Monte Carlo inference methods in population genetics

R.C. Griffiths
Monash University *

Simon Tavaré
University of Southern California

Mathl. Comput. Modelling, **23**, 141–158, 1996

Abstract

We study the distribution of summary statistics of the sample configuration of DNA sequences taken from a large population that has evolved with determistically varying population size. We study the information available in the number of alleles and segregating sites for estimating the substitution rate, and for making inferences about the time to the most recent common ancestor of the sample. We exploit a Monte Carlo method for solving recurrence equations that define the requisite sampling probabilities. The methods are illustrated with some mitochondrial control region data.

Running Head: Monte Carlo methods in genetics
Keywords: Coalescent, Varying environment, Infinitely-many-sites, Ancestral inference

*R.C. Griffiths, Department of Mathematics, Monash University, Clayton 3168, Australia. S. Tavaré, Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, CA 90089-1113, USA. ST is supported in part by NSF grant DMS 90-05833. The authors would like to thank the referees for their thoughtful comments on an earlier draft.

1 Introduction

Comparisons of DNA sequence information from individuals within a species can be used to study aspects of the evolutionary history of that species. In the study of human evolution one particular molecule, mitochondrial DNA, has played a central role; cf. Cann et al. [1], Stoneking [2], Ward et al. [3], Shields et al. [4]. Mitochondria are small circular DNA molecules about 16,500 base pairs in length. They live outside the nucleus of cells and they play a role in energy production. Mammalian mitochondria are maternally inherited, which makes them ideal for studying the maternal lineages in which they arise. One part of the molecule, an 1100 base pair stretch known as the control region or D-loop, has been particularly important for human evolutionary studies, primarily because its very high mutation rate produces sequence differences between individuals who are (evolutionarily) quite closely related. Samples of D-loop sequences are now available from many diverse human populations. The data used in this paper are a sample of sequences from the beginning of the control region, taken from the Nuu-Chah-Nulth Indians (Ward et al. [3]). These sequences, one from each individual in the sample, may be used to track the maternal ancestry of the populations under study.

Suppose that we take a random sample of n such sequences. We assume the sequences are aligned, so that each of the n sequences has the same length, s base pairs. The sequences can be thought of as an $n \times s$ matrix D, each row corresponding to the sample sequence of one individual. In our data, $s = 360$ and $n = 63$. Each column is called a *site*. There will be some sites at which the sample sequences have an identical letter (either A, C, G, or T, the four bases in DNA), and some sites at which the sample sequences vary. These sites are called *segregating sites*. Moreover, there will be some sequences in the sample that are identical. Distinct sequences are called *alleles* or *lineages*. The number of segregating sites and the number of alleles observed in a sample of sequences are important summary

descriptions of the variability in the data.

Over time, changes in these sequences are passed on from generation to generation. In this paper, we suppose that the changes are caused by *substitutions*, the replacement of one base by another. In particular, we ignore the effects of length variations caused by insertions and deletions, and we assume there is no recombination in the region, so that the sites in the sequences are *completely linked*, being passed on intact from generation to generation. Our aim is to model the substitution process, and use it to estimate the rate at which such substitutions occur. This rate calibrates the molecular clock of the region, and so allows us to make inferences about many ancestral features of the sample and of the population. One quantity of great interest in human evolution is the time to the most recent common ancestor (MRCA) of the sample or population.

The result of each mutation can be modeled in many different ways, depending largely on the type of data being analyzed. The simplest assumption is that each mutation produces an allele never seen before in the population. Under this *infinitely-many-alleles* model the labels of alleles are arbitrary, being used merely to distinguish different types from each other. For some DNA sequence data, the *infinitely-many-sites* assumption is often used, particularly in theoretical studies of sequence variability. Under this model, each mutation occurs at a site that has never had a mutation before. This means that each site in the DNA sequence is binary, the two possible types at a site corresponding to the ancestral or mutant base. Each mutation introduces a new segregating site into the sample. For this model, the matrix D conventionally describes only the segregating sites, so that s is the number of segregating sites in the data. If each distinct sequence of sites is identified as an allele, then the structure of the allele frequencies is precisely that of the infinitely-many-alleles model. A more detailed model for DNA sequence data specifies different mutation rates and probabilities for different sites, and in particular allows mutations to occur at a given site more than once.

In earlier work [5-8] we studied the sampling theory of these substitution models. We derived the probability distribution of the data matrix D under particular mutation models, and used this distribution to find maximum-likelihood estimators of the substitution parameters. Typically, these estimators cannot be found in explicit form or by standard numerical analysis techniques for maximizing functions, due primarily to the extremely complicated nature of the distributions. We resorted instead to a Monte Carlo approach in which likelihoods are simulated. Likelihoods for the full joint distribution of the data D can be very time consuming to approximate in this way, particularly when the substitution mechanism is complicated.

In this paper we assess the effects of using simpler summary statistics about the data, in particular the number of alleles and segregating sites in the sample, for inferences about the mutation rate in the infinitely-many-sites model. This allows us to assess the trade-off between statistically more precise estimates based on the full data D, and the time required to generate the estimates. We also assess the effects that the use of summary statistics have on ancestral inference. Specifically, we approximate the conditional distribution of the time T_{MRCA} to the MRCA of the sample, given the number of segregating sites and alleles, and we compare this to the conditional distribution based on the full data D.

This paper is organized as follows. The ancestry of the individuals in the sample is not known in any detail, and therefore has to be modeled. In Section 2, we describe the coalescent (Kingman [9], Hudson [10]), a stochastic process used by population geneticists to describe the random ancestry of such a sample. We allow for deterministic variation in population size (Kingman [11], Slatkin and Hudson [12], Griffiths and Tavaré [5]). The effects of mutation are superimposed on this ancestral tree. In Section 3 we show how this approach can be used to derive (recursions satisfied by) the distribution of the number of alleles and the number of segregating sites, their joint distribution, and the joint distribution of these quantities and T_{MRCA}. In Section 4 we describe a computer-intensive approach in which

the recurrence equation is used to construct a time-inhomogeneous Markov chain in such a way that the required sampling probability is the mean of a functional of the chain up to a hitting time. This provides a conceptually simple Monte Carlo technique for approximating complex sampling probabilities. The methods are illustrated using a sample of mitochondrial D-loop sequences in Section 5.

2 The coalescent

2.1 Ancestral trees

In the population genetics literature the ancestry of the (female) individuals in a random sample from a population is often modeled by a continuous-time stochastic process known as the *coalescent*. This process was introduced by Kingman [9, 11] as an approximation, valid in the limit of large population size, to the ancestral structure of a wide variety of reproduction models. Assume that a random sample of n individuals is taken from a large population with non-overlapping generations that have been of constant size N. The reproduction mechanism can be described as follows: Suppose that in a particular generation the individuals are labeled 1,2, ..., N, and let $\nu_1, \nu_2, \ldots, \nu_N$ be the number of offspring they have. We assume the ν_i are exchangeable random variables with sum N, and that their joint distribution is constant over time. Label the sampling generation as 0, and let $A_n^N(r)$ be the number of distinct ancestors the sample has r generations into the past. To avoid degenerate limiting behavior, we assume that

$$\sigma^2 \equiv \lim_{N\to\infty} \mathrm{Var}(\nu_1) \in (0, \infty). \tag{1}$$

To define the limiting process, begin with a continuous-time pure death process $\{A_n(t), t \geq 0\}$ on the integers $n, n-1, \ldots, 1$. $A_n(\cdot)$ starts from n, moves from state k to $k-1$ at rate $k(k-1)/2$, and is eventually absorbed in the state 1. Under assumption (1) and an additional mild regularity condition, Kingman [11] showed that as $N \to \infty$ the process

$\{A_n^N(\lfloor Nt \rfloor), t \geq 0\}$ converges in distribution to $\{A_n(\sigma^2 t), t \geq 0\}$. The parameter σ^2 has the effect of calibrating the time scale of the ancestral process. For the Wright-Fisher model often considered in the population genetics literature the ν_i have a symmetric multinomial distribution, and hence $\sigma^2 = 1$. Without loss of generality we assume this scaling in what follows.

