COMPUTATIONAL METHODS FOR THE COALESCENT*

ROBERT C. GRIFFITHS† AND SIMON TAVARÉ‡

Abstract. This paper describes recent work on computational methods for the coalescent. We show how integro-recurrence relations for sampling distributions and related quantities may be solved by a simple Markov chain Monte Carlo method. We describe the method in the context of the coalescent process for a population that is evolving according to a deterministic population size function. The usual constant population size models appear as a special case of this approach. One of the appealing features of the approach is its generic nature: many apparently different problems may be attacked with this one approach. A wide variety of examples are discussed, among them maximum likelihood estimation of parameters.

Key words. Coalescent process, Markov chain Monte Carlo, Population genetics, Sampling distributions, Variable population size.

AMS(MOS) subject classifications. 60G35, 92A05, 92A10.

1. Introduction. Kingman's introduction of the coalescent in 1982 [15] [16] has had the effect of focusing attention on the role played by genealogy in the evolution of populations. Coalescent arguments are now standard in the population genetics literature, and many novel applications and extensions continue to be found. For recent reviews of a range of applications, see [13], [14] and [26] for example. One important feature of this genealogical approach is that it provides a simple way to simulate the behavior of samples taken from populations undergoing very complicated mutation mechanisms, usually without having to simulate the structure of the entire population. In general terms, the idea is to generate the genealogy of the sample back to the common ancestor, and then simulate the effects of mutation down the ancestral tree from this most recent common ancestor to the individuals in the sample. This approach is very useful for studying the statistical behavior of allelic configurations in a sample.

In contrast, we discuss methods that intrinsically use the mutation process in the other direction, from the individuals in the sample back to the common ancestor. This approach has proven useful for computing the probability that a sample of genes has a *particular* allelic configuration. Such sampling probabilities, epitomized by the Ewens Sampling Formula [4], play an important role in the statistical analysis of genomic variability, as they form the basis of a likelihood approach to inference in the coalescent.

Our approach to sampling probabilities and related quantities is via the

* Supported in part by Australian Research Grant A19131517, and NSF grant DMS 90-05833.

† Mathematics Department, Monash University, Clayton 3168, Australia.

‡ Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, CA 90089-1113, USA.

derivation of certain integro-recurrence equations that they satisfy. These recursions are determined by what happens as we look back up the ancestral tree towards the root. Either we see a coalescence event or a mutation (or perhaps a recombination) event, and these different possibilities lead to a recursive formula for the sampling probability. These recursions are usually difficult to solve, either explicitly or by conventional numerical analysis techniques. We have developed a Markov chain Monte Carlo technique by which the solutions to such recursions can be approximated. The idea, typical of Monte Carlo methods, is to represent the quantity of interest as the mean of a functional of a non-homogeneous Markov chain, and to estimate this mean by repeated simulations of the chain.

There are many variants on this theme, among them a surface simulation technique that uses a single Markov chain to generate solutions of recurrences with different parameters (this being particularly useful to compute Monte Carlo approximants to likelihood surfaces), and a version that can be used to solve non-homogeneous recursions. We discuss these issues, and give a variety of examples, later in the paper.

The setting for the subsequent development is the coalescent evolving with a deterministically varying population size [16] [25]. We present the structure of this process as a deterministic time change of the usual coalescent, and use this representation to derive the appropriate sampling distributions. The results extend those in [8]. In particular, we show how the method can be used to study models for DNA sequence data in which the stationary distribution of a sequence is arbitrary, for example a high-order Markov measure. This extends earlier work in [6].

2. The coalescent. The ancestry of the genes 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 [15] [16] as an approximation, valid in the limit of large population size, to the ancestral structure of a wide variety of selectively neutral reproduction models, including the Wright-Fisher model. Specifically, assume that a random sample of n genes is taken from a large random mating population with non- overlapping generations that have been of constant size N genes for a long time. Label the sampling generation as generation 0, and let $A_n^N(r)$ be the number of distinct ancestors the sample has r generations into the past. For the Wright-Fisher model, which can be thought of as the case where genes choose their parent genes uniformly and at random, Kingman [15] showed that as $N \to \infty$ the process $\{A_n^N(\lfloor Nt \rfloor), t \geq 0\}$ converges in distribution to a Markov pure death process $\{A_n(t), t \geq 0\}$ on the integers $n, n-1, \ldots, 1$. $A_n(\cdot)$ moves from state k to $k-1$ at rate $k(k-1)/2$.

The same limiting process applies to many other neutral models of reproduction, as long as the genealogy does not collapse or expand too fast [15]. Let σ^2 being the (limiting) variance of the number of offspring of a

typical gene, and suppose $0 < \sigma^2 < \infty$. When time is scaled in units of $\sigma^{-2}N$ generations, then the process $\{A_n^N(\lfloor \sigma^{-2}Nt \rfloor), t \geq 0\}$ converges in distribution to $\{A_n(t), t \geq 0\}$.

The coalescent may be thought of as a random tree, with leaves representing the n sample genes, and vertices where ancestral lines join. In this continuous-time approximation the tree is binary, and the topology of the tree is obtained by randomly merging pairs of individuals. The root of the tree is the most recent common ancestor (MRCA) of the sample of genes. Let T_j be the amount of time the sample has j distinct ancestors, for $j = n, n-1, \ldots, 2$. 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

$$\mathbb{E}T_{MRCA} = 2\left(1 - \frac{1}{n}\right). \tag{2.1}$$

The coalescent may be modified to account for the effects of stochastic or deterministic variation in the population size [16]. In the determinsitic case, suppose that the population is of size $M(0) \equiv N$ genes at the time of sampling, and is of size $M(r)$ in the rth generation before sampling. Let $\sigma^2(r)$ be the variance of the number of offspring born to an individual in generation r. We concentrate here on fluctuations in the population size that are of order N. We assume that there is a relative size function v, and a variance function τ^2 such that for all $x \geq 0$

$$v(x) \equiv \lim_{N\to\infty} \frac{M(\lfloor Nx \rfloor)}{M(0)} > 0,$$

and

$$\tau^2(x) = \lim_{N\to\infty} \sigma^2(\lfloor Nx \rfloor).$$

For the Wright-Fisher model, $\tau^2(x) \equiv 1$. Define the population size intensity function Λ and its density λ by

$$\Lambda(t) = \int_0^t \frac{\tau^2(x)}{v(x)}dx, \quad \lambda(t) = \frac{\tau^2(t)}{v(t)}, \quad t > 0. \tag{2.2}$$

