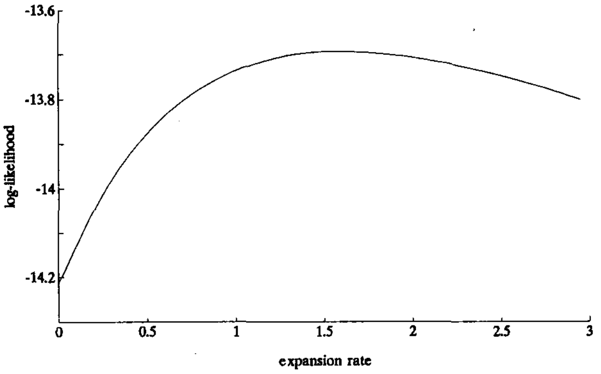

Figure 4. Estimated log-likelihood curve as a function of the expansion parameter β.

6. DISCUSSION

The sampling theory of population genetics processes in a varying environment is, somewhat surprisingly, still in its infancy. The Monte Carlo likelihood technique we have illustrated here provides one approach to estimation and inference for such processes. One appealing feature is that the method can readily be modified to cover other problems of interest. For example, we can study the distribution of other summary statistics such as the number of alleles or the number of segregating sites seen in a sample. For applications to other human population data, it may also be important to assess the effects of population subdivision. Nath & Griffiths (1993) have shown how these techniques may be used to study the sampling distribution of an island model with equal population sizes. Ancestral inference is another important application: we can use this approach to study the distribution of the time to the most recent common ancestor of a sample, conditional on the observed data. It is also possible to derive a corresponding sampling theory for the finite-sites model, a process that is perhaps more useful for the analysis of DNA sequence data. Suppose then that there are d possible alleles, and that the probability of a mutation from type i to type j is p_{ij}, $1 \leqslant i, j \leqslant d$. The state space comprises d-dimensional vectors $x = \boldsymbol{n} = (n_1, \ldots, n_d)$, where n_i is the number of times the allele i is observed in the sample. The jump time intensity γ in (17) for a state $\boldsymbol{n}$ satisfying $n_1 + \ldots + n_d = n > 1$ is given by

$$\gamma(s, \boldsymbol{n}) = \frac{n(n-1)}{2}\lambda(s) + \frac{n\theta}{2}\left(1 - \sum_{i=1}^{d} \frac{n_i}{n} p_{ii}\right),$$

and the non-zero entries of the kernel r in (16) are

$$\begin{aligned} r(s;\boldsymbol{n},\boldsymbol{m}) &= \frac{\theta}{2\gamma(s,\boldsymbol{n})}(n_i+1)p_{ij}, \\ &\quad \boldsymbol{m}=\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j, \qquad 1 \leqslant i,j \leqslant d, \qquad n_j>0, i \neq j \\ &= \frac{n\lambda(s)}{2\gamma(s,\boldsymbol{n})}(n_j-1), \\ &\qquad \boldsymbol{m}=\boldsymbol{n}-\boldsymbol{e}_j, \qquad 1 \leqslant j \leqslant d, \qquad n_j>0. \end{aligned}$$

The initial conditions are $q(t,\boldsymbol{e}_i)=\pi_i$, the initial frequency of allele i. We shall report on this in more detail elsewhere.

The computational tools outlined here are available on request in portable C code. There are two programs, ALLELES and PTREESIM. The first covers data analysis for the infinite-alleles model, the second for the infinite-sites model. Both allow for varying environments, as illustrated for the mitochondrial data by an exponential population size function $\lambda(s)=e^{\beta x}$, and the limiting case $\beta=0$ of constant population size. In both cases, surface simulation to estimate θ for fixed β, or β for fixed θ is provided. Of course, many other forms of expansion might be appropriate and these can be handled by precisely the same techniques. We are currently adding other types of population expansion to the programs.

S.T. was supported in part by NSF grant DMS 90-05833.

REFERENCES

Chakraborty, R. 1977 Distribution of nucleotide differences between two randomly chosen cistrons in a population of variable size. *Theor. Popul. Biol.* **11**, 11–22.

Chakraborty, R. & Nei, M. 1977 Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. *Evolution* **31**, 347–356.

Donnelly, P. 1986 A genealogical approach to variable population size models in population genetics. *J. appl. Prob.* **23**, 283–296.

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

Griffiths, R.C. 1987 An algorithm for constructing genealogical trees. Statistics Research Report 163, Department of Mathematics, Monash University, Australia.

Griffiths, R.C. 1989 Genealogical-tree probabilities in the infinitely-many site model. *J. math. Biol.* **27**, 667–680.

Griffiths, R.C. & Pakes, A.G. 1988 An infinite-alleles version of the simple branching process. *Adv. appl. Prob.* **20**, 489–524.

Griffiths, R.C. & Tavaré, S. 1994*a* Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* (In the press.)

Griffiths, R.C. & Tavaré, S. 1994*b* Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosci.* (Submitted.)

Karlin, S. & McGregor, J. 1972 Addendum to a paper of W. Ewens. *Theor. Popul. Biol.* **3**, 113–116.

Kingman, J.F.C. 1982*a* On the genealogy of large populations. *J. appl. Prob.* **19A**, 27–43.

Kingman, J.F.C. 1982*b* Exchangeability and the evolution of large populations. In *Exchangeability in probability and statistics* (ed. G. Koch & F. Spizzichino), pp. 97–112. North-Holland Publishing Company.

Lundstrom, R., Tavaré, S. & Ward, R.H. 1992 Estimating mutation rates from molecular data using the coalescent. *Proc. natn. Acad. Sci. U.S.A.* **89**, 5961–5965.

Marjoram, P. & Donnelly, P. 1994*a* Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics* **136**, 673–683.

Marjoram, P. & Donnelly, P. 1994*b* Genealogical structure in populations of variable size, and the time since Eve. (Submitted.)

Nath, H.B. & Griffiths, R.C. 1993 Estimation in an island model undergoing a multidimensional coalescent process. Statistics Research Report 226, Monash University, Australia.

Nei, M., Maruyama, T. & Chakraborty, R. 1977 The bottleneck effect and genetic variability in populations. *Evolution* **29**, 1–10.

Pollak, E. 1984 The Ewens sampling formula in a population that varies in size. In *Experimental design, statistical models and genetic statistics. Essays in honor of Oscar Kempthorne* (ed. K. Hinkelmann), pp. 385–400. New York and Basel: Marcel Dekker, Inc.

Rogers, A.R. & Harpending, H. 1992 Population growth makes waves in the distribution of pairwise genetic differences. *Molec. Biol. Evol.* **9**, 552–569.

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

Taib, Z. 1992 *Branching processes and neutral evolution.* (*Lect. Notes Biomath.* **93**.) Berlin: Springer-Verlag.

Tavaré, S. 1993 Calibrating the clock: using stochastic processes to measure the rate of evolution. In *Calculating the secrets of life* (ed. E. S. Lander). Washington DC: National Adacemy Press. (In the press.)

Ward, R.H., Frazier, B.L., Dew, K. & Pääbo, S. 1991 Extensive mitochondrial diversity within a single Amerindian tribe. *Proc. natn. Acad. Sci. U.S.A.* **88**, 8720–8724.

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

Watterson, G.A. 1984 Allele frequencies after a bottleneck. *Theor. Popul. Biol.* **26**, 387–407.

Watterson, G.A. 1989 The neutral alleles model with bottlenecks. In *Mathematical evolutionary theory* (ed. M. W. Feldman), pp. 26–40. Princeton University Press.