Let T_j be the amount of time the sample has j distinct ancestors, for $j = n, n-1, \ldots, 2$. When $\sigma^2 = 1$, the T_j are independent exponential random variables, with means $\mathbb{E}(T_j) = 2/(j(j-1))$, from which it follows that the time $T_{MRCA} = T_n + \cdots + T_2$ back to the MRCA has mean

$$\mathbb{E}(T_{MRCA}) = 2\left(1 - \frac{1}{n}\right). \tag{2}$$

The coalescent itself may be thought of as a random rooted tree, with leaves representing the n sample individuals, and vertices where ancestral lines join. In our continuous-time approximation the tree is binary, and the topology of the tree is obtained by randomly merging pairs of individuals. The individual at the root of the tree is the most recent common ancestor of the sample. The tree has j branches of length T_j, for $j = n, n-1, \ldots, 2$.

Kingman's original formulation [9] of the coalescent applied to populations of constant size. The case of deterministically varying population size in which all the generations are large can be accomodated as follows. Suppose the generation at the time of sampling had size $M(0) = N$, and that the size of the population r generations earlier was $M(r)$. Assume once more that in generation r the number of offspring $\nu_i^{(r)}$ born to individual i, $i = 1, 2, \ldots, M(r)$, are exchangeable random variables, and that $\mathrm{Var}(\nu_1^{(r)}) = \sigma^2(r)$. We measure time in units of N generations, and assume there is a strictly positive function $v(\cdot)$ such that

$$\lim_{N\to\infty} \sum_{j=1}^{\lfloor Nt \rfloor} \frac{\sigma^2(r)}{M(r-1)} = \Lambda(t) \equiv \int_0^t \frac{1}{v(s)} ds, \tag{3}$$

satisfying $0 < \Lambda(t) < \infty$ for $t > 0$. Informally, when $\sigma^2(r) \equiv 1$, the function $v(t)$ may be thought of as the relative size of the population Nt generations ago. It will be convenient in

what follows to denote the density of Λ by λ:

$$\lambda(t) = 1/v(t),\ t > 0.$$

Let $A_n^v(t)$ be the number of distinct ancestors time t ago in this (limiting) variable population size model. Griffiths and Tavaré [5, 7] show that the distribution of the process $\{A_n^v(t), t \geq 0\}$ can be defined by the coupling

$$A_n^v(t) = A_n(\Lambda(t)), t \geq 0. \tag{4}$$

See also Kingman [11]. We assume that $\Lambda(\infty) = \infty$, so that the sample may be traced back to a common ancestor with probability one. Most interesting properties of the variable-size process can be calculated using the representation in (4). For example, it defines the joint distribution of the times $T_j, j = n, n-1, \ldots, 2$, and so allows us to study properties of T_{MRCA}. The topology of the coalescent tree in the variable size case is just as before; just the distributions of the lengths of the branches change.

2.2 Mutations in the coalescent

We assume that mutations occur independently among all offspring, the probability that a mutation occurs in a given offspring in a given generation being u. It is conventional to suppose that mutation rates are of the order of $1/N$, in that

$$\lim_{N\to\infty} 2Nu = \theta \in (0, \infty).$$

Mutations (in our case base substitutions) are then superimposed on the coalescent tree of the sample as follows: Conditional on the tree, place mutations according to independent Poisson processes of rate $\theta/2$ in each branch.

3 Sampling equations

3.1 Estimating θ

We begin by studying the distributions of the number of alleles K_n and the number of segregating sites S_n in a sample of n genes from the infinitely-many-sites model. Our method is to derive a recurrence equation satisfied by the distribution of interest. To illustrate how such equations arise we give three examples, the first being the marginal distribution of K_n. A similar approach may be used to derive sampling equations for many other distributions of interest; see [5] for example.

Let $q(t,(n,k))$ be the probability that a sample of n genes taken at time t in the past has k distinct alleles. As we look back into the past at the history of the genes in the sample, we will eventually see either a coalescence event or a mutation event. The time W_t of this first event has distribution

$$\mathbb{P}(W_t > s) = \exp\left(-\int_t^s \gamma(u,n)du\right), \tag{5}$$

where

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

(5) being the probability that no coalescence events or mutations occur in the interval (t,s). We denote the probability density function of W_t by

$$g(t,n;s) = \gamma(s,n)\exp\left(-\int_t^s \gamma(u,n)du\right), \quad s \geq t. \tag{6}$$

If the event occurs at time s, it is a mutation with probability

$$\frac{n\theta}{2\gamma(s,n)} = \frac{\theta}{\theta + (n-1)\lambda(s)},$$

and a coalescence with probability

$$\frac{(n-1)\lambda(s)}{\theta + (n-1)\lambda(s)}.$$

If the mutation occurs, the $n-1$ other genes at time s must have comprised $k-1$ alleles, whereas if the coalescence occurs, the $n-1$ ancestors at time s must have comprised k alleles. Combining these possibilities, we see that for $n \geq 2,\ 1 \leq k \leq n$

$$\begin{aligned} q(t,(n,k)) &= \int_t^\infty \left\{ \frac{\theta}{\theta + (n-1)\lambda(s)} q(s,(n-1,k-1)) \right. \\ &\quad \left. + \frac{(n-1)\lambda(s)}{\theta + (n-1)\lambda(s)} q(s,(n-1,k)) \right\} g(t,n;s)ds, \end{aligned} \tag{7}$$

with the convention that $q(s,(n,k)) = 0$ if $k > n$ or $k < 1$. The initial condition is $q(s,(1,1)) = 1$, and the sampling distribution of interest is given by $q((n,k)) \equiv q(0,(n,k))$.

We may derive an analogous equation for the distribution of S_n. Let $\tilde{q}(t,(n,m))$ be the probability that the sample of size n taken at time t in the past has m segregating sites. Similar reasoning shows that for $n \geq 2, 0 \leq m \leq n$

$$\begin{aligned} \tilde{q}(t,(n,m)) &= \int_t^\infty \left\{ \frac{\theta}{\theta + (n-1)\lambda(s)} \tilde{q}(s,(n,m-1)) \right. \\ &\quad \left. + \frac{(n-1)\lambda(s)}{\theta + (n-1)\lambda(s)} \tilde{q}(s,(n-1,m)) \right\} g(t,n;s)ds, \end{aligned} \tag{8}$$

with the proviso that $\tilde{q}(t,(n,m)) = 0$ if $m < 0$, and initial condition $\tilde{q}(t,(1,0)) = 1$. It is the quantity $\tilde{q}((n,m)) \equiv \tilde{q}(0,(n,m))$ that provides the sampling distribution of the number of segregating sites in our sample.

The third example provides the joint distribution of K_n and S_n. At time t, choose a random sample of size n and a random subsample of size r from those n. Introducing the subsample is a device to produce a recursive system of equations. Let $q^*(t,(n,m;r,k))$ be the probability that the sample has m segregating sites and the subsample has k alleles. We want to find the sampling distribution $q^*((n,m;n,k)) \equiv q^*(0,(n,m;n,k))$. Once more we consider what happens at the first event in the history of the sample. If it was a mutation, then with probability r/n the mutation occurred in the subsample, in which case the remaining $r-1$ genes in the subsample must have formed $k-1$ alleles, and the n genes at time s must

have contained $m-1$ segregating sites. If the mutation occurred outside the subsample (probability $1-r/n$), then the n genes at time s must have had $m-1$ segregating sites, and the r ancestors of the subsample must have had k alleles. The other possibility is that the coalescence occurred first. If it involved ancestors of two genes in the subsample, then the subsample of $r-1$ ancestors at time s must have contained k alleles, and the sample of $n-1$ genes at time s must have had m segregating sites. In the event that the coalescence at time s involved less than 2 of the ancestors of individuals in the subsample, the subsample comprising the r distinct ancestors of the subsample must have had k alleles, and the sample of $n-1$ ancestors at s must have had m segregating sites. Combining these possibilities produces

$$
\begin{aligned}
&q^*(t,(n,m;r,k)) = \\
&\quad \int_t^\infty \left[\frac{\theta}{\theta+(n-1)\lambda(s)} \left\{ \frac{r}{n} q^*(s,(n,m-1;r-1,k-1)) + \left(1-\frac{r}{n}\right) q^*(s,(n,m-1;r,k)) \right\} \right. \\
&\quad + \frac{(n-1)\lambda(s)}{\theta+(n-1)\lambda(s)} \left\{ \frac{r(r-1)}{n(n-1)} q^*(s,(n-1,m;r-1,k)) \right. \\
&\quad \left. \left. + \left(1-\frac{r(r-1)}{n(n-1)}\right) q^*(s,(n-1,m;r,k)) \right\} \right] g(t,n;s)ds \qquad (9)
\end{aligned}
$$

for $k=1,2,\ldots,r;\ m \geq k-1$.