With time measured in units of N generations, we may once more approximate the distribution of the ancestral process by a continuous time non-homogeneous death process $\{A_n^v(t), t \geq 0\}$, whose structure is most simply defined as a deterministic time change of the process $A_n(\cdot)$:

$$A_n^v(t) = A_n(\Lambda(t)), \ t \geq 0. \tag{2.3}$$

All of the properties of the variable-size process can be calculated using the representation in (2.3). For example, this shows immediately that if

$\Lambda(\infty) = \infty$ the sample may be traced back to a common ancestor with probability one. It also defines the joint distribution of the times $T_j, j = n, n-1, \ldots, 2$ (which are no longer independent), and so allows us to study properties of T_{MRCA}. These distributions are somewhat unmanageable, but they are very simple to simulate. Let $S_n < S_{n-1} < \cdots < S_2$ be the times at which $A_n^v(\cdot)$ moves to $n-1, n-2, \ldots, 1$, and set $S_{n+1} \equiv 0$. The sequence $S_n, S_{n-1}, \ldots, S_2$ is Markovian, and the coupling in (2.3) shows that

$$\Lambda(S_j) - \Lambda(S_{j+1}) \stackrel{\mathrm{d}}{=} E_j, j = n, n-1, \ldots, 2$$

where the E_j are independent exponential random variables with parameter $j(j-1)/2$. To simulate the sequence of jump times, we need only

(i) Set $s = 0, j = n$
(ii) Generate E_j exponential $j(j-1)/2$
(iii) Solve for t the equation $\Lambda(t) - \Lambda(s) = E_j$
(iv) Set $S_j = t, s = t, j = j-1$
(v) If j is greater than 1, go to step (ii).

The topology of the ancestral tree in the variable population size case is formed just as before: merge a random pair of individuals at each of the times $S_n, \ldots, S_2$.

3. The effects of mutation. The effects of mutation are superimposed on the ancestral tree of the sample. These mutations occur according to Poisson processes of rate $\theta/2$ along each edge of the tree, the processes in different edges being conditionally independent given the length of the edges. As a limit from the Wright-Fisher model, if u is the probability of a mutation per gene per generation, then $\theta = \lim_{N\to\infty} 2Nu$, where N is the size of the population of genes from which the sample was taken.

The effects of each mutation may be modeled in many different ways. For example, to describe the evolution of a sample of DNA sequences in which the sites in the sequence are completely linked, it is convenient to consider a general mutation scheme in which there are d possible types of gene, labeled $1, 2, \ldots, d$. When a mutation arises in a lineage, a transition is made from type i to j according to the entry p_{ij} in a transition matrix P. It is convenient to allow entries on the diagonal of P to be non-zero, thereby providing for different overall mutation rates for the different types. The mutation rates in the model are uniquely determined by the generator matrix

$$R = (r_{ij}) \equiv \frac{\theta}{2}(P - I). \tag{3.1}$$

The configuration of types in the sample is determined by the mutations in the tree from the root to the leaves. It is usually assumed that P is a regular matrix with a stationary distribution π. If the common ancestor is chosen from a stationary population, then her type has the distribution π.

3.1. Mutation models for DNA sequences. If the sequences are of length s, and the alphabet of bases at each site has a elements (typically $a = 2$ or 4), then $d = a^s$. Types are denoted by sequences $\boldsymbol{i} = (i_1, \ldots, i_s)$ with entries in $[a] \equiv \{1, 2, \ldots, a\}$. We assume that there is no recombination between sites.

The standard model assumes that mutations cause just a single base change, site l being chosen with probability $h_l > 0$, $l = 1, 2, \ldots, s$. The lth site has transition matrix M_l, with stationary distribution $\boldsymbol{\pi}^l$. The mutation matrix P is then given by

$$P = \sum_{l=1}^{s} h_l I \otimes I \otimes \cdots \otimes M_l \otimes \cdots \otimes I \,, \tag{3.2}$$

where $\otimes$ denotes direct product, I is the identity matrix, and

$$\boldsymbol{\pi} = \boldsymbol{\pi}^1 \otimes \cdots \otimes \boldsymbol{\pi}^s.$$

This model allows for variable rates at different sites (the overall rate at site l being $\theta h_l/2$), and for arbitrary substitution probabilities at each site. Nonetheless, the stationary distribution of this mutation mechanism corresponds to independent trials, so that a single gene sampled at random from a stationary population will appear to have independent sites. In its simplest form, the model takes $h_l \equiv 1/s$ and $M_l \equiv M$, so that there are no mutational hotspots, and a gene sampled at stationarity appears to have i.i.d. sites.

For many gene regions, this independence of sites feature is clearly violated (cf. [1], [28]), and a more complex model of the substitution process is required to fit observed data. These authors note that many DNA sequences exhibit a Markovian structure. With this in mind, we describe a simple model, explored in more detail in Tavaré [27], that may be arranged to have an arbitrary stationary measure. To illustrate, suppose that we require the stationary measure $\boldsymbol{\pi}$ to be a Markov measure, that is for each $\boldsymbol{i} = (i_1, \ldots, i_s)$

$$\pi(\boldsymbol{i}) = \mu(i_1) \prod_{j=2}^{s} t(i_{j-1}, i_j) \tag{3.3}$$

where $T = (t(l, m), 1 \le l, m \le a)$ is a strictly positive stochastic matrix, and $(\mu(l), 1 \le l \le a)$ is a probability distribution.

