J. R. Statist. Soc. B (2000)
62, *Part* 4, *pp.* 605–655

Inference in molecular population genetics

Matthew Stephens and Peter Donnelly

University of Oxford, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, March 15th, 2000*, Professor P. J. Diggle *in the Chair*]

Summary. Full likelihood-based inference for modern population genetics data presents methodological and computational challenges. The problem is of considerable practical importance and has attracted recent attention, with the development of algorithms based on importance sampling (IS) and Markov chain Monte Carlo (MCMC) sampling. Here we introduce a new IS algorithm. The optimal proposal distribution for these problems can be characterized, and we exploit a detailed analysis of genealogical processes to develop a practicable approximation to it. We compare the new method with existing algorithms on a variety of genetic examples. Our approach substantially outperforms existing IS algorithms, with efficiency typically improved by several orders of magnitude. The new method also compares favourably with existing MCMC methods in some problems, and less favourably in others, suggesting that both IS and MCMC methods have a continuing role to play in this area. We offer insights into the relative advantages of each approach, and we discuss diagnostics in the IS framework.

Keywords: Ancestral inference; Coalescent; Computationally intensive inference; Importance sampling; Markov chain Monte Carlo methods; Population genetics

1. Introduction

There has been a long and mutually beneficial history of interaction between population genetics and statistical science, dating back principally to Fisher, Wright and Haldane (for background, see for example Ewens (1979)). For most of this history, population genetics was largely theory driven, with the theory depending often on applied probability modelling. Markov chains, diffusion processes and more recently measure-valued diffusions and the coalescent have played important roles in the study of stochastic models for genetic evolution within a population.

Recent experimental advances have led to an explosion in data which document genetic variation at the level of deoxyribonucleic acid (DNA) within and between populations. These kinds of data lead to challenging inference problems. In the abstract, they consist of a high dimensional, but partial, snapshot, taken at a single point in time, from the evolution of a complicated stochastic process. Whereas the structure of the stochastic models may be well understood, explicit expressions for probability distributions are typically not available. Further, distinct data points — typically genetic information from sampled chromosomes at a particular region of interest — are highly positively correlated, exactly because the sampled chromosomes share ancestral history. As a consequence, there is limited information even in very large samples. (In many problems, this information grows only as the logarithm of the sample size or worse. In others the amount of information in a sample is bounded as the

Address for correspondence: Matthew Stephens, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK.
E-mail: stephens@stats.ox.ac.uk

 1369–7412/00/62605

sample size increases.) The limited information in data puts a premium on efficient inferential methods. In practice though, inference has typically been based on low dimensional summary statistics. There has thus been growing interest in the development of full likelihood-based inference methods.

Although there are several directions from which inference may be approached, all have the flavour and structure of missing data problems. Two principal schools have pioneered full likelihood-based inference in population genetics via computationally intensive techniques. The first, due to Griffiths and Tavaré (Griffiths and Tavaré, 1994a, b, c, 1999), exploited a clever computational method for approximating the solution to recursive equations for quantities of interest. The second, due to Felsenstein and colleagues (Kuhner *et al.*, 1995, 1998), takes a Markov chain Monte Carlo (MCMC) approach. Subsequent developments on each theme have either adapted the methods to different genetic systems or, in the MCMC context, adopted alternative formulations or proposal distributions (Griffiths and Marjoram, 1996; Nielsen, 1997; Wilson and Balding, 1998; Beaumont, 1999; Bahlo and Griffiths, 2000; Slade, 2000).

The Griffiths–Tavaré approach has recently been shown (Felsenstein *et al.*, 1999) to be a version of importance sampling (IS). This observation is the starting-point for our analysis. The optimal IS distribution can be characterized. It effectively depends on the time reversal of certain stochastic processes. Although this optimal solution is tantalizingly inaccessible in most (but not all) settings, insights into the stochastic models suggest natural approximations for the optimal proposal distribution. We use these as the basis for a new IS approach.

All methods for these problems are extremely computationally intensive, and often on or beyond the borders of practicability for realistically sized problems. Our new approach is most naturally compared with the original method of Griffiths and Tavaré. It represents a substantial improvement in efficiency (typically by several orders of magnitude) and accuracy. We also compare our method with existing MCMC approaches. In problems in which the genetic structure is 'constrained' (in a sense which will become clearer later) IS seems competitive with, or superior to, these MCMC approaches. The latter seem to have an advantage for less constrained problems. There are parallel difficulties with implementation of either MCMC or IS methods in complex problems such as these, effectively because one can never be sure that the algorithm has been run for sufficiently long. These difficulties seem to have been underestimated in early applications within genetics. Our analysis may offer some useful practical insights.

At a more generic level, inference in population genetics may provide a useful model system in which to gain insight for inference procedures in other complicated settings, and in particular in high or infinite dimensional stochastic processes. The success of IS methods in this context may provide a useful counterpoint to the routine use of MCMC sampling as the first choice for implementing computationally intensive inference. More specifically, there are useful general guidelines on which sorts of MCMC or IS approaches may be more successful than others.

We set the scene in the next section, describing the simplest versions of the genetics and demographic models of interest. Section 3 describes the simplest generic inference problem and the potential for both IS and MCMC methods in its solution. Motivated by genealogical arguments, and the structure of the optimal IS function in this context, we introduce a new IS function in Section 4. We implement this for several different types of genetic data and compare its performance with existing IS and MCMC methods (Section 5). The final section offers insights into the strengths and weaknesses of the various methods and discusses possible extensions and improvements.

The flavour of the inference problem on which this paper is focused is reminiscent of the problem addressed by Edwards (1970). In each case, the structure in genetic data arises through the action of stochastic processes superimposed on an evolutionary tree. In the phylogenetic context addressed by Edwards the central question relates to inference for the unobserved tree. In the population genetics setting considered in this paper, much or all of the structure of the tree is not of primary interest, instead playing the role of missing data. Of course, computationally intensive statistical methods were in their infancy in 1970, but we note an analogy between the approach suggested by Whittle (1970) in the discussion of Edwards (1970) and the approach which we adopt here.