3.2 The time to MRCA

The second part of our study focusses on the information available in summary statistics of the data D when they are used for ancestral inference. A problem of some current interest in anthropology concerns the estimation of the age and geographical location of the mitochondrial MRCA of humans; Cann et al. [1], Stoneking [2]. In our setting, this issue involves inferences about the conditional distribution of T_{MRCA} for a sample, given the summary statistics, and a comparison of it with the corresponding conditional distribution given the full data set D.

The approach to finding sampling distributions used above can be modified to find the joint distribution of the summary statistic and T_{MRCA}. We illustrate the approach by finding the joint distribution of the number of segregating sites in a sample of n genes taken at time t in the past, and the time from t until the most recent common ancestor of the sample is reached. Let $\tilde{q}(t,(n,m),w)$ be the probability of m segregating sites and a waiting time to the MRCA of at most w. The argument that leads to (8) readily produces the recursion

$$\begin{aligned}\tilde{q}(t,(n,m),w) &= \int_t^\infty \left\{ \frac{\theta}{\theta+(n-1)\lambda(s)}\tilde{q}(s,(n,m-1),t+w-s) \right. \\ &\quad \left. + \frac{(n-1)\lambda(s)}{\theta+(n-1)\lambda(s)}\tilde{q}(s,(n-1,m),t+w-s) \right\} g(t,n;s)ds, \qquad (10)\end{aligned}$$

where $\tilde{q}(t,(n,m),s)=0$ if $m<0$ or $s<0$. The distribution of interest is then

$$\mathbb{P}(T_{MRCA} \le w | S_n = m) = \frac{\tilde{q}(0,(n,m),w)}{\tilde{q}((n,m))}, \ w \ge 0. \qquad (11)$$

For the joint distribution of S_n, K_n and T_{MRCA}, let $q^*(t,(n,m;r,k),w)$ be the probability that a sample of n genes taken at time t in the past has m segregating sites, the subsample of r genes has k distinct alleles, and the time to the MRCA is at most w time units further into the past. Then q^* satisfies an equation of the form (9), with terms on the right side of (9) of the form $q^*(s,(n,m;r,k))$ replaced by $q^*(t,(n,m;r,k),t+w-s)$. The distribution we are interested in is

$$\mathbb{P}(T_{MRCA} \le w | S_n = m, K_n = k) = \frac{q^*(0,(n,m;n,k),w)}{q^*((n,m;n,k))}, \ w \ge 0. \qquad (12)$$

4 Monte Carlo methods

In the case of constant population size, when $\lambda(s) \equiv 1$, the recurrence equations in the previous section simplify considerably because the sampling time t plays no role in the equations. Writing $\mathbb{P}(K_n = k) = q((n,k)) \equiv q(t,(n,k))$, we see from (7) that

$$(\theta+n-1)\mathbb{P}(K_n=k) = (n-1)\mathbb{P}(K_{n-1}=k) + \theta\mathbb{P}(K_{n-1}=k-1),$$

with initial condition $\mathbb{P}(K_1 = 1) = 1$. The solution of this recurrence is well known to be

$$\mathbb{P}(K_n = k) = \frac{\theta^k |S_n^k|}{\theta_{(n)}}, \quad k = 1, 2, \ldots, n,$$

where S_n^k is a Stirling number of the first kind, and $x_{(n)} = x(x+1)\cdots(x+n-1)$; see Ewens [13]. The recursion for $\mathbb{P}(S_n = m) = \tilde{q}((n,m)) \equiv \tilde{q}(t,(n,m))$ reduces to

$$(\theta + n - 1)\mathbb{P}(S_n = m) = (n-1)\mathbb{P}(S_{n-1} = m) + \theta\mathbb{P}(S_n = m-1).$$

This distribution has been studied by Watterson [14], and an explicit formula is known for it; cf. [15]. Griffiths [16] has studied the joint law of K_n and S_n, deriving in particular the constant population size analog of the recurrence in (9), but few explicit results are available.

Although these distributions can be found explicitly (or at least computed simply) in the constant population size case, this seems far from the case for the variable population size case, especially if the sample size n 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 require.

4.1 The basic method

The recursions in (7) – (9) have a common structure that may be written in the form

$$\begin{aligned} q(t,x) &= \int_t^\infty \sum_{y \in \mathcal{A}} r(s;x,y) q(s,y) g(t,x;s) ds \\ &\quad + \int_t^\infty \sum_{y \in \mathcal{B}} r(s;x,y) q(s,y) g(t,x;s) ds, \quad x \in \mathcal{B}, \end{aligned} \tag{13}$$

where $q(t,x)$ is known explicitly (or is easy to compute, perhaps numerically) for $x \in \mathcal{A}$, $r(s;x,y) \geq 0$ and $g(t,x;s)$ is a probability density satisfying $\int_0^\infty g(t,x;s)ds = 1$. For example, in equation (7) the states x are of the form $x = (n,k)$, we can take $\mathcal{A} = \{(1,1)\}$, and for $1 \leq k \leq n, n \geq 2$ we have

$$r(s,x;y) = \frac{\theta}{\theta + (n-1)\lambda(s)}, \quad y = (n-1, k-1)$$

$$= \frac{(n-1)\lambda(s)}{\theta + (n-1)\lambda(s)}, \quad y = (n-1, k),$$

and $g(t, x; s)$ is given by (6). For states of the form $x = (n, 1)$ only the second term applies, since $q(t, (n, 0)) = 0$.

Let $P(s; x, y)$ be a transition probability kernel on the discrete state space $\mathcal{X} = \mathcal{A} \cup \mathcal{B}$ satisfying $\sum_{y \in \mathcal{X}} P(s; x, y) = 1$ for all $s \geq 0, x \in \mathcal{X}$ and

$$P(s; x, y) > 0 \text{ if } r(s; x, y) > 0.$$

P and g determine a non-homogeneous Markov chain $X(\cdot)$ on $\mathcal{X}$ as follows: Given that $X(t) = x \in \mathcal{B}$, the time of the next change of state has density $g(t, x; s)$, and given that this change occurs at time s, the probability that the next state is y is $P(s; x, y)$. We are interested in the process up to the time τ that it reaches the set $\mathcal{A}$. We assume that P has been constructed so that $\mathbb{P}_x(\tau < \infty) = 1$ for all $x \in \mathcal{B}$.