Suppose that the sequence is currently of type $\boldsymbol{i}$. A potential mutation changes a sequence of type $\boldsymbol{i}$ to a sequence of type $\boldsymbol{j}$ with probability $p(\boldsymbol{i}, \boldsymbol{j})$ determined by (3.2). Thus the potential mutant sequence differs from its parent in at most a single coordinate. The net effect of this potential mutation depends, however, on the bases at neighboring sites. For two sequences $\boldsymbol{i}$ and $\boldsymbol{j}$ that differ at a single coordinate, define

$$h(\boldsymbol{i}, \boldsymbol{j}) = \frac{\pi(\boldsymbol{j}) p(\boldsymbol{j}, \boldsymbol{i})}{\pi(\boldsymbol{i}) p(\boldsymbol{i}, \boldsymbol{j})} \wedge 1.$$

The mutation mechanism then makes a mutant of type j with probability $h(i,j)$, and otherwise makes a 'mutant' of type i, the original type.

When the mutation matrices M_l are identical at each site with $M_l \equiv M = (m(i,j))$, the stationary distribution of this model is π determined by (3.3). To check this, we need only observe that this construction is a variant of Hasting's algorithm [12], familiar to Markov chain Monte Carlo enthusiasts. The form of $h(i,j)$ simplifies considerably for the present example. Simple algebra shows that when i and j differ in just the lth coordinate

$$h(i,j) = \begin{cases} \dfrac{m(j_1,i_1)\mu(j_1)t(j_1,i_2)}{m(i_1,j_1)\mu(i_1)t(i_1,i_2)}, & l=1 \\ \dfrac{m(j_l,i_l)t(i_{l-1},j_l)t(j_l,i_{l+1})}{m(i_l,j_l)t(i_{l-1},i_l)t(i_l,i_{l+1})}, & 2 \le l \le s-1 \\ \dfrac{m(j_s,i_s)t(i_{s-1},j_s)}{m(i_s,j_s)t(i_{s-1},i_s)}, & l=s \end{cases}$$

Thus a mutation mechanism determined by the bases at sites adjacent to the target site can readily produce a Markov dependent stationary distribution. Clearly, this scheme can be generalized in many ways to produce stationary measures of great complexity, for example those with high order Markov dependence, and those with non-homogeneous Markov structure.

In the next section, we show how the distribution of the allelic configuration of a sample of genes undergoing such a mutation mechanism in a population of deterministically varying size can be calculated.

3.2. Sampling distributions. In this section, we return to our original labeling of types as $1,2,\ldots,d$, with mutation matrix $P = (p_{ij})$. Let $q(t,\boldsymbol{n})$ be the probability that a sample of n genes taken at time t in the past has a type configuration of $\boldsymbol{n} = (n_1,\ldots,n_d)$, where n_i is the number of copies of type i in the sample. The fundamental integro-recurrence relation for $q(t,\boldsymbol{n})$ is derived by considering the configuration of genes at the time of the first event (either a coalescence or a mutation) in the ancestry of the sample prior to time t. The time W_t of this first event has distribution determined by

$$\mathbb{P}(W_t > s) = \exp\left(-\int_t^s \gamma(u,\boldsymbol{n})du\right), \quad s \ge t, \tag{3.4}$$

where

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

The integro-recurrence for the sampling formula takes the form

$$q(t,\boldsymbol{n}) = \int_t^\infty \left\{ \frac{n\theta}{2\gamma(s,\boldsymbol{n})} \sum_{\substack{i,j\in[d] \\ n_j>0, i\neq j}} \frac{n_i+1}{n} p_{ij} q(s,\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j) \right.$$

$$+\frac{n\theta}{2\gamma(s,\boldsymbol{n})}\sum_{i\in[d]}\frac{n_i}{n}p_{ii}q(s,\boldsymbol{n})$$

$$+\frac{n(n-1)\lambda(s)}{2\gamma(s,\boldsymbol{n})}\sum_{j\in[d],n_j>0}\frac{n_j-1}{n-1}q(s,\boldsymbol{n}-\boldsymbol{e}_j)\Bigg\}g(t,\boldsymbol{n};s)ds, \tag{3.5}$$

where $\{\boldsymbol{e}_i\}$ are the d unit vectors, and

$$g(t,\boldsymbol{n};s)=\gamma(s,\boldsymbol{n})\exp\left(-\int_t^s\gamma(u,\boldsymbol{n})du\right)$$

is the density of W_t. Boundary conditions are required to determine the solution to (3.5). These have the form

$$q(t,\boldsymbol{e}_i)=\pi_i^*,\ i=1,\ldots,d, \tag{3.6}$$

where π_i^* is the probability that the most recent common ancestor is of type i. It is often assumed that

$$\pi_i^*=\pi_i,\ i=1,\ldots,d, \tag{3.7}$$

where $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_d)$ is the stationary distribution of P. Of particular interest is the solution $q(\boldsymbol{n})\equiv q(0,\boldsymbol{n})$.