2. Demographic and genetic models

To illustrate our methods we shall consider what is effectively the simplest demographic and genetic scenario. Although rather simple, this model is nevertheless the basis of most existing inference from the growing amounts of molecular population genetics data. We shall give a brief informal description of the underlying model. For more applied background see Donnelly and Tavaré (1995) and references therein. For a rigorous treatment see Ethier and Kurtz (1993) or Donnelly and Kurtz (1996a). The methods discussed here have been extended to several more general models.

Many organisms are diploid, carrying chromosomes in pairs. We focus attention on a chromosomal region that is sufficiently small that the possibility of recombination over relevant timescales can be ignored. As a consequence, each chromosomal segment is descended from a single chromosome in the previous generation, and it is enough to consider haploid demographic models, i.e. those in which 'individuals' have a single parent.

Our methods apply to a wide range of specific demographic models (formally, those which are in the domain of attraction of the Fleming–Viot diffusion; Donnelly and Kurtz (1999)). For example they apply to the *Wright–Fisher* model, in which a population that has been of constant (large) size N chromosomes throughout its history evolves in non-overlapping generations, with the joint distribution of offspring numbers in a particular generation being symmetric multinomial and independent of offspring numbers in other generations. For definiteness we shall focus the description here on populations of constant size N chromosomes which evolve in non-overlapping generations.

Within the region of interest each chromosome has an associated genetic type. We denote the set of possible types by E, which we assume for now to be countable. (In practice, the choice of E will depend on the genetic system being modelled; a model for which E is uncountable is considered in Section 5.5.) Independently of all other events, the genetic type of each progeny of a chromosome of type $\alpha \in E$ is α with probability $1 - \mu$, and $\beta \in E$ with probability $\mu P_{\alpha\beta}$, i.e. mutations occur at rate μ per chromosome per generation, according to a Markov chain with transition matrix P. We assume that the evolution is neutral, in that the demography is independent of the genetic types of the chromosomes. To simplify the description we focus on the case where P has a unique stationary distribution.

Now consider a random sample of n chromosomes, taken from the population at stationarity. In the region of interest, each chromosome has a unique ancestor in any previous generation. The ancestral relationships among the sample back to its most recent common ancestor (MRCA) can then be described by a genealogical tree (see for example Fig. 1). It turns out that the natural timescale on which to view the evolution of the population is in units of N/ν^2 generations, where ν^2 is the variance of the number of progeny of an arbitrary chromosome (see for example Wright (1969), page 215, and references therein). On this scale the distribution of

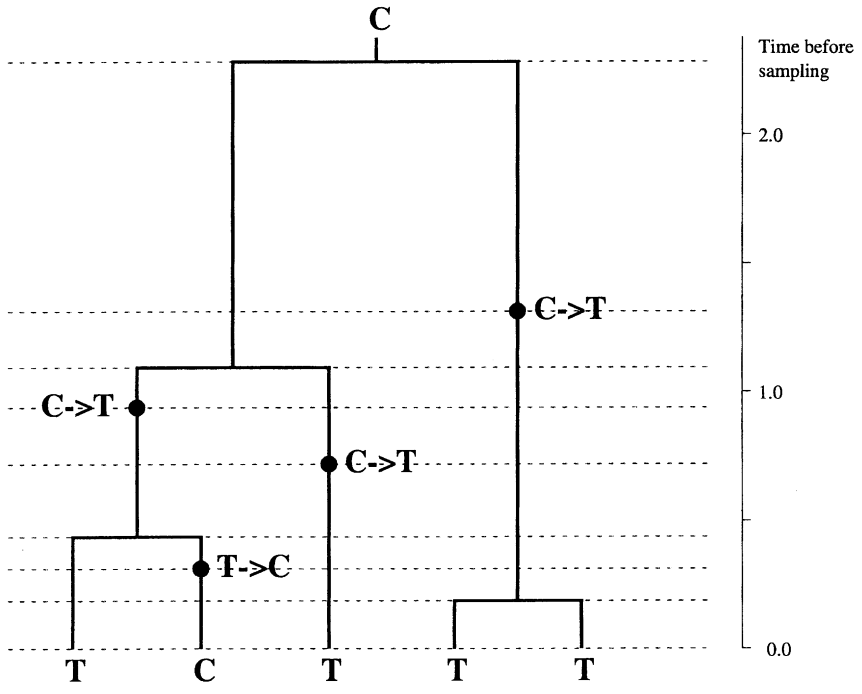

Fig. 1. Illustration of a genealogical tree $\mathcal{G}$, typed ancestry $\mathcal{A}$ and corresponding history $\mathcal{H}$, for five individuals with genetic types in $E = \{C, T\}$: the sampled individuals are represented by the tips at the bottom of the tree, and their ancestral lineages are represented by thick lines running up the page (time runs backwards up the page and is measured in units of N/ν^2 generations); ancestral lineages are joined by horizontal lines (and are said to *coalesce*) when they share a common ancestor; the dots represent mutations, and horizontal dotted lines indicate the times at which events (coalescences and mutations) occur; the typed ancestry $\mathcal{A}$ consists of all the information in the figure, whereas the genealogy $\mathcal{G}$ consists of only the full lines (including times of coalescences); for this typed ancestry, the history $\mathcal{H} = (H_{-m}, H_{-(m-1)}, \ldots, H_1, H_0)$ may be represented as $(\{C\}, \{C, C\}, \{C, T\}, \{C, C, T\}, \{C, T, T\}, \{T, T, T\}, \{T, T, T, T\}, \{C, T, T, T\}, \{C, T, T, T, T\})$

the genealogical tree of the sample is well approximated by that of a random tree called the *n-coalescent* (Kingman, 1982a, b, c). At stationarity, under the neutrality assumption, the distribution of the type of the MRCA is given by the stationary distribution of the (mutation) Markov chain P. Conditionally on the genealogical tree and the type of the MRCA, types change along different branches of the tree as independent continuous time Markov chains with rate $\theta/2 = N\mu/\nu^2$ and transition matrix P (see for example Donnelly and Tavaré (1995)). Throughout the paper we shall adopt the coalescent approximation as the model underlying the data.