We can rewrite (13) as

$$\begin{aligned} q(t, x) &= \int_t^\infty \sum_{y \in \mathcal{A}} f(t, x; s, y) q(s, y) P(s; x, y) g(t, x; s) ds \\ &\quad + \int_t^\infty \sum_{y \in \mathcal{B}} f(t, x; s, y) q(s, y) P(s; x, y) g(t, x; s) ds, \quad x \in \mathcal{B}, \end{aligned} \tag{14}$$

where

$$f(t, x; s, y) = \frac{r(s; x, y)}{P(s; x, y)}. \tag{15}$$

It is convenient in what follows to include the variable t in the definition of f given above. The term $q(s, y)$ in (14) for $y \in \mathcal{B}$ can be evaluated iteratively, to provide for $x \in \mathcal{B}$

$$\begin{aligned} q(t, x) = & \\ & \int_t^\infty \sum_{y_1 \in \mathcal{A}} f(t, x; s_1, y_1) q(s_1, y_1) P(s_1; x, y_1) g(t, x; s_1) ds_1 \\ & + \int_t^\infty \sum_{y_1 \in \mathcal{B}} f(t, x; s_1, y_1) P(s_1; x, y_1) g(t, x; s_1) \left[\int_{s_1}^\infty \sum_{y_2 \in \mathcal{A}} f(s_1, y_1; s_2, y_2) \right. \\ & \left. q(s_2, y_2) P(s_2; y_1, y_2) g(s_1, y_1; s_2) ds_2 \right] ds_1 \quad + \cdots. \end{aligned} \tag{16}$$

This provides a probabilistic representation of $q(t,x)$. Let $\tau_1 < \tau_2 \cdots < \tau_K = \tau$ be the (random number of) jump times of $X(\cdot)$ until entering $\mathcal{A}$, and define $\tau_0 = t$. Then

$$q(t,x) = \mathbb{E}_{(t,x)} \left(q(\tau, X(\tau)) \prod_{j=1}^{K} f(\tau_{j-1}, X(\tau_{j-1}); \tau_j, X(\tau_j)) \right), \tag{17}$$

where $\mathbb{E}_{(t,x)}$ denotes expectation with respect to $X(t) = x$.

The representation in (17) provides a Markov chain Monte Carlo approximant to $q(t,x)$: Simulate many independent copies of the process $X(\cdot)$ starting from $X(t) = x$, and compute the observed value of the functional under the expectation sign in (17) for each of them. The average of these values is an unbiased estimate of $q(t,x)$, and we may then use standard theory to see how accurately $q(t,x)$ has been estimated. To be useful in practice we must check in each case that the variance of the estimator is finite.

There is a natural candidate for P, obtained by defining

$$\begin{aligned} f(x;s) &= \sum_y r(s;x,y) \\ P(s;x,y) &= \frac{r(s;x,y)}{f(x;s)}, \end{aligned} \tag{18}$$

and then (15) shows that $f(t,x;s,y)$ reduces to

$$f(t,x;s,y) = f(x;s).$$

It is also important in practice, particularly in the context of variance reduction, to have some flexibility in choosing the stopping time τ, or, equivalently, the set $\mathcal{A}$. In the present context it is often possible to calculate probabilities once there are two or three distinct ancestors, rather than tracing the genealogy back to just a single individual. Several examples are given in [6, 8].

It is also clear that this method can be adapted to deal with more complicated problems in which the $X(\cdot)$ process might be semi-Markov. In principle, Markov chain Monte Carlo methods like this date back at least to the late 1940s, where they were used to solve matrix

equations of the form $Ax = b$; Forsythe and Leibler [17]. Halton [18] discusses several variants on this theme. These methods are similar in spirit to the Metropolis algorithm (Metropolis et al. [19]), further developed by Hastings [20]. In that method, the quantity of interest is represented as the mean (under the stationary distribution) of a function of an ergodic Markov chain, and this mean is estimated by computing an ergodic average. This uses a single run of the chain to produce estimates, the observations within the run being correlated. In the present approach we use independent runs of random lengths, which in principle makes the subsequent analysis of the output somewhat simpler.

4.2 Monte Carlo likelihoods

In the general setting, 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. Often we are interested in finding the solution q_Γ for a variety of values of Γ, for example when using q as a likelihood function. One way to do this is to perform several independent simulations of the process controlled by each of the values of Γ. In practice this usually proves to be too time consuming, and we use the following approach based on importance sampling. We 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. Write (12) in the form

$$q_\Gamma(t,x) = \int_t^\infty \sum_{y\in\mathcal{X}} h_{\Gamma,\Gamma 0}(t,x;s,y) P_{\Gamma_0}(s;x,y) q_\Gamma(s,y)\; g_{\Gamma_0}(t,x;s) ds \tag{19}$$

where

$$h_{\Gamma,\Gamma 0}(t,x;s,y) = \frac{f_\Gamma(t,x;s,y) g_\Gamma(t,x;s) P_\Gamma(s;x,y)}{g_{\Gamma_0}(t,x;s) P_{\Gamma_0}(s;x,y)}$$

The representation of $q_\Gamma(t,x)$ is, from (17),

$$q_\Gamma(t,x) = \mathbb{E}_{(t,x)} \left(q_\Gamma(\tau, X(\tau)) \prod_{j=1}^{k} h_{\Gamma,\Gamma_0}(\tau_{j-1}, X(\tau_{j-1}); \tau_j, X(\tau_j)) \right). \tag{20}$$

Estimates of $q_\Gamma(t,x)$ may be now obtained as described above. This method proves to be faster when the cost of producing observations on the process $X(\cdot)$ outweighs the cost

of calculating the functionals in (20). In exchange for this time saving, the estimates for different Γ are no longer independent, but rather they are correlated because of the common generating process. This makes the analysis of the output somewhat more complicated than in the independent replicates case. 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.

Monte Carlo likelihood and Bayesian methods using the Hastings-Metropolis approach are also popular; see Geyer and Thompson [21], Besag and Green [22], Smith and Roberts [23], and Thompson [24] for some examples.

4.3 The distribution of time to MRCA

The quantities required to find the conditional distributions in Section 3.2 satisfy recursions like (10), and these have the common form

$$q(t,x,w) = \int_t^\infty \sum_{y\in\mathcal{X}} r(s;x,y)q(s,y,t+w-s)g(t,x;s)ds. \tag{21}$$

Iterating as in Section 4.1 produces the representation

$$q(t,x,w) = \mathbb{E}_{(t,x)}\left(q(\tau,X(\tau),t+w-\tau)\prod_{j=1}^{k} f(\tau_{j-1},X(\tau_{j-1});\tau_j,X(\tau_j))\right),\ w\geq 0, \tag{22}$$

where f is defined in (15). The natural initial condition in these recursions takes the form

$$q(t,x,w) = I\{w\geq 0\},\ x\in\mathcal{A},$$

where $I\{A\}$ denotes the indicator of the event A. Thus the term $q(\tau,X(\tau),t+w-\tau)$ in (22) reduces to $I\{\tau\leq t+w\}$. It follows that if we simulate the process $X(\cdot)$ R times, and define

$$F_l = \prod_{j=1}^{k_l} f(\tau_{j-1},X(\tau_{j-1});\tau_j,X(\tau_j)), \tag{23}$$

the value of the functional under the expectation sign in (17) for the lth simulation, $l = 1,2,\ldots,R$, then a conditional probability of the form $q(t,x,w)/q(t,x)$ can be approximated

by the ratio

$$\frac{\sum_{l=1}^{R} F_l I\{\tau_l \leq t+w\}}{\sum_{l=1}^{R} F_l},$$

where τ_l is the time the lth simulation hits $\mathcal{A}$. In particular, the conditional distribution function $q(t,x,w)/q(t,x), w \geq 0$ can be approximated by the empirical distribution which jumps a height $F_{(l)}/\sum F_j$ at the point $\tau_{(l)}$, the lth smallest of the simulated values $\tau_1, \ldots, \tau_R$. Conditional moments can be computed in a similar way.

5 Applications

In this section we use the Monte Carlo Markov chain method described in Section 4 to study a particular mitochondrial data set. Our aim is to explore the extent to which using just part of the data set for estimation and inference changes our conclusions. One reason this is important is that it allows us to assess the trade-off between computational complexity (using the Markov chain method on complicated state spaces $\mathcal{X}$) and statistical accuracy (presumed to come from use of the full data set).