To derive (3.4), we first calculate the probability of no coalescence events in a sample of size n in time (t,s). This event has probability

$$\exp\left(-\frac{n(n-1)}{2}(\Lambda(s)-\Lambda(t))\right).$$

Given no coalescences in (t,s), the conditional probability of no mutations in (t,s) is just the chance that no mutations occur on the n branches of the ancestral tree in time (t,s). This is

$$\exp\left(-\frac{n\theta}{2}(s-t)\right).$$

To verify (3.5) suppose that the first event prior to time t occurred at time $W_t=s$. The relative rates of mutation and coalescence for the n genes are $n\theta/2:n(n-1)\lambda(s)/2$, so the probability that the event at time s is a mutation is $n\theta/2\gamma(s,\boldsymbol{n})$. To obtain a configuration of $\boldsymbol{n}$ after a mutation the configuration at time $s+$ must be either $\boldsymbol{n}$, and a transition $i\to i$ takes place for some $i\in[d]$ (the mutation resulted in no observable change), or $\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j$, $i,j\in[d],n_j>0,i\neq j$ and a transition $i\to j$ takes place. On the other hand, the probability that the event at s is a coalescence is $n(n-1)/2\gamma(s,\boldsymbol{n})$. To obtain a configuration $\boldsymbol{n}$ the configuration must be $\boldsymbol{n}-\boldsymbol{e}_j$ for some $j\in[d]$ with $n_j>0$ and the ancestral lines involved in the

coalescence must be of type j. Averaging over the density of W_t produces (3.5).

An alternative (and equivalent) integro-recurrence relation for $q(t, \boldsymbol{n})$ can be derived by considering the configuration of genes at the time of the first event *that changes the configuration* in the ancestry of the sample prior to time t. The time W_t^* of this first event has distribution determined by

$$\mathbb{P}(W_t^* > s) = \exp\left(-\int_t^s \gamma^*(u, \boldsymbol{n})du\right), \quad s \geq t,$$

where

$$\gamma^*(u, \boldsymbol{n}) = \frac{n(n-1)}{2}\lambda(u) + \frac{n\theta}{2}\left(1 - \sum_{i \in [d]} \frac{n_i}{n} p_{ii}\right). \tag{3.8}$$

The corresponding recurrence is

$$\begin{aligned} q(t, \boldsymbol{n}) &= \int_t^\infty \left\{ \frac{n\theta}{2\gamma^*(s, \boldsymbol{n})} \sum_{\substack{i,j \in [d] \\ n_j > 0, i \neq j}} \frac{n_i + 1}{n} p_{ij} q(s, \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_j) \right. \\ &\left. + \frac{n(n-1)\lambda(s)}{2\gamma^*(s, \boldsymbol{n})} \sum_{j \in [d], n_j > 0} \frac{n_j - 1}{n-1} q(s, \boldsymbol{n} - \boldsymbol{e}_j) \right\} g^*(t, \boldsymbol{n}; s)ds, \end{aligned} \tag{3.9}$$

where g^* is the density of W_t^*. This equation is somewhat simpler to solve than (3.5), and so may be preferred in practice.

To see that the solutions to (3.5) and (3.9) are the same, note that (3.5) is equivalent to the differential equation

$$\begin{aligned} \frac{dq(u, \boldsymbol{n})}{du} &= \gamma(u, \boldsymbol{n})q(u, \boldsymbol{n}) - \frac{n\theta}{2} \sum_{i \in [d]} \frac{n_i}{n} p_{ii} q(u, \boldsymbol{n}) \\ &\quad - \frac{n\theta}{2} \sum_{\substack{i,j \in [d] \\ n_j > 0, i \neq j}} \frac{n_i + 1}{n} p_{ij} q(u, \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_j) \\ &\quad - \frac{n(n-1)\lambda(u)}{2} \sum_{j \in [d], n_j > 0} \frac{n_j - 1}{n-1} q(u, \boldsymbol{n} - \boldsymbol{e}_j). \end{aligned}$$

Both (3.5) and (3.9) can be found as the solution of this differential equation with integrating factors $\exp(-\int_0^s \gamma(u, \boldsymbol{n})du)$ and $\exp(-\int_0^s \gamma^*(u, \boldsymbol{n})du)$, respectively.

When the population size is constant through time, equations (3.5) and (3.9) reduce to a discrete recurrence for the configuration probability

$q(\boldsymbol{n}) \equiv q(t, \boldsymbol{n})$:

$$q(\boldsymbol{n}) = \frac{\theta}{n+\theta-1}\left\{\sum_{\substack{i,j\in[d]\\ n_j>0, i\neq j}} \frac{n_i+1}{n} p_{ij} q(\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j) + \sum_{i\in[d]} \frac{n_i}{n} p_{ii} q(\boldsymbol{n})\right\}$$

$$+ \frac{(n-1)}{\theta+n-1} \sum_{j\in[d], n_j>0} \frac{n_j-1}{n-1} q(\boldsymbol{n}-\boldsymbol{e}_j). \tag{3.10}$$

This recursion has been studied in various forms by several authors, among them Sawyer, Dykhuizen, and Hartl (1987), and Lundstrom (1990). Given $\{q(\boldsymbol{m}); m < n\}$, simultaneous equations for the $\binom{n+d-1}{d-1}$ unknown probabilities $\{q(\boldsymbol{m}); m = n\}$ are non-singular, and in theory can be solved. In practice a numerical solution is difficult because of the large number of equations, and the situation is considerably more difficult for the recursion in (3.5). In the next section, we describe the Markov chain Monte Carlo approach that we have used to solve systems like (3.5), and describe some of the applications of the technique.

4. Markov chain Monte Carlo. The recursions for sampling probabilities, typified by (3.5) and (3.9), have a common structure. Let $\mathcal{X}$ denote the discrete set of states of the recursion. In (3.5) for instance, $\mathcal{X}$ is the set of d-dimensional vectors $\boldsymbol{n} = (n_1, \ldots, n_d)$ with nonnegative integer entries and sum $m = 1, 2, \ldots, n$, n being the size of the sample. The recursions may be written in the form

$$q(t,x) = \int_t^\infty \sum_{y\in\mathcal{A}} r(s;x,y) q(s,y) g(t,x;s) ds$$

$$+ \int_t^\infty \sum_{y\in\mathcal{B}} r(s;x,y) q(s,y) g(t,x;s) ds, \quad x \in \mathcal{B}, \tag{4.1}$$

where $q(t,x)$ is known explicitly for $x \in \mathcal{A}$, $r(s;x,y) \geq 0$, $g(t,x;s)$ is a probability density satisfying $\int_0^\infty g(t,x;s)ds = 1$, and $\mathcal{X} = \mathcal{A} \cup \mathcal{B}$.