Define the *typed ancestry* $\mathcal{A}$ of a sample to be its genealogical tree $\mathcal{G}$, together with the genetic type of the MRCA, and the details and positions of the mutation events which occur along the branches of $\mathcal{G}$ (see Fig. 1). The distribution of the typed ancestry of a random sample of n chromosomes taken from the population at stationarity depends on N, μ and ν^2 only through θ. Later we regard θ as unknown, and the transition matrix P as known, and focus on inference for θ, so we shall suppress dependence on P and denote probability distributions by $P_\theta(\cdot)$. The distribution of $\mathcal{A}$ may be sampled from (directly) using the following algorithm.

Algorithm 1. To sample from $P_\theta(\mathcal{A})$ we follow these steps. (This is a continuous time

version of the discrete time urn model discussed by Ethier and Griffiths (1987); its properties are investigated by Donnelly and Kurtz (1996b).)

Step 1: choose a type in E at random according to the stationary distribution $\psi(\cdot)$ of the transition matrix P. Start the ancestry with a gene of this type which splits immediately into two lines of this type.
Step 2: if there are currently k lines in the ancestry, wait a random amount of time which is exponentially distributed with rate parameter $\lambda_k = k(k-1+\theta)/2$ and then choose an ancestral line at random from the k. Split this line into two lines (each with the type of the progenitor) with probability $(k-1)/(k-1+\theta)$; otherwise mutate it (according to P).
Step 3: if there are fewer than $n+1$ lines in the ancestry return to step 2. Otherwise go back to the last time at which there were n lines in the ancestry and stop.

For mutation models such as the infinite sites model (Section 5.5), in which the transition matrix P does not have a stationary distribution, it is usually sufficient to measure types relative to the type of the MRCA, with an arbitrary type being assigned to the MRCA in step 1 of algorithm 1.

One aspect of $\mathcal{A}$ which will play a key role in what follows is the *history* $\mathcal{H}$, which we may describe informally as $\mathcal{A}$ with the time and topology information removed. Formally we define $\mathcal{H}$ to be the type of the MRCA, together with an ordered list of the split and mutation events which occur in the typed ancestry (including the details of the genetic type or types involved in each event, but not including the details of exactly which line was involved in each event). The history $\mathcal{H}$ thus includes a record of the states $(H_{-m}, H_{-(m-1)}, \ldots, H_1, H_0)$ visited by a Markov process beginning with the genetic type $H_{-m} \in E$ of the MRCA and ending with the genetic types $H_0 \in E^n$ corresponding to the genetic types of a random sample. Here m is random, and the H_i are *unordered* lists of genetic types. For notational convenience we shall write $\mathcal{H} = (H_{-m}, H_{-(m-1)}, \ldots, H_1, H_0)$, although $\mathcal{H}$ actually contains details of the transitions which occur between these states (which may not be uniquely determined by the list of states if $P_{\alpha\alpha} > 0$ for more than one $\alpha \in E$). If H_i is obtained from H_{i-1} by a mutation from a type α to a type β then we write $H_i = H_{i-1} - \alpha + \beta$. If H_i is obtained from H_{i-1} by a split of a line of type α then we write $H_i = H_{i-1} + \alpha$. The concept of a history is illustrated in Fig. 1.

It follows from algorithm 1 that the distribution $P_\theta(\mathcal{H})$ is defined by the distribution $\psi(\cdot)$ of the type of the MRCA, the stopping procedure (step 3 of the algorithm) and the Markov transition probabilities

$$p_\theta(H_i|H_{i-1}) = \begin{cases} \dfrac{n_\alpha}{n}\dfrac{\theta}{n-1+\theta}P_{\alpha\beta} & \text{if } H_i = H_{i-1} - \alpha + \beta, \\ \dfrac{n_\alpha}{n}\dfrac{n-1}{n-1+\theta} & \text{if } H_i = H_{i-1} + \alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where n is the number of chromosomes in H_{i-1} and n_α is the number of chromosomes of type α in H_{i-1}.

Consider now the distribution $\pi_\theta(\cdot)$ of the genetic types $A_n = (a_1, a_2, \ldots, a_n) \in E^n$ of a random (ordered) sample of n chromosomes. Since the history $\mathcal{H}$ includes (in H_0) the genetic types of a sample, algorithm 1 also provides a straightforward way of sampling from $\pi_\theta(\cdot)$. To give an ordered sample, a random order must be assigned to the n types in H_0, so

$$\pi_\theta(A_n|\mathcal{H}) = \pi_\theta(A_n|H_0) = \begin{cases} \left(\prod_{\alpha\in E} n_\alpha!\right) \Big/ n! & \text{if } H_0 \text{ is consistent with } A_n, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

3. Approaches to inference

We now focus on the problem of performing inference based on the genetic types $A_n = (a_1, a_2, \ldots, a_n) \in E^n$ of a random (ordered) sample of n chromosomes taken from the population at stationarity. Questions of interest may relate to inference for

(a) the 'evolutionary parameters' (in our case the scaled mutation rate θ, or the mutation matrix P),
(b) aspects of the unobserved ancestry of the sampled chromosomes (e.g. the time to the MRCA of the sample) and
(c) the demographic history of the population.

For ease of exposition we focus on the problem of likelihood inference for θ (assuming P is known). We define

$$L(\theta) \equiv \pi_\theta(A_n) \quad (3)$$

and refer to this as the *likelihood*. It is typically sufficient to know the likelihood up to an arbitrary multiplicative constant (which may depend on the data, but not the parameter), and indeed likelihood is often defined only up to such a constant. We shall distinguish below between approaches which approximate the probability of the data as a function of the parameter and approaches which approximate a function proportional to this (where the constant of proportionality is unknown). To facilitate this distinction we shall reserve the term likelihood for definition (3) and refer to any function proportional to this as a *relative likelihood*. We note that in some circumstances, e.g. when a model is to be compared with other (non-nested) models, it can be helpful to know the likelihood rather than only the relative likelihood.

For most mutational models of interest, no explicit expression for $L(\theta)$ is available. Writing

$$L(\theta) = \pi_\theta(A_n) = \int \pi_\theta(A_n|\mathcal{H})\, P_\theta(\mathcal{H})\, \mathrm{d}\mathcal{H}, \quad (4)$$

and noting that $\pi_\theta(A_n|\mathcal{H})$ is easily calculated (using equations (2)), suggests viewing this as a missing data, or data augmentation, problem with $\mathcal{H}$ being the missing data. (Some of the many generic methods that are now available for such problems, including MCMC methods and variants on the EM algorithm, are described in Tanner (1993) and Gilks *et al.* (1996).) As is common in this context, there is a choice of what to include in the missing data, with the usual trade-offs (typically, the lower the dimension of the missing data $\mathcal{T}$, the harder it is to calculate $\pi_\theta(A_n|\mathcal{T})$). We note though that for all potential data augmentation strategies the dimension of the space in which the missing data live will be enormous. For example, there are $n!(n-1)!/2^{n-1}$ different possible topologies for the underlying trees. The high dimensional nature of the missing data is one of the main reasons that these problems pose such a computational challenge.