5.1 A mitochondrial data set

We begin with a brief description of the data that motivated much of our work. A much more comprehensive description appears in [7], but the present outline should suffice to set the scene. We focus on mitochondrial data sampled from a single North American Indian tribe, the Nuu-Chah-Nulth from Vancouver Island. The original data appeared in Ward et al. [3]. They comprise a random sample of mitochondrial DNA sequences from 63 individuals. Each sequence is the first 360 base pair segment of the control region, comprising 201 pyrimidine sites (bases A or G) and 159 purine sites (bases C or T). 21 of the pyrimidine sites are segregating, i.e. not identical in all 63 sequences in the sample. In contrast, only 5 of the purine sites are segregating. There are 28 distinct DNA sequences in the data. Each site in the data is binary, being either a purine site or a pyrimidine site. Furthermore, because there

Table 1: Mitochondrial DNA sequence data

Position	106	190	251	296	344	88	91	124	149	162	166	194	233	267	271	275	313	339	allele freqs.
Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
allele																			
a	A	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	2
b	A	G	G	A	A	T	C	C	T	T	T	T	C	T	C	T	T	C	2
c	G	A	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	1
d	G	G	A	G	A	C	C	C	C	C	T	T	C	C	C	T	T	C	3
e	G	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	19
f	G	G	G	A	G	T	C	C	T	C	T	T	C	T	C	T	T	C	1
g	G	G	G	G	A	C	C	C	T	C	C	C	C	C	C	T	T	T	1
h	G	G	G	G	A	C	C	C	T	C	C	C	T	C	C	T	T	T	1
i	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	C	T	4
j	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	T	T	8
k	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	C	5
l	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	4
m	G	G	G	G	A	C	C	T	T	C	T	T	C	C	C	T	T	C	3
n	G	G	G	G	A	C	T	C	T	C	T	T	C	C	T	T	T	C	1

Data from Ward et al. ([3], Figure 1). Variable purine and pyrimidine positions in the control region.

is no recombination in mitochondrial DNA, each site in the sample has the same ancestral history.

In [7] we studied one part of the original data that seems to have a relatively simple mutation structure. The subsample we used comprised 55 of the original 63 sequences, and 352 of the original 360 DNA sites. Eight of the pyrimidine segregating sites were removed, resulting in a set of 18 segregating sites in all; 13 of these sites are pyrimidines, and 5 are purines. These data are given in Table 1, reproduced in modified form from [7]. Only the segregating sites are shown, subdivided into sites containing purines and pyrimidines. Each row of the table represents a distinct DNA sequence, and the frequencies of these alleles are given in the right hand column of the table.

In [7] we used the data described in Table 1 to estimate the substitution rate θ in the control region, using purine and pyrimidine sites both separately and combined. We modeled the evolution of the sequences using the infinitely-many-sites model described in Section 2, and based our estimation on the complete tree structure of the data. We also studied the conditional distribution of the time to the MRCA given the full tree structure. In the next

section, we study the analogous problems using just the summary statistics based on the number of alleles and segregating sites seen in the data.

5.2 Constant population size

We argued in [7] that to a first approximation, the population size of the Nuu-Chah-Nulth has been constant for roughly the last 6,000 years. We make the same assumption here, and so take $v(x) = 1$ for all x.

5.2.1 Estimating the substitution rate

We begin by estimating the substitution rate θ using the Monte Carlo likelihood method described in Section 4.2, with the canonical choice of $X(\cdot)$ determined by (18). 50,000 replicates of the $X(\cdot)$ process were made. The results are summarized in Table 2. The table compares estimates based on Watterson's moment method [14] that uses the number of segregating sites, the estimates based on the distribution of the number of segregating sites (determined by (8)), the joint distribution of the number of alleles and segregating sites (determined by (9)), and the maximum likelihood estimate obtained in [7] using the full sequence data. The standard deviations (sd) for the last three rows of the table were found from (estimates of) the observed Fisher information. The rows of the table present estimates in order of what is, intuitively, increasing information. The estimates increase with increasing information content, but have essentially constant standard deviations. This is consistent with recent results of Fu and Li [25], who showed that Watterson's estimator has good variance properties when θ is small. From a practical point of view, it is interesting to note that essentially the same results are obtained by any of the methods when θ is small (as in the purine data), whereas there are more substantial differences when θ is large (as for the combined data set).

Table 2: Estimates of θ

Region	pur			pyr			all		
Estimator	θ	sd	θ/sd	θ	sd	θ/sd	θ	sd	θ/sd
W[1]	1.09	0.58	1.89	2.84	1.12	2.54	3.93	1.44	2.73
S[2]	1.11	0.58	1.93	2.94	1.13	2.61	4.10	1.45	2.83
AS[3]	1.21	0.61	1.98	3.25	1.14	2.85	4.69	1.43	3.28
MLE[4]	1.22	0.61	2.00	3.31	1.14	2.90	4.76	1.48	3.21

[1] Watterson moment estimator using number of segregating sites
[2] Likelihood using number of segregating sites
[3] Likelihood using number of segregating sites and alleles
[4] Likelihood using full data from [7]

5.2.2 Ancestral inference

We shall compare estimates of the distribution of T_{MRCA} conditional on (a) the number of segregating sites; (b) the number of segregating sites and alleles; and (c) the full data. In Table 3 we give the mean and standard deviation of these conditional distributions for the purine and pyrimidine sites combined. These are found by using the approach outlined in Section 4.3. The distribution function was estimated by binning the observations in expression (23), from which the mean and standard deviation were then computed. The values of θ used in the table correspond to the maximum likelihood estimates (plus and minus one standard deviation) presented in Table 2 based on the full data.

It can be seen from Table 3 that for these data both the mean and standard deviation of the conditional distribution decrease with increasing information. Notice in particular that for $\theta = 3.3$, the mean conditional time to the MRCA using just the number of segregating sites is some 60% larger than the corresponding quantity using the full data. Notice also

Table 3: Conditional mean (standard deviations) of T_{MRCA}

θ	Data: none[1]	sites[2]	sites, alleles[3]	all[4]
3.3	1.96 (1.08)	2.22 (0.84)	1.68 (0.57)	1.40 (0.55)
4.8		1.52 (0.56)	1.28 (0.41)	1.20 (0.39)
6.3		1.16 (0.40)	1.04 (0.32)	0.96 (0.12)

[1] Unconditional moments from discussion at (2), $n = 55$.
[2] Uses method in (23) and recursion of form (8)
[3] Uses method in (23) and recursion of form (9)
[4] Results from [7], Table 3.

that when the substitution rate is large, the means tend to be very similar, although the precision decreases dramatically with increasing information.

In Table 4, we give the results of a similar analysis for the purine and pyrimidine data separately. The values of θ used in the table correspond once more to the maximum likelihood estimators given in Table 2. Much the same qualitative conclusions apply as in Table 3.

Finally, we compare the estimates of the conditional distributions themselves, using the purine and pyrimidine sites combined. We estimate these distributions for three values of the substitution rate θ, namely 3.3, 4.8 and 6.3. The unconditional distribution is that of $T_{MRCA} = T_n + \cdots + T_2$ from Section 2.1. The conditional distributions were found by using the method in Section 4.3, and the results are plotted in Figures 1 - 3. The plots confirm the observation that the larger the value of θ, the closer the conditional distributions based on the number of segregating sites and on the number of alleles and segregating sites are to each other, and the smaller the variability in the distributions. Note also that the unconditional distribution has a markedly different shape.