For the sampling probabilities determined by (3.9) we have $x = \boldsymbol{n}, n = \sum_{i=1}^d n_i, y = \boldsymbol{m}$, and the non-zero entries of the kernel r in (4.1) are

$$r(s;\boldsymbol{n},\boldsymbol{m}) = \frac{\theta}{2\gamma^*(s,\boldsymbol{n})}(n_i+1)p_{ij}, \quad \boldsymbol{m} = \boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j,\ i,j\in[d],\ n_j>0,\ i\neq j$$

$$= \frac{n\lambda(s)}{2\gamma^*(s,\boldsymbol{n})}(n_j-1), \quad \boldsymbol{m} = \boldsymbol{n}-\boldsymbol{e}_j,\ j\in[d],\ n_j>0,$$

where γ^* is given in (3.8).

These recursions are typically impossible to solve explicitly except perhaps for very small sample sizes. Standard numerical solutions are often

not practicable either because of the enormous dimension of $\mathcal{X}$ or the difficulty in evaluating multiple integrals. Instead, we have developed a Markov chain Monte Carlo method that can be used to find approximants to $q(t,x)$. The idea, as in all Monte Carlo methods, is to express $q(t,x)$ as the mean of a functional of $X(\cdot)$, and then repeatedly simulate from $X(\cdot)$. This can be achieved in the following way.

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

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

We use P and g to define a non-homogeneous Markov chain $X(\cdot)$ on $\mathcal{X}$ that evolves 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 chosen so that $\mathbb{P}_x(\tau < \infty) = 1$ for all $x \in \mathcal{B}$.

Now write (4.1) as

$$\begin{aligned} q(t,x) &= \int_t^\infty \sum_{y\in\mathcal{A}} f(t,x;s,y)q(s,y)P(s;x,y)g(t,x;s)ds \\ &\quad + \int_t^\infty \sum_{y\in\mathcal{B}} f(t,x;s,y)q(s,y)P(s;x,y)g(t,x;s)ds, \quad x \in \mathcal{B}, \end{aligned} \tag{4.2}$$

where

$$f(t,x;s,y) = \frac{r(s;x,y)}{P(s;x,y)}. \tag{4.3}$$

Let $\tau_1 < \tau_2 \cdots < \tau_k = \tau$ be the jump times of $X(\cdot)$ and define $\tau_0 = t$. It is shown in [10] that

$$q(t,x) = \mathbb{E}_{(t,x)} q(\tau, X(\tau)) \prod_{j=1}^{k} f(\tau_{j-1}, X(\tau_{j-1}); \tau_j, X(\tau_j)), \tag{4.4}$$

where $\mathbb{E}_{(t,x)}$ denotes expectation with respect to $X(t) = x$. This representation provides a simple 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 (4.4) for each of them. The average of these values is an unbiased estimate of $q(t,x)$, and standard theory may be used to assess how accurately $q(t,x)$ has been estimated.

There is a canonical candidate for P, obtained by setting

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

and

$$P(s;x,y) = \frac{r(s;x,y)}{f(x;s)}. \tag{4.5}$$

Equation (4.3) shows that $f(t,x;s,y)$ reduces to

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

For the special case in (3.9), if we define

$$w^*(s;\boldsymbol{n}) = \theta \sum_{\substack{i,j\in[d]\\ n_j>0, i\neq j}} (n_i+1)p_{ij} + n\lambda(s) \sum_{j\in[d],n_j>0} (n_j-1),$$

then

$$\begin{aligned} P(s;\boldsymbol{n},\boldsymbol{m}) &= \frac{\theta}{w^*(s;\boldsymbol{n})}(n_i+1)p_{ij}, \quad \boldsymbol{m} = \boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j,\ i,j\in[d],\\ &\qquad\qquad n_j>0,\ i\neq j\\ &= \frac{n\lambda(s)}{w^*(s;\boldsymbol{n})}(n_j-1), \quad \boldsymbol{m} = \boldsymbol{n}-\boldsymbol{e}_j,\ j\in[d],\ n_j>0. \end{aligned}$$

We have found it important in practice, particularly in the context of variance reduction, to have some flexibility in choosing the stopping time τ, or, equivalently, the set $\mathcal{A}$. For instance, the natural choice for the case (3.9) has $\mathcal{A} = \{\boldsymbol{m} : \sum_{i=1}^d m_i = 1\}$. This corresponds to tracing the ancestry back to a single individual. However, it is sometimes possible to calculate sampling probabilities, either explicitly or perhaps numerically, when there are two or three distinct ancestors, rather than tracing the genealogy back to just a single individual. In this case we can take $\mathcal{A} = \{\boldsymbol{m} : \sum_{i=1}^d m_i = 2\}$ for example.

4.1. Surface simulation and Monte Carlo likelihoods. The sampling probability $q(t,x)$ is usually a function of some unknown parameters, denoted 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. To compute q on a surface of Γ-values, 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 (4.1) in the form

$$q_\Gamma(t,x) = \int_t^\infty \sum_{y\in\mathcal{X}} h_{\Gamma,\Gamma_0}(t,x;s,y) P_{\Gamma_0}(s;x,y) q_\Gamma(s,y)\ g_{\Gamma_0}(t,x;s)ds \tag{4.6}$$

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 [10],

$$q_\Gamma(t,x) = \mathbb{E}_{(t,x)} q_\Gamma(\tau, X(\tau)) \prod_{j=1}^{k} h_{\Gamma,\Gamma_0}(\tau_{j-1}, X(\tau_{j-1}); \tau_j, X(\tau_j)). \tag{4.7}$$

Estimates of $q_\Gamma(t,x)$ may be now obtained as described above. This method is faster than simulating independent runs at a variety of grid points when the cost of producing observations on the process $X(\cdot)$ outweighs the cost of calculating the functionals in (4.7). In exchange for this time saving, the estimates 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.