Expression (4) suggests a naïve estimator of $L(\theta)$:

$$L(\theta) = \pi_\theta(A_n) \approx \frac{1}{M}\sum_{i=1}^{M} \pi_\theta(A_n|\mathcal{H}^{(i)}) \tag{5}$$

where $\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(M)}$ are independent samples from $P_\theta(\mathcal{H})$. Since $\pi_\theta(A_n|\mathcal{H})$ is 0 unless $\mathcal{H}$ is consistent with A_n, each term of the sum will be 0 with very high probability, and reliable estimation will require values of M (in the range 10^6–10^{100} for the examples that we consider here) which are too large for the method to be practicable. Similar problems appear to persist for any other potential choice of missing data. It is therefore necessary to turn to more sophisticated computational methods to approximate the likelihood.

3.1. The Griffiths–Tavaré scheme

The first such method in this context was developed by Griffiths and Tavaré (1994a). They showed that by finding recursions for probabilities of interest the likelihood could be written in the form

$$L(\theta) = \pi_\theta(A_n) = E\left\{ \prod_{j=0}^{\tau} F(B_j) \middle| B_0 = A_n \right\} \tag{6}$$

where $B_0, B_1, \ldots, B_\tau$ is a particular set-valued Markov chain stopped the first time τ that it has cardinality 1. This leads to a natural Monte Carlo approximation for $L(\theta)$: simply evaluate the expectation above by repeatedly simulating the chain started from A_n, and averaging the realized values of $\Pi_{j=0}^{\tau} F(B_j)$ across the simulated realizations of the chain.

Naïve application of this method to estimate the likelihood function $L(\cdot)$ requires simulation from a different Markov chain for each value of θ. Griffiths and Tavaré (1994a) showed how to use realizations of a single Markov chain to estimate the likelihood function. They subsequently extended the method to more general models and inference questions (see for example Griffiths and Tavaré (1994b, c, 1999)).

3.2. Markov chain Monte Carlo method I

A natural alternative to the Griffiths–Tavaré scheme, which was first pursued by Kuhner *et al*. (1995, 1998), is to deal with the missing data by using MCMC methods. If $\mathcal{T}$ represents missing data for which $\pi_\theta(A_n|\mathcal{T})$ is (relatively) easy to calculate, then it is straightforward to use the Metropolis–Hastings algorithm to construct a Markov chain with stationary distribution $P_\theta(\mathcal{T}|A_n)$. If $\mathcal{T}^{(1)}, \mathcal{T}^{(2)}, \ldots$ is a sample (after burn-in, and possibly thinning) from a Markov chain with stationary distribution $P_{\theta_0}(\mathcal{T}|A_n)$, then a relative likelihood surface for θ can be estimated by using

$$\begin{aligned} \frac{L(\theta)}{L(\theta_0)} &= \frac{\int P_\theta(\mathcal{T},\, A_n)\, \mathrm{d}\mathcal{T}}{P_{\theta_0}(A_n)} \\ &= \int \frac{P_\theta(\mathcal{T},\, A_n)}{P_{\theta_0}(\mathcal{T},\, A_n)} P_{\theta_0}(\mathcal{T}|A_n)\, \mathrm{d}\mathcal{T} \\ &\approx \frac{1}{M}\sum_{i=1}^{M} \frac{P_\theta(A_n,\, \mathcal{T}^{(i)})}{P_{\theta_0}(A_n,\, \mathcal{T}^{(i)})}, \end{aligned} \tag{7}$$

as in Geyer and Thompson (1992) (see also Geyer (1996)).

3.3. Markov chain Monte Carlo method II

Wilson and Balding (1998) developed an alternative MCMC approach. They took a fully Bayesian view, putting a prior distribution $P(\theta)$ on θ, and constructing a Markov chain with stationary distribution $P(\mathcal{T}, \theta|A_n)$. Although they base inference on the posterior distribution of θ (an approach which we fully endorse), we note that, since the posterior density is proportional to the product of the prior and the likelihood, an estimate of the relative likelihood surface for θ may be obtained by estimating the posterior density and dividing by the prior density. The posterior density is most easily estimated by smoothing a sample $\theta^{(1)}, \ldots, \theta^{(M)}$ from the Markov chain, although potentially more efficient methods exist (see for example Chen (1994)).

In contrast with the Griffiths–Tavaré scheme described above, and the IS methods described below, these MCMC approaches give only relative, rather than absolute, likelihood surfaces (although methods such as those reviewed by Raftery (1996) may allow the likelihood itself to be computed).

3.4. Importance sampling

IS (see Ripley (1987) for background) is a standard method of reducing the variance of Monte Carlo estimators such as expression (5). If $Q_\theta(\cdot)$ is any distribution on ancestries whose support includes $\{\mathcal{H}: \pi_\theta(A_n|\mathcal{H}) > 0\}$, then equation (4) may be rewritten as

$$L(\theta) = \int \pi_\theta(A_n|\mathcal{H}) \frac{P_\theta(\mathcal{H})}{Q_\theta(\mathcal{H})} Q_\theta(\mathcal{H})\, \mathrm{d}\mathcal{H} \tag{8}$$

$$\approx \frac{1}{M} \sum_{i=1}^{M} \pi_\theta(A_n|\mathcal{H}^{(i)}) \frac{P_\theta(\mathcal{H}^{(i)})}{Q_\theta(\mathcal{H}^{(i)})} = \frac{1}{M} \sum_{i=1}^{M} w^{(i)} \text{ say,} \tag{9}$$