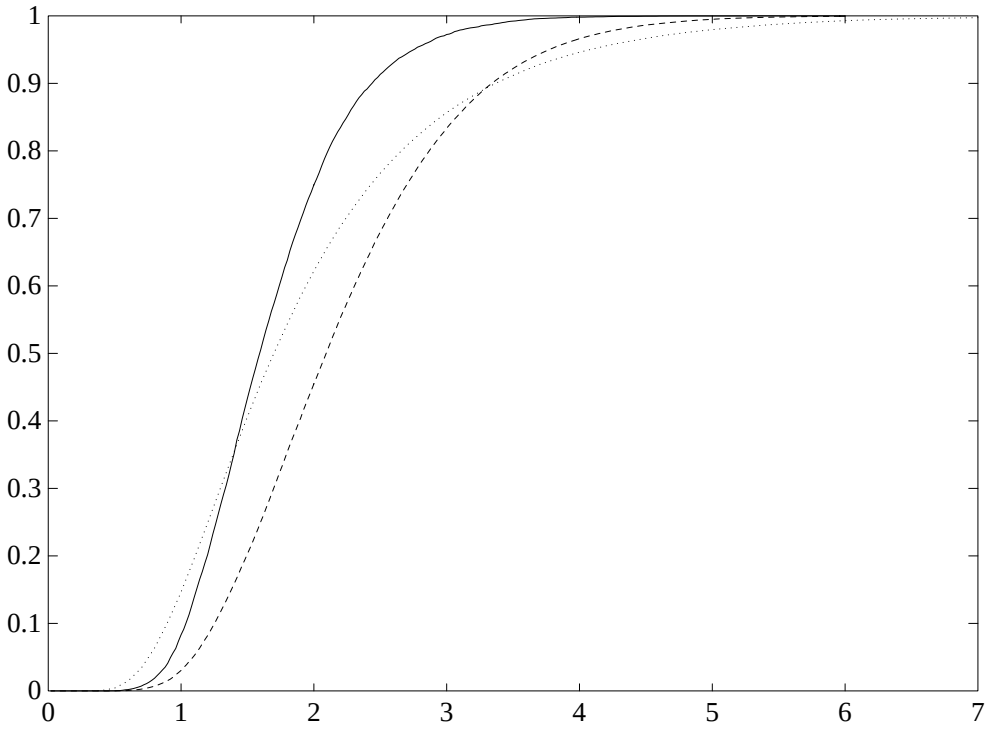

Figure 1: Distribution functions of T_{MRCA}, $\theta = 3.3$

x-axis is time
solid curve: distribution conditional on segregating sites and alleles
dashed curve: distribution conditional on segregating sites
dotted curve: unconditional distribution

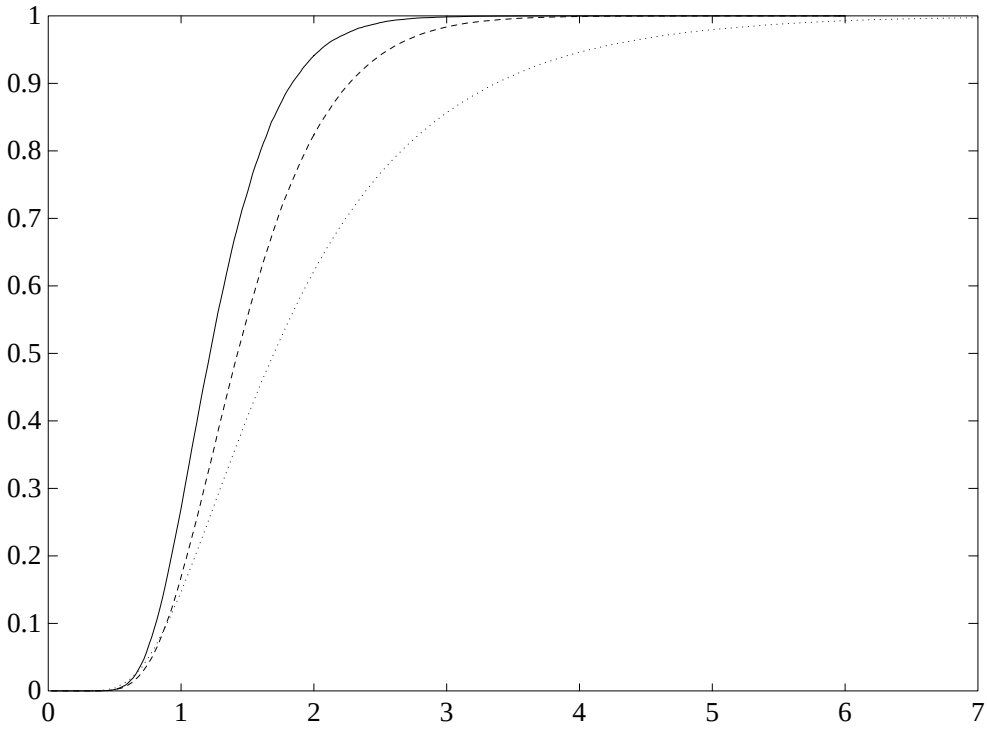

Figure 2: Distribution functions of T_{MRCA}, $\theta = 4.8$

x-axis is time
solid curve: distribution conditional on segregating sites and alleles
dashed curve: distribution conditional on segregating sites
dotted curve: unconditional distribution

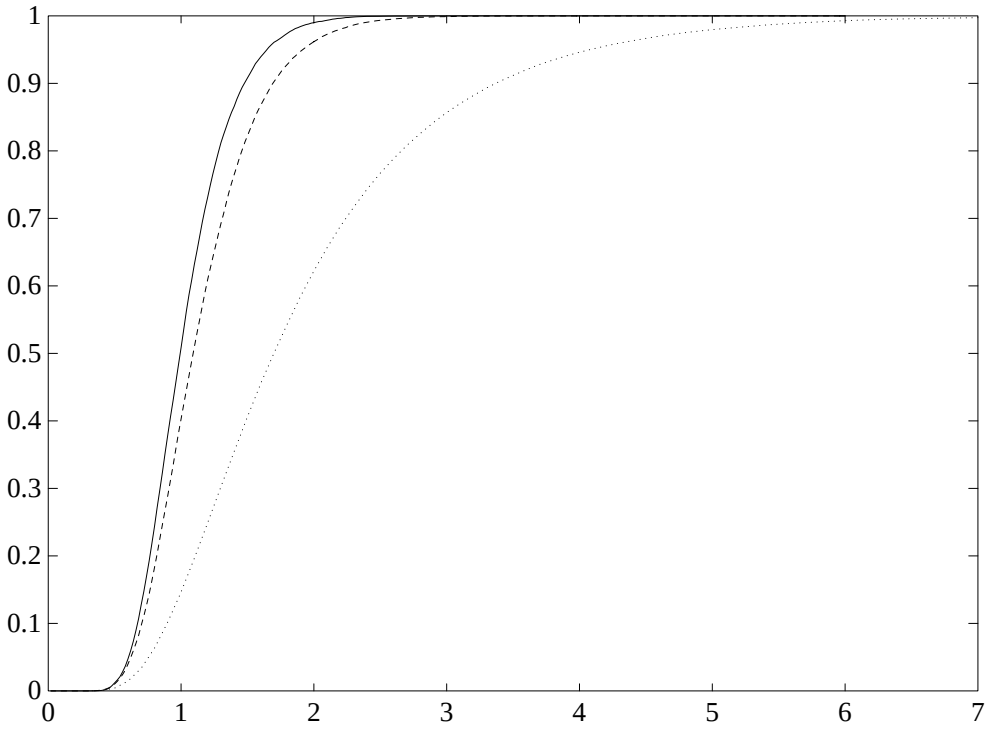

Figure 3: Distribution functions of T_{MRCA}, $\theta = 6.3$

x-axis is time
solid curve: distribution conditional on segregating sites and alleles
dashed curve: distribution conditional on segregating sites
dotted curve: unconditional distribution

Table 4: Conditional mean (standard deviation) of T_{MRCA}

θ	Data: none	sites	sites, alleles	all[1]
1.2[2]	1.96 (1.08)	1.79 (0.83)	1.62 (0.71)	1.54 (0.65)
3.3[3]		1.63 (0.64)	1.42 (0.50)	1.26 (0.41)

[1] results from [7]
[2] $\theta = 1.2$ corresponds to purine sites
[3] $\theta = 3.3$ corresponds to pyrimidine sites

5.3 Variable population size

One of the advantages of the Markov chain Monte Carlo approach advocated here is that it is relatively straightforward to adapt it to study models with variable population size. In Sections 3 and 4 we showed how this could be done for fairly simple summary statistics of the data. In [5] we developed the analogous theory for the full sampling distribution of the data, assuming infinitely-many-alleles, infinitely-many-sites, or finitely-many-sites models for the sequences. We used these results to estimate θ for a given population size function v, and to estimate parameters of v for given substitution rate θ. Rather than focus further on the issue of rate estimation, we turn instead to ancestral inference once more.