4.2. Non-homogeneous recursions. Another class of problems that arise in studying probabilistic aspects of the coalescent involves recursions that are non-homogeneous. These may be written in the form

$$m(t,x) = w(t,x) + \int_t^\infty \sum_{y \in \mathcal{X}} r(s;x,y) m(s,y) g(t,x;s) ds, \tag{4.8}$$

where $w(t,x)$ is a known function, and where $m(t,x)$ is known on the set $\mathcal{A}$. Recursions of this form may be solved in a similar way to their homogeneous counterparts, as follows. For $x \in \mathcal{B}$, write (4.8) as

$$\begin{aligned} m(t,x) &= w(t,x) + \int_t^\infty \sum_{y \in \mathcal{A}} f(t,x;s,y) m(s,y) P(s;x,y) g(t,x;s) ds \\ &\quad + \int_t^\infty \sum_{y \in \mathcal{B}} f(t,x;s,y) m(s,y) P(s;x,y) g(t,x;s) ds, \end{aligned} \tag{4.9}$$

and iterate to obtain

$$\begin{aligned} m(t,x) &= \int_t^\infty \sum_{y_1 \in \mathcal{A}} f(t,x;s_1,y_1) m(s_1,y_1) P(s_1;x,y_1) g(t,x;s_1) ds_1 \\ &\quad + \int_t^\infty \sum_{y_1 \in \mathcal{B}} \int_{s_1}^\infty \sum_{y_2 \in \mathcal{A}} f(t,x;s_1,y_1) f(s_1,y_1;s_2,y_2) m(s_2,y_2) \\ &\qquad P(s_1;x,y_1) P(s_2;y_1,y_2) g(t,x;s_1) g(s_1,y_1;s_2) ds_2 ds_1 \\ &\quad + \cdots \\ &\quad + w(t,x) + \int_t^\infty \sum_{y_1 \in \mathcal{B}} f(t,x;s_1,y_1) w(s_1,y_1) P(s_1;x,y_1) g(t,x;s_1) ds_1 \\ &\quad + \int_t^\infty \sum_{y_1 \in \mathcal{B}} \int_{s_1}^\infty \sum_{y_2 \in \mathcal{B}} f(t,x;s_1,y_1) f(s_1,y_1;s_2,y_2) w(s_2,y_2) \end{aligned} \tag{4.10}$$

$$P(s_1; x, y_1)P(s_2; y_1, y_2)g(t, x; s_1)g(s_1, y_1; s_2)ds_2 ds_1$$
$$+ \cdots$$

In terms of the Markov process $X(\cdot)$, $m(t, x)$ may be represented as follows. Let $\tau_1 < \tau_2 < \cdots < \tau_k = \tau$ be the times of the jumps $X(\cdot)$ makes until it reaches the set $\mathcal{A}$, and set $\tau_0 = t$. Then

$$\begin{aligned} m(t, x) &= \mathbb{E}_{(t,x)} \sum_{l=0}^{k-1} w(\tau_l, X(\tau_l)) \prod_{j=1}^{l} f(\tau_{j-1}, X(\tau_{j-1}); \tau_j, X(\tau_j)) \\ &\quad + \mathbb{E}_{(t,x)} m(\tau, X(\tau)) \prod_{j=1}^{k} f(\tau_{j-1}, X(\tau_{j-1}); \tau_j, X(\tau_j)). \end{aligned} \tag{4.11}$$

Once more, independent replicates of $X(\cdot)$ starting from $X(t) = x$ may be used to estimate the expectation in (4.11), values of the sum on the right being accumulated as each simulation progresses. It is straightforward to adapt this scheme to the surface simulation setting of the last section.

4.3. Other sampling properties: the distribution of the time to MRCA. Sampling distributions are not the only quantities that produce recursions to which the Markov chain Monte Carlo method can be applied. One example arises in studying the joint distribution of the sample configuration and the time to the most recent common ancestor. Let $q(t, x, w)$ be the probability that a sample taken at time t has configuration x, and the (further) time to the MRCA is at most w. Of particular interest is the distribution function

$$\mathbb{P}(T_{MRCA} \le w | x) = \frac{q(t, x, w)}{q(t, x)}. \tag{4.12}$$

It is shown in [9] and [10] that $q(t, x, w)$ satisfies a recursion of the form

$$q(t, x, w) = \int_t^\infty \sum_{y \in \mathcal{X}} r(s; x, y) q(s, y, t + w - s) g(t, x; s) ds, \tag{4.13}$$

and that for $w > 0$

$$q(t, x, w) = \mathbb{E}_{(t,x)} q(\tau, X(\tau), t + w - \tau) \prod_{j=1}^{k} f(\tau_{j-1}, X(\tau_{j-1}); \tau_j, X(\tau_j)). \tag{4.14}$$

Under the initial condition

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

the term $q(\tau, X(\tau), t + w - \tau)$ in (4.14) reduces to $I\{\tau \le t + w\}$. If we simulate the process $X(\cdot)$ R times, and define

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

the value of the functional under the expectation sign in (4.4) for the lth simulation, then the distribution in (4.12) 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 $\tau^{(l)}$ is the time the lth simulation hits $\mathcal{A}$. Conditional moments can be computed in a similar way.

4.4. Applications. In this section, we review briefly some of the applications of this computational approach. Further details may be found in the original papers. Computer code is available from the authors on request.