where $\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(M)}$ are independent samples from $Q_\theta(\cdot)$. The distribution $Q_\theta(\cdot)$ is known as the *IS function* or the *proposal distribution*, and the $w^{(i)}$ are known as the *IS weights*. For a careful choice of Q_θ the variance of the estimator (9) will be much smaller than that of estimator (5). The optimal choice Q_θ^* of Q_θ is the conditional distribution of histories given the data,

$$Q_\theta^*(\mathcal{H}) = P_\theta(\mathcal{H}|A_n), \tag{10}$$

as then every term of the sum (9) is identical:

$$\pi_\theta(A_n|\mathcal{H}) \frac{P_\theta(\mathcal{H})}{Q_\theta^*(\mathcal{H})} = \pi_\theta(A_n|\mathcal{H}) \frac{P_\theta(\mathcal{H})}{P_\theta(\mathcal{H}|A_n)} = \frac{P_\theta(\mathcal{H} \cap A_n)}{P_\theta(\mathcal{H}|A_n)} = \pi_\theta(A_n) = L(\theta), \tag{11}$$

and the variance of the estimator is 0. Unfortunately the conditional distribution of histories given the data, $P_\theta(\cdot|A_n)$, is not known in most cases of interest. Indeed, it follows from equation (11) that knowledge of $P_\theta(\mathcal{H}|A_n)$, for any $\mathcal{H}$, is equivalent to knowledge of the likelihood $L(\theta)$.

Expression (9) could be used to estimate the likelihood independently for many different values of θ, using samples from a different IS function Q_θ for each value of θ. However, it appears to be more efficient to reuse samples from a single IS function. This is in theory straightforward: for any fixed θ_0 we have

$$L(\theta) \approx \frac{1}{M} \sum_{i=1}^{M} \pi_\theta(A_n|\mathcal{H}^{(i)}) \frac{P_\theta(\mathcal{H}^{(i)})}{Q_{\theta_0}(\mathcal{H}^{(i)})}, \tag{12}$$

where $\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(M)}$ are independent samples from $Q_{\theta_0}(\cdot)$. This approach is due to Griffiths and Tavaré (1994a), which refers to θ_0 as the 'driving value' of θ.

In summary, three frameworks can be distinguished: the use of IS to estimate the likelihood at one particular value of θ, with further IS (via equation (12)) to estimate the likelihood surface (we shall see below that the Griffiths–Tavaré method is a particular instance of this approach); the use of MCMC methods to sample from the conditional distribution of the missing data, again at one particular θ-value, with IS (via equation (7)) to estimate the likelihood surface; MCMC methods which sample from the joint conditional distribution of the missing data and θ.

4. Towards a more efficient proposal distribution

Felsenstein *et al.* (1999) recently noted that the Griffiths–Tavaré scheme is a version of IS with a particular proposal distribution. (As noted by Griffiths and Tavaré (1994a), their method is a particular instance of a Monte Carlo approach to solving linear equations, which has a long history. The connection with IS in this context had already been observed; see for example Ripley (1987), section 7.3.) It is thus natural to ask whether other IS schemes, corresponding to different proposal distributions, may be more efficient.

A natural class of proposal distributions on histories arises if we consider randomly constructing histories backwards in time, in a Markov way, from the sample A_n to an MRCA, i.e. a random history $\mathcal{H} = (H_{-m}, \ldots, H_{-1}, H_0)$ may be sampled by choosing $H_0 = A_n$, and successively choosing H_{i-1}, $i - 1 = -1, -2, \ldots, -m$, according to prespecified backward transition probabilities $q_\theta(H_{i-1}|H_i)$. The process stops at the first time $-m$ at which the configuration H_{-m} consists of a single type.

For equation (8) to hold it is necessary to restrict attention to the subclass $\mathcal{M}$ of these distributions for which, for each i, the support of $q_\theta(\cdot|H_i)$ is the set

$$\{H_{i-1}: p_\theta(H_i|H_{i-1}) > 0\},$$

where the forward transition probabilities p_θ are defined at equations (1). Specifying such backward transition probabilities q_θ then defines a distribution Q_θ with support the set of histories consistent with A_n. Furthermore it is straightforward to simulate from Q_θ and to evaluate the ratio $P_\theta(\mathcal{H})/Q_\theta(\mathcal{H})$, so that approximation (9) may be applied.

When viewed as an IS method, the Griffiths–Tavaré scheme corresponds to a proposal distribution Q_θ^{GT} in the class $\mathcal{M}$, with

$$q_\theta(H_{i-1}|H_i) \propto p_\theta(H_i|H_{i-1}), \tag{13}$$

where the forward transition probabilities p_θ are defined at equations (1). In fact there is a whole family of Markov chains satisfying equation (6), each with its own function F. There is a one-to-one correspondence between the Markov chains in this family and the class of Markov proposal distributions described two paragraphs above. (Note that this is different from the generation of alternative Griffiths–Tavaré schemes via 'renormalized sampling probabilities' suggested in Griffiths and Tavaré (1994a), page 157.) Implementations of the Griffiths–Tavaré scheme (by themselves and others) have routinely used the particular Markov chain leading to expression (13), which Griffiths and Tavaré (1997), page 174, refer to as the 'canonical candidate'.

We begin our search for a more efficient proposal distribution by characterizing the optimal proposal distribution (10) in theorem 1 below. Although it is not possible to sample from this distribution directly, the key to our IS approach is to exploit the characterization in approximating the optimal proposal distribution.

Theorem 1. Write $\pi(\cdot|A_n)$ for the conditional distribution of the type of an $(n+1)$th sampled chromosome, given the types, A_n, of the first n sampled chromosomes:

$$\pi(\alpha|A_n) = \frac{\pi_\theta\{(A_n, \alpha)\}}{\pi_\theta(A_n)}. \tag{14}$$

The optimal proposal distribution Q_θ^* is in the class $\mathcal{M}$, with

$$q_\theta^*(H_{i-1}|H_i) = \begin{cases} C^{-1}\dfrac{\theta}{2} n_\alpha \dfrac{\pi(\beta|H_i - \alpha)}{\pi(\alpha|H_i - \alpha)} P_{\beta\alpha} & \text{if } H_{i-1} = H_i - \alpha + \beta, \\ C^{-1}\dbinom{n_\alpha}{2} \dfrac{1}{\pi(\alpha|H_i - \alpha)} & \text{if } H_{i-1} = H_i - \alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

where n_α denotes the number of chromosomes of type α in H_i. The constant of proportionality C is given by

$$C = \frac{n(n-1+\theta)}{2},$$

where n is the number of chromosomes in H_i.