5.3.1 Ancestral inference

We do not have a very accurate picture of the contractions and expansions of the population size that the Nuu-Chah-Nulth population had prior to about 6,000 years ago. Rather than make further assumptions in this direction, we shall use simulated data to address the issue of what is learned about the time to the MRCA using the assumption of constant population size, when in fact the population has undergone exponential decrease (looking back into the

Table 5: Simulated sequence data

Site	1	2	3	4	5	6	7	8	9	Allele freqs.
allele										
a	1	1	0	0	0	0	0	0	0	5
b	0	0	0	0	0	0	0	0	0	25
c	1	0	0	0	0	0	0	0	0	12
d	0	0	1	0	0	0	0	0	0	7
e	0	0	0	1	0	0	0	0	0	1
f	1	1	0	0	1	0	0	0	0	1
g	1	0	0	0	0	1	0	0	0	1
h	1	1	0	0	0	0	1	0	0	1
i	1	0	0	0	0	0	0	1	0	1
j	0	0	0	0	0	0	0	0	1	1

Simulated data from infinitely-many-sites model
Substitution rate $\theta = 4.5$, expansion rate 5.0
1 denotes mutant base, 0 denotes ancestral base

past). To do this, we simulated a data set of $n = 55$ genes from the infinitely-many-sites model with exponential growth function given by

$$v(x) = e^{-\beta x}, x \geq 0$$

with $\beta = 5.0$ and $\theta = 4.5$, the value of θ being chosen to reflect the size observed in the mitochondrial data. The data are shown in Figure 5 in a form consistent with that in Table 1.

We compared the estimates of the conditional distribution of T_{MRCA} given the number of segregating sites, the number of segregating sites and alleles, and the full data under two assumptions: (a) no population expansion, and $\theta = 4.5$; and (b) exponential population expansion with $\beta = 5.0$ and $\theta = 4.5$. This should provide a rough assessment of how

ancestral inference based on models that ignore population expansion compare to analyses that include the possibility of expansion.

In Table 6 we give the summary statistics of the analyses. The row for $\beta = 0.0$ gives the results for the analysis based on the assumption of no population expansion. The same qualitative behavior as in Table 3 is observed, the mean and standard deviation decreasing with increasing information. The conditional mean time is much larger than in the case that assumes exponential population expansion. There is a heuristic explanation of why this behavior might be anticipated: The effect of the expansion is to drastically decrease the unconditional time to the MRCA. Assuming a constant population size has the effect of reducing the substitution rate. Hence in assuming that $\theta = 4.5$, the fixed population size analysis is taking the substitution rate too small, and hence the conditional mean will tend to be too big.

The row for $\beta = 5.0$ gives the corresponding result for the analysis assuming exponential population expansion (the same as used to generate the synthetic data). In this case the means and standard deviation are essentially identical in all cases, presumably because the data are 'typical' of the model being used.

In Figures 4 – 6, we present some more detailed information about the conditional distributions. Figure 5 shows that the conditional distributions and the unconditional distribution are essentially identical. In this case, inference based on the distribution of the number of segregating sites does as well as the distribution based on the full distribution.

In practice, of course, we do not know the values of substitution rates and population expansion parameters. Further simulation studies of the correspondence between conditional distributions based on estimated parameters and the conditional distribution for the parameters that simulated the data are underway.

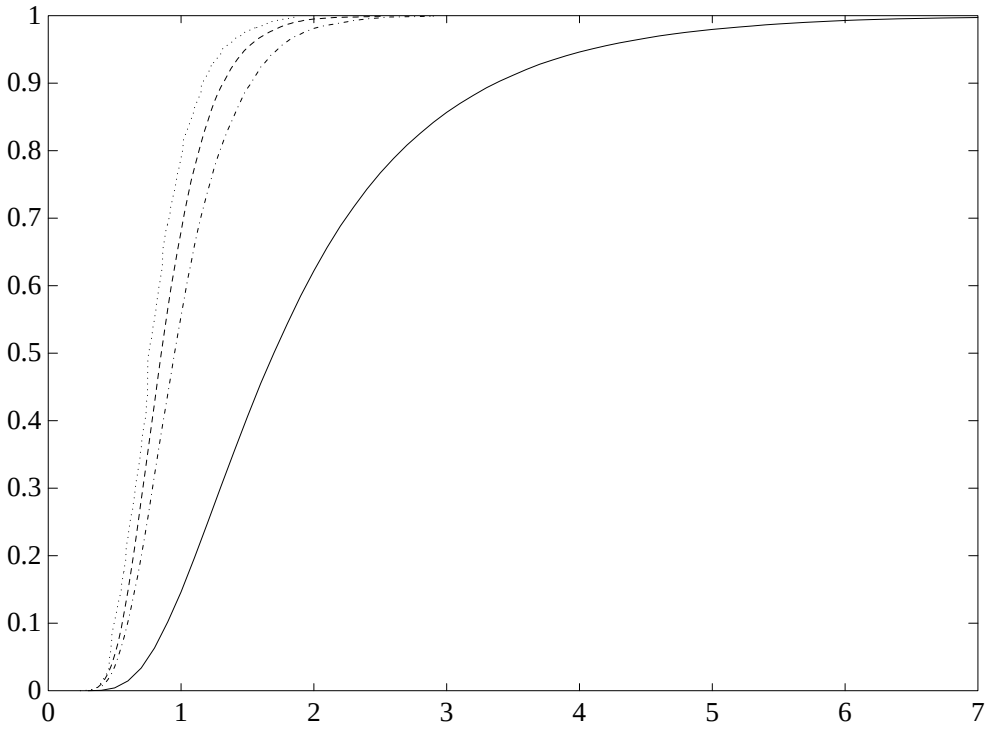

Figure 4: Distribution functions of T_{MRCA}, $\beta = 0.0$

x-axis is time
dotted curve: distribution conditional on full data
dashed curve: distribution conditional on segregating sites and alleles
dotdash curve: distribution conditional on segregating sites
solid curve: unconditional distribution

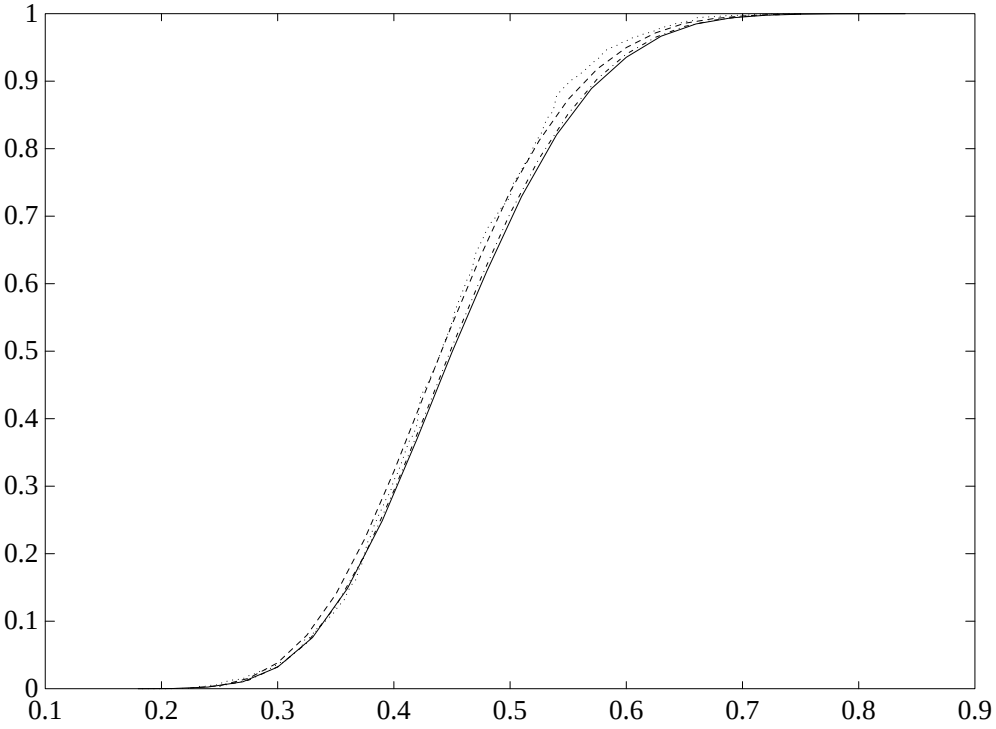

Figure 5: Distribution functions of T_{MRCA}, $\beta = 5.0$

x-axis is time
dotted curve: distribution conditional on full data
dashed curve: distribution conditional on segregating sites and alleles
dotdash curve: distribution conditional on segregating sites
solid curve: unconditional distribution