4.4.1. The infinitely-many-sites model. The simplest mutation structure is the infinitely-many-sites model of sequence evolution, in which every mutation in the ancestral tree of the sample produces a new segregating site in the sample. Hudson [13] [14] gives a variety of applications. The sample may be described by a collection of sequences of zeros and ones. If the labeling of the ancestral base at each site is known, we can suppose that the ones denote mutant bases at a site, while the zeros denote sites at which the ancestral type is still present. Typically, this labeling is unknown. The distribution of the sample is determined by certain rooted and unrooted genealogical trees that are embedded in the process; rooted trees correspond to known labeling of sites, unrooted trees to unknown labeling of sites. The theory of these trees is developed in [7]. Markov chain Monte Carlo methods are used to estimate parameters in the varying population size model in [8]. population Inference about the distribution of the time to the most recent common ancestor, conditional on the structure of a sample, is discussed in [9], where applications to mitochondrial sequence data are given. These computer-intensive methods are sometimes time- consuming, and it is therefore of some interest to know how inferences based on simpler summary statistics of the data (for example, the number of segregating sites and alleles) compare to inferences based on the full data. Inference about θ and the time to the MRCA are addressed in [10].

If distinct sequences in the sample are identified as alleles, the sampling theory of the allele frequencies in the constant population size case is given by the Ewens sampling formula [4]. The analogous theory for the variable population size case appears in [8], where the Markov chain Monte Carlo method is also explored. See also [10].

4.4.2. The finitely-many-sites model. Of central interest in the analysis of DNA sequence data is the development of methods for estimating parameters of the substitution process. In the population genetics setting, this can be thought of as the problem of estimating the parameters of the mutation rate matrix R in (3.1). One method, developed by Lundstrom [19] and extended in [20], uses a method of moments approach. In

the simplest mutation model determined by (3.2) with identical substitution matrices $M_l \equiv M$ and equal rates $h_l = 1/s$ at each site, the vectors the count the number of each type of base observed at each site are exchangeable. In particular, they have the same distribution (but of course they are not independent). This observation provides a simple moment method for estimating the entries of the rate matrix θM: equate observed and expected counts, and minimize the sum of squares of the differences. A detailed study of the behavior of this method appears in [19] and [20]; the extension to hypervariable sites is described in [21].

Our development of the Markov chain Monte Carlo method for coalescents was motivated in part by trying to assess whether the estimation methods described above had good statistical properties. The simulation method, together with the surface simulation for likelihoods, is developed for the sampling distribution (3.10) in [6]. Among the issues addressed is the effect on variance reduction of choosing the stopping time τ, and a variety of suggestions for speeding up the method. Note that it is simple to use the same Monte Carlo approach to estimate parameters for the more complicated sequence models described in Section 3.1, and the effects of variable size can be accommodated simply as well [8].

Notice that in the model of sequence evolution determined by (3.2), the mutation processes at different sites are conditionally independent given the genealogy. This means that if the genealogical tree is known, the probability of a set of sequences may be computed by, in effect, reducing the problem to the computation of sampling probabilities at a single site. For simple models for M_l, the mutation matrix at the lth site, it is possible to compute the probability that a base that is of type i at time 0 is type j at time t, and so compute the probability that a site has a particular set of types at the tips of the ancestral tree.

Kuhner, Yamato and Felsenstein [17] [18] have developed an alternative approach to maximum likelihood estimation of θ in this constant population size model. They use a Metropolis-Hastings sampler to sample genealogies, and compute the probability of the set of sequences by using the conditional independence property.

4.4.3. The effects of recombination. The previous examples have been concerned with samples in which the effects of recombination can be ignored. However, the same principles can be applied to study recombination as well. The simplest case is the one with completely unlinked loci, for which computational aspects of the sampling theory can be found in [24], [20], and [21]. For the linked case, think of two finitely-many-alleles loci, A and B, with K alleles at the first locus, L at the second, and mutation rate matrices

$$R_A = \frac{\theta_A}{2}(P^A - I), \quad R_B = \frac{\theta_B}{2}(P^B - I).$$

The analog of the sampling equation (3.10) is a linear system satisfied by the probability $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ of ordered configurations of the form $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$, where $\boldsymbol{a} = (a_1, \ldots, a_K), \boldsymbol{b} = (b_1, \ldots, b_L)$, and $\boldsymbol{c} = (c_{ij}, i \in [K], j \in [L])$. Here, a_i gametes have type i at the A locus and unspecified alleles at the B locus, b_j gametes have type j at the B locus and unspecified alleles at the A locus, and c_{ij} gametes have allele i at the A locus and allele j at the B locus, for $i \in [K], j \in [L]$. The linear system can be derived from a simple coalescent argument, and the sampling formula $q(\mathbf{0}, \mathbf{0}, \boldsymbol{c})$ of the gamete configuration $\boldsymbol{c}$ found by the Markov chain Monte Carlo approach. The same method works to solve the analogous linear system for two infinitely-many-alleles loci that is discussed by Ethier and Griffiths [2], [3]. The methods can also be extended to allow for variable population size, more loci and more complex mutation schemes.

4.4.4. The effects of migration. Nath and Griffiths (1993) derive a recursion analogous to (3.10) in an island model with migration among L islands. $q(\boldsymbol{n})$ is then replaced by $q(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_L)$ the configuration probability of samples of sizes $n_1, \ldots, n_L$ taken from the L islands.

The Markov chain Monte Carlo technique in Section 4 is developed, and the estimated surface of probabilities with the migration rate varying is used to study likelihood estimation of the migration rate in the case of $L = 2$ islands with $d = 2$ possible alleles.

5. Discussion. In this paper, we have reviewed one computational approach for calculating sampling probabilities and related quantities for models arising from versions of the coalescent. The progenitor of this approach dates back at least to the late 1940s, where it was used to solve matrix equations of the form $A\boldsymbol{x} = \boldsymbol{b}$; see Forsythe and Leibler [5] and Halton [11] for example. The techniques advocated here are similar in spirit to the Hastings- Metropolis method [22], [12], where 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.