Proof. That Q_θ^* is in the class $\mathcal{M}$ follows from the Markov nature of $\mathcal{H}$ (in particular, the fact that $\{H_{-m}, H_{-(m-1)}, \ldots, H_{i-1}\}$ and $\{H_{i+1}, H_{i+2}, \ldots, H_0$, including the fact that the history ends at $H_0\}$ are conditionally independent given H_i).

We now derive the backward transition rates for $\mathcal{A}$ using the 'particle representation' of Donnelly and Kurtz (1996a) which is a convenient embedding of both $P_\theta(\mathcal{A})$ and $\pi_\theta(A_n)$ within the following continuous time Markov process on E^n.

(a) At time $t = 0$ assign types from E randomly to $A_n(0) = (a_1(0), \ldots, a_n(0))$ exchangeably.
(b) Let $a_1(\cdot)$ evolve from $a_1(0)$ according to a continuous time Markov process with rate $\theta/2$ and transition matrix P (the *mutation process*).
(c) Let $a_k(\cdot)$ $(k = 2, \ldots, n)$ evolve as follows: at points of a Poisson process of rate $k-1$, a_k chooses i uniformly at random from $1, \ldots, k-1$, and 'looks down and copies' the value of a_i at that time. Between these events it evolves according to the mutation process.
(d) All the above 'look down' and mutation processes are independent, and independent of (a).

Donnelly and Kurtz (1996a) showed that this process has stationary distribution $\pi_\theta(\cdot)$, and that looking backwards in time (at stationarity) in this process, treating the 'look down and copy' moves as coalescence events, gives the typed ancestry of a sample.

Suppose that at time t the ancestry consists of k lineages. Since the lines of the particle model are exchangeable, we can assume without loss of generality that these ancestral lineages correspond to the first k lines of the particle model $A_k(t) = (a_1(t), a_2(t), \ldots, a_k(t))$. If the configuration of types at time t is $A_k(t) = (\alpha_1, \alpha_2, \ldots, \alpha_{k-1}, \alpha)$ then the probability of the event Υ_m that in the last δ time units there was a mutation from $a_k(t-\delta) = \beta$ to $a_k(t) = \alpha$ is given by

$$P\{\Upsilon_m \cap A_k(t-\delta) = (\alpha_1, \ldots, \alpha_{k-1}, \beta) | A_k(t) = (\alpha_1, \ldots, \alpha_{k-1}, \alpha)\}$$
$$= \frac{P\{\Upsilon_m \cap A_k(t-\delta) = (\alpha_1, \ldots, \alpha_{k-1}, \beta) \cap A_k(t) = (\alpha_1, \ldots, \alpha_{k-1}, \alpha)\}}{P\{A_k(t) = (\alpha_1, \ldots, \alpha_{k-1}, \alpha)\}}$$

$$= \frac{\pi(\alpha_1, \ldots, \alpha_{n-1}, \beta)\delta\theta P_{\beta\alpha}/2}{\pi(\alpha_1, \ldots, \alpha_{n-1}, \alpha)} + o(\delta)$$

$$= \delta\frac{\theta}{2}\frac{\pi(\beta|A_k - \alpha)}{\pi(\alpha|A_k - \alpha)}P_{\beta\alpha} + o(\delta).$$

By exchangeability this result holds for every line of type α, and so multiplying through by n_α gives the total rate at which mutations occur backwards in time from α to β.

Similarly, if $A_k(t) = (\alpha_1, \ldots, \alpha_{k-2}, \alpha, \alpha)$, then the probability of the event Υ_c that in the last δ time units there was a coalescence of lineages k and $k-1$ (i.e. line k looked down and copied line $k-1$) is given by

$$\frac{\sum_\beta P\{\Upsilon_c \cap A_k(t-\delta) = (\alpha_1, \ldots, \alpha_{k-2}, \alpha, \beta) \cap A_k(t) = (\alpha_1, \ldots, \alpha_{k-2}, \alpha, \alpha)\}}{P\{A_k(t) = (\alpha_1, \ldots, \alpha_{k-2}, \alpha, \alpha)\}}$$

$$= \frac{\sum_\beta P\{A_k(t-\delta) = (\alpha_1, \ldots, \alpha_{k-2}, \alpha, \beta)\}\delta}{P\{A_k(t) = (\alpha_1, \ldots, \alpha_{k-2}, \alpha, \alpha)\}} + o(\delta)$$

$$= \frac{\delta}{\pi(\alpha|A_k - \alpha)} + o(\delta).$$

Again, by exchangeability every pair of lineages of type α coalesces at this rate, and so the total rate is obtained by multiplying by $n_\alpha(n_\alpha - 1)/2$.

These results imply that $Q^*_\theta(\cdot)$ is in $\mathcal{M}$ and give equations (15) up to the constant of proportionality C. The value of C follows from the fact that events must occur at total rate $k(k-1+\theta)/2$ when there are k lineages in the ancestry (see Stephens (2000)). □

Theorem 1 shows how to find the optimal IS function Q^*_θ from the conditional probabilities $\pi(\cdot|\cdot)$. We have seen at equation (11) that knowledge of Q^*_θ is equivalent to knowledge of the likelihood $L(\theta)$. It is also easy to see that knowing the conditional probabilities $\pi(\cdot|\cdot)$ appearing in equations (15) is equivalent to knowing $L(\theta)$. Not surprisingly then, these conditional probabilities cannot be found explicitly in most cases of interest. We can, however, gain considerable insight into their structure from knowledge of the underlying genealogical processes. Theorem 1 then has two helpful consequences. For any particular proposal distribution $Q_\theta \in \mathcal{M}$, such as that used by Griffiths and Tavaré, we can assess the associated q_θ in the light of equations (15) to gain insight into when, or whether, it may behave well. Secondly, we can hope to construct well-behaved proposal distributions by developing good approximations for $\pi(\cdot|\cdot)$ and substituting these into equations (15) (renormalizing if necessary).