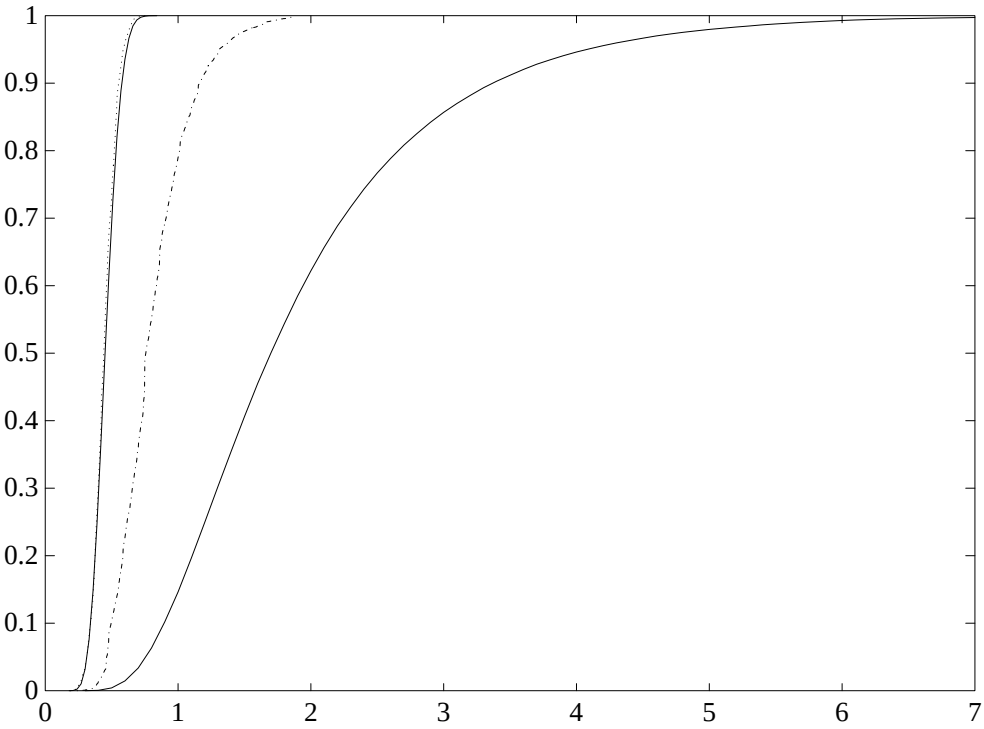

Figure 6: Distribution functions of T_{MRCA}, combined

x-axis is time
solid curves: unconditional distribution, from Figure 4 (right-hand curve), from Figure 5 for left-hand curve.
dotted curve: distribution conditional on full data ($\beta = 5.0$)
dotdash curve: distribution conditional on full data ($\beta = 0.0$)

Table 6: Conditional mean (standard deviation) of T_{MRCA}

β	Data: none	sites	sites, alleles	all[1]
0.0[2]	1.96 (1.08)	1.02 (0.38)	0.90 (0.31)	0.81 (0.27)
5.0[3]	0.46 (0.09)	0.45 (0.09)	0.45 (0.09)	0.45 (0.08)

[1] Uses theory from [5, 7].
[2] Analysis based on assumption that there was no population expansion.
[3] Analysis based on assumption that there was population expansion.

6 Discussion

The computational method for solving recursive systems described in this paper has proved extremely useful for addressing issues that have previously proved intractable by explicit methods. Perhaps its most compelling feature is that it is entirely generic; it is simple to adapt to many different problems. The advantage of this approach over more conventional numerical methods becomes more pronounced as the complexity of the recursion increases. Several examples arising in population genetics are cited in the paper.

These schemes differ from the 'usual' Markov chain Monte Carlo (MCMC) methods like Hastings-Metropolis in that the process being simulated is not ergodic, and there are no 'start-up' problems in ascertaining when stationarity is deemed to hold. However, the technique shares several common features with other MCMC methods: variants on the theme are among the oldest of Monte Carlo methods, variance reduction and time reduction are important practical issues, and the method may be extended to generate approximants to complete likelihood surfaces from a single generating process.

7 References

1. R.M. Cann, M. Stoneking, and A. Wilson, Mitochondrial DNA and human evolution, *Nature* **325**, 31-36 (1987).

2. M. Stoneking, DNA and recent human evolution, *Evolutionary Anthropology* **2**, 60-73 (1993).

3. R.H. Ward, B.L. Frazier, K. Dew, and S. Pääbo, Extensive mitochondrial diversity within a single Amerindian tribe, *Proc. Natl. Acad. Sci. USA* **88**, 8720-8724 (1991).

4. G.F. Shields, A.M. Schmiechen, B.L. Frazier, A. Redd, M.I. Voevoda, J.K. Reed, and R.H. Ward, mtDNA sequences suggest a recent evolutionary divergence for Beringian and Northern North American populations, *Am. J. Hum. Genet.* **53**, 549-562 (1993).

5. R.C. Griffiths and S. Tavaré, Sampling theory for neutral alleles in a varying environment, *Phil. Trans. R. Soc. Lond. B* **344**, 403-410 (1994).

6. R.C. Griffiths and S. Tavaré, Unrooted genealogical tree probabilities in the infinitely-many-sites model, *Math. Biosci.* **126**, in press (1995).

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

8. R.C. Griffiths and S. Tavaré, Simulating probability distributions in the coalescent, *Theoret. Popn. Biol.* **46**, 131-159 (1994).

9. J.F.C. Kingman, On the genealogy of large populations, *J. Appl. Prob.* **19A**, 27-43 (1982).

10. R.R. Hudson, Gene genealogies and the coalescent process, In *Oxford Surveys in Evolutionary Biology*, Volume 7. Edited by D. Futuyma and J. Antonovics, 1-44 (1991).

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

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

13. W.J. Ewens, The sampling theory of selectively neutral alleles, *Theoret. Popn. Biol.* **3**, 87-112 (1972).

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

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

16. R.C. Griffiths, The number of alleles and segregating sites in a sample from the infinite-alleles model, *Adv. Appl. Prob.* **14**, 225-239 (1982).

17. G.E. Forsythe and R.A. Leibler, Matrix inversion by the Monte Carlo method, *Math. Comp.* **26**, 127-129 (1950).

18. J.H. Halton, A retrospective and prospective study of the Monte Carlo method, *SIAM Review* **12**, 1-63 (1970).

19. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, Equations of state calculations by fast computing machines, *J. Chem. Phys.* **21**, 1087-1092 (1953).

20. W.K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97-109 (1970).

21. C.J. Geyer and E.A. Thompson, Constrained Monte Carlo maximum likelihood for dependent data (with discussion), *J. R. Statist. Soc.* **54**, 657-700 (1992).

22. J. Besag and P.J. Green, Spatial statistics and Bayesian computation, *J. R. Statist. Soc. B* **55**, 25-38 (1993).

23. A.F.M. Smith and G.O. Roberts, Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *J. R. Statist. Soc. B* **55**, 3-24 (1993).

24. E.A. Thompson, Monte Carlo likelihood in genetic mapping, *Statistical Science* **9**, 355-366 (1994).

25. Y.-X. Fu and W.-H. Li, Maximum likelihood estimation of population parameters, *Genetics* **134**, 1261-1270 (1993).