These techniques may also be applied to other variants on the population genetics theme. The models are described here in terms of 'alleles' and 'mutations', but these may be interpreted in other ways as well. For example, imagine a population of individuals reproducing according to the coalescent, but now think of the 'alleles' as describing the structure of a population of mitochondria within each individual. For example, the parameter θ may be interpreted as the birth-and-death rate of the individual mitochondrial populations, and the transition matrix P describes how a given population reproduces at the birth-and-death times. If each of the mitochondria is labeled as type A or type B, then a plausible model for the

evolution of the individual populations is the two-type Moran model with mutation. More complicated within-individual reproduction mechanisms could of course be used. This provides a simple model for the evolution of a mitochondrial lineage within a reproducing human population. These methods also work for other models in which the branching structure of the coalescent is replaced by other branching processes, such as the binary splitting, or Yule, process. In this case, all that changes in equation (3.10) is the relative rate of 'splits' and 'mutations'.

REFERENCES

[1] BORODOVSKY, M.Y., SPRIZHITSKY, Y., GOLOVANOV, E. and ALEXANDROV, A. *Statistical patterns in the primary structures of functional regions in the genome of E. coli: II Nonuniform Markov models*, Mol. Biol., **20**, 1024-1033, 1986.

[2] ETHIER, S.N. and GRIFFITHS, R.C., *The neutral two-locus model as a measure-valued diffusion*, Adv. Appl. Prob., **22**, 773-786, 1990.

[3] ETHIER, S.N. and GRIFFITHS, R.C., *On the two-locus sampling distribution*, J. Math. Biol., **29**, 131-159, 1990.

[4] EWENS, W.J., *The sampling theory of selectively neutral alleles*, Theoret. Popul. Biol., **3**, 87-112, 1972.

[5] FORSYTHE, G.E. and LEIBLER, R.A, *Matrix inversion by the Monte Carlo method*, Math. Comp., **26**, 127-129, 1950.

[6] GRIFFITHS, R.C. and TAVARÉ, S., *Simulating probability distributions in the coalescent*, Theoret. Popul. Biol., **46**, 131-159, 1994.

[7] GRIFFITHS, R.C. and TAVARÉ, S., *Unrooted genealogical tree probabilities in the infinitely-many-sites model*, Math. Biosci., **127**, 77-98, 1995.

[8] GRIFFITHS, R.C. and TAVARÉ, S., *Sampling theory for neutral alleles in a varying environment*, Phil Trans. R. Soc. Lond. B, **344**, 403-410, 1994.

[9] GRIFFITHS, R.C. and TAVARÉ, S., *Ancestral inference in population genetics*, Statistical Science, **9**, 307-319, 1994.

[10] GRIFFITHS, R.C. and TAVARÉ, S., *Monte Carlo inference methods in population genetics*, Mathl. and Comput. Modelling, in press, 1996.

[11] HALTON, J.H., *A retrospective and prospective study of the Monte Carlo method*, SIAM Review, **12**, 1-63, 1970.

[12] HASTINGS, W.K., *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, **57**, 97-109, 1970.

[13] HUDSON, R.R., *Gene genealogies and the coalescent process*, In: Oxford Surveys in Evolutionary Biology, Volume 7. Edited by D. Futuyma and J. Antonovics, 1-44, 1991.

[14] HUDSON, R.R., *The how and why of generating gene genealogies*, In: Mechanisms of molecular evolution, N. Takahata and A.G. Clark (editors), 23-36, 1992. Sinauer.

[15] KINGMAN, J.F.C., *On the genealogy of large populations*, J. Appl. Prob., **19A**, 27-43, 1982.

[16] KINGMAN, J.F.C., *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.

[17] KUHNER, M.K., YAMATO, J. and FELSENSTEIN, J., *Estimating effective population size from sequence data using Metropolis-Hastings sampling*, Genetics, submitted, 1994.

[18] KUHNER, M.K., YAMATO, J. and FELSENSTEIN, J., *Applications of Metropolis-Hastings genealogy sampling*, IMA volume, in press, 1994.

[19] LUNDSTROM, R. *Stochastic models and statistical methods for DNA sequence data.*

Ph.D. thesis, Mathematics Department, University of Utah, 1990.

[20] LUNDSTROM, R., TAVARÉ, S. and WARD, R.H., *Estimating mutation rates from molecular data using the coalescent*, Proc. Natl. Acad. Sci. USA, **89**, 5961-5965, 1992.

[21] LUNDSTROM, R., TAVARÉ, S. and WARD, R.H., *Modelling the evolution of the human mitochondrial genome*, Math. Biosci, **112**, 319-335, 1992.

[22] METROPOLIS, N., ROSENBLUTH, A.W., ROSENBLUTH, M.N., TELLER, A.H., and TELLER, E., *Equations of state calculations by fast computing machines*, J. Chem. Phys., **21**, 1087-1092, 1953.

[23] NATH, H.B. and GRIFFITHS, R.C., *Estimation in an island model undergoing a multidimensional coalescent process*, Statistics Research Report 226, Monash University, 1993.

[24] SAWYER, S., DYKHUIZEN, D. and HARTL, D., *Confidence interval for the number of selectively neutral amino acid polymorphisms*, Proc. Natl. Acad. Sci. USA, **84**, 6225-6228, 1987.

[25] SLATKIN, M. and HUDSON, R.R., *Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations*, Genetics, **129**, 555-562, 1991.

[26] TAVARÉ, S., *Calibrating the clock: using stochastic processes to measure the rate of evolution*, Chapter 5 in "Calculating the secrets of life", E.S. Lander and M.S. Waterman (editors). National Academy Press, Washington DC, pp. 114-152, 1995.

[27] TAVARÉ, S., *The effects of site dependence on estimating the topology of a tree from DNA sequence data*, in preparation, 1995.

[28] WATTERSON, G.A., *A stochastic analysis of three viral sequences*, Mol. Biol. Evol., **9**, 666-677, 1992.