We now consider the conditional probabilities $\pi(\cdot|\cdot)$ in more detail. These are known exactly for the special case of *parent-independent mutation* (PIM), in which the type of a mutant does not depend on the type of its parent:

$$P_{\alpha\beta} = P_\beta \qquad \text{for all } \alpha, \beta.$$

In this case

$$\pi(\beta|A_n) = \frac{n_\beta + \theta P_\beta}{n + \theta}, \tag{16}$$

where n_β is the number of chromosomes of type β in A_n. (See Hoppe (1984) for the infinite alleles case; the general result then follows as in Donnelly (1986).) Equivalently, with probability

$n/(n+\theta)$ the type of the $(n+1)$th sampled chromosome is given by that of a chromosome sampled uniformly at random from A_n, and with probability $\theta/(n+\theta)$ its type is $\beta \in E$ with probability P_β.

In the more general setting in which the type of a mutant depends on the type of its parent, this simple structure is lost. Nevertheless we would expect types which have high frequency in A_n to have reasonable probability in $\pi(\cdot|A_n)$. Among types that are not present in the sample there are two different effects. Loosely speaking, those types which are closer (in the sense of requiring fewer mutational changes) to the sample should have higher conditional probability in $\pi(\cdot|A_n)$ than those which are more distant. In contrast, types which are more likely under the stationary distribution of the mutation process should have higher conditional probability in $\pi(\cdot|A_n)$ than those which are not. These two effects may work in different directions; the relative importance of the second effect increases as the mutation rate increases. (In the limit as $\theta \to \infty$, $\pi(\cdot|A_n)$ converges to the stationary distribution of the mutation process for any A_n.)

It follows from the discussion in the previous paragraph and theorem 1 that in the optimal proposal distribution there will be a tendency for mutations to occur towards the rest of the sample, and that coalescences of 'unlikely' types are more likely than those of 'likely' types. The proposal distribution (13) underlying the Griffiths–Tavaré approach does not, in general, enjoy these properties. As an example suppose that $E = \{1, 2, \ldots, K\}$ (where K is large), and that the mutation process is a symmetric random walk with reflecting boundaries. If $A_n = (5, 5, 5, 5, 5, 5, 11)$ then, under the Griffiths–Tavaré proposal distribution, if the most recent event is a mutation which gave rise to the allele 11, the progenitor is equally likely to be 10 or 12. It is clear intuitively here, and borne out by the form of equations (15), that under the optimal proposal distribution (unless θ is large) the progenitor is more likely to be 10. Similarly, under the optimal proposal distribution the progenitor of this 10 allele is more likely to be 9 than 11, and so on, whereas under Q^{GT} each such choice remains equally likely. As a consequence histories sampled from Q^{GT} will tend to involve lengthy excursions through unlikely configurations. These histories will be computationally expensive and contribute negligible importance weights. This kind of effect will be most marked in settings where (as in the example just given) there are few long-term constraints on the types of ancestral alleles, and it explains the instances of unreliable performance by their method in our examples (Section 5) for mutation models other than infinite sites.

Having reflected on the underlying genealogical processes, we propose the following approximation to the conditional sampling probabilities.

Definition 1. Let $\hat{\pi}(\cdot|A_n)$ be the distribution which is defined by choosing an individual from A_n uniformly at random, and then mutating it according to the mutation matrix P a geometric number of times, with parameter $\theta/(n+\theta)$, i.e.

$$\hat{\pi}(\beta|A_n) = \sum_{\alpha \in E} \sum_{m=0}^{\infty} \frac{n_\alpha}{n} \left(\frac{\theta}{n+\theta}\right)^m \frac{n}{n+\theta} (P^m)_{\alpha\beta}. \tag{17}$$

We now summarize some of the properties of $\hat{\pi}(\cdot|A_n)$ when considered as an approximation to $\pi(\cdot|A_n)$.

Proposition 1. The distribution $\hat{\pi}(\cdot|A_n)$ defined by definition 1 has the following properties.

(a) $\hat{\pi}(\cdot|A_n) = \pi(\cdot|A_n)$, for all A_n, in the case of PIM.
(b) $\hat{\pi}(\cdot|A_n) = \pi(\cdot|A_n)$ for the case $n = 1$, provided that P is reversible.
(c) The distribution $\hat{\pi}(\cdot|A_n)$ is of the form

$$\hat{\pi}(\beta|A_n) = \sum_{\alpha \in E} \frac{n_\alpha}{n} M^{(n)}_{\alpha\beta} \tag{18}$$

for some $M^{(n)}$, i.e. a chromosome β may be sampled from $\hat{\pi}(\cdot|A_n)$ by first sampling a chromosome uniformly at random from A_n, and then choosing β from a distribution which depends only on the sample size n, and on the type of the sampled chromosome, and not otherwise on A_n. In our case

$$M^{(n)} = (1 - \lambda_n)(I - \lambda_n P)^{-1} \tag{19}$$

where $\lambda_n = \theta/(n + \theta)$.

(d) The approximation $\hat{\pi}$ is the only distribution of the form (18) which both satisfies condition (b) above and is consistent in the sense that the conditional distribution of the $(n + 1)$th observation given the first n observations is the same as the (marginal) conditional distribution of the $(n + 2)$th given the first n observations:

$$\hat{\pi}(\beta|A_n) = \sum_{\alpha \in E} \hat{\pi}(\alpha|A_n)\, \hat{\pi}\{\beta|(A_n, \alpha)\} \tag{20}$$

for all A_n. (We note in passing that in general our approximation lacks the stronger property $\hat{\pi}(\alpha|A_n)\, \hat{\pi}\{\beta|(A_n, \alpha)\} = \hat{\pi}(\beta|A_n)\, \hat{\pi}\{\alpha|(A_n, \beta)\}$.)

(e) The distribution $\hat{\pi}(\cdot|A_n)$ is the stationary distribution of the Markov chain on E with transition matrix

$$T_{\beta\alpha} = \frac{\theta}{n + \theta} P_{\beta\alpha} + \frac{n_\alpha}{n + \theta}, \tag{21}$$

i.e.

$$\hat{\pi}(\alpha|A_n) = \sum_{\beta \in E} \hat{\pi}(\beta|A_n) \left(\frac{\theta}{n + \theta} P_{\beta\alpha} + \frac{n_\alpha}{n + \theta} \right). \tag{22}$$

Remark 1. Properties (a), (b) and (d) give grounds for optimism that $\hat{\pi}$ is a sensible approximation of π. Property (c) can be computationally very convenient, as it means that $\hat{\pi}(\cdot|H)$ may be computed efficiently for any H with fewer than n chromosomes, once the matrices $M^{(1)}, M^{(2)}, \ldots, M^{(n)}$ have been found, which need only be done once. (Appendix A exploits further computationally attractive features of $\hat{\pi}(\cdot|\cdot)$ to facilitate implementation when these matrices are not easily found.) Property (e) provides a characterization of π which is, in some more general settings not considered in this paper, more amenable to generalization than definition 1. The Markov transition matrix (21) is the transition matrix associated with the $(n + 1)$th line of the particle representation described in Section 4 (or the Moran model) when the first n lines are fixed to be A_n.

Proof of property (a). The conditional probabilities $\pi(\cdot|A_n)$ for PIM are defined at equation (16), and the result follows from the fact that, for PIM, $P^m = P$ for $m = 1, 2, \ldots$.

Proof of property (b). Consider the coalescence tree which describes the ancestry of two sampled chromosomes (labelled 1 and 2), and denote by m_1 and m_2 respectively the number of mutations between the MRCA and chromosomes 1 and 2. It follows from the reversibility of P that at stationarity, conditionally on m_1 and the type α_1 of chromosome 1, the type of the MRCA is β with probability $(P^{m_1})_{\alpha_1\beta}$. Conditionally on the type γ of the MRCA, and m_2, the type of chromosome 2 is β with probability $(P^{m_2})_{\gamma\beta}$. Thus, conditionally on α_1, m_1 and

m_2, the type of chromosome 2 is β with probability $(P^{m_1+m_2})_{\alpha_1\beta}$. The result follows from the fact that $m_1 + m_2$ is geometric with parameter $\theta/(1+\theta)$, independently of α_1.

Proof of property (c). Definition 1 gives

$$\begin{aligned}\hat{\pi}(\beta|A_n) &= \sum_{\alpha\in E}\sum_{m=0}^{\infty}\frac{n_\alpha}{n}\left(\frac{\theta}{n+\theta}\right)^m\frac{n}{n+\theta}(P^m)_{\alpha\beta}\\ &= \sum_{\alpha\in E}\sum_{m=0}^{\infty}\frac{n_\alpha}{n}(1-\lambda_n)(\lambda_n P)^m_{\alpha\beta}\\ &= \sum_{\alpha\in E}\frac{n_\alpha}{n}(1-\lambda_n)(I-\lambda_n P)^{-1}_{\alpha\beta},\end{aligned}$$

from which the result follows.

Proof of property (d). Suppose that $\tilde{\pi}$ is an approximation of the form (18):

$$\tilde{\pi}(\beta|A_n) = \sum_{\alpha}\frac{n_\alpha}{n}M^{(n)}_{\alpha\beta}.$$

Define η $(=\eta(A_n))$ to be the vector (n_α/n), so $\tilde{\pi}(\beta|A_n) = [\eta M^{(n)}]_\beta$. Substituting into equation (20) gives

$$\begin{aligned}[\eta M^{(n)}]_\beta &= \sum_{\alpha\in E}\tilde{\pi}(\alpha|A_n)\,\tilde{\pi}\{\beta|(A_n,\alpha)\}\\ &= \sum_{\alpha\in E}\sum_{\gamma\in E}\frac{n_\gamma}{n}M^{(n)}_{\gamma\alpha}\sum_{\xi\in E}\frac{n_\xi+\delta_{\alpha\xi}}{n+1}M^{(n+1)}_{\xi\beta}\\ &= \sum_{\gamma\in E}\sum_{\xi\in E}\left\{\frac{n_\gamma}{n}\frac{n_\xi}{n+1}M^{(n+1)}_{\xi\beta}+\frac{n_\gamma}{n(n+1)}M^{(n)}_{\gamma\xi}M^{(n+1)}_{\xi\beta}\right\}\\ &= \frac{1}{n+1}[\eta(nM^{(n+1)}+M^{(n)}M^{(n+1)})]_\beta.\end{aligned}$$

This must hold for all η (since equation (20) holds for all A_n), and so

$$(n+1)M^{(n)} = nM^{(n+1)} + M^{(n)}M^{(n+1)}. \tag{23}$$

If $\tilde{\pi}$ satisfies property (b) of the proposition, then $M^{(1)} = (1-\lambda_1)(I-\lambda_1 P)^{-1}$. It then follows from equation (23) and mathematical induction that $M^{(n)} = (1-\lambda_n)(I-\lambda_n P)^{-1}$, and so $\tilde{\pi} = \hat{\pi}$.

Proof of property (e). Expressions (18) and (19) give

$$\hat{\pi}(\alpha|H) = \sum_{\beta\in E}\frac{n_\beta}{n}\frac{n}{n+\theta}\left(I-\frac{\theta}{n+\theta}P\right)^{-1}_{\beta\alpha}. \tag{24}$$

For notational convenience write $\hat{\pi}_\alpha$ for $\hat{\pi}(\alpha|H)$, and write $\hat{\pi}$ for the vector $(\hat{\pi}_\alpha)$. Then equation (24) becomes

$$\hat{\pi} = \frac{n}{n+\theta}\eta\left(I-\frac{\theta}{n+\theta}P\right)^{-1},$$

and on post-multiplying each side by $I-\{\theta/(n+\theta)\}P$ and rearranging we obtain

$$\hat{\pi} = \frac{\theta}{n+\theta}\hat{\pi}P + \frac{n}{n+\theta}\eta,$$

from which equation (22) follows. □

We now define our proposal distribution and describe how it may be sampled from efficiently.

Definition 2. Define the proposal distribution Q_θ^{SD} as the distribution in $\mathcal{M}$ corresponding to the backward transition probabilities $\hat{q}_\theta$ obtained by substituting $\hat{\pi}(\cdot|\cdot)$ for $\pi(\cdot|\cdot)$ into equations (15):

$$\hat{q}_\theta(H_{i-1}|H_i) = \begin{cases} C^{-1}\dfrac{\theta}{2}n_\alpha\dfrac{\hat{\pi}(\beta|H_i-\alpha)}{\hat{\pi}(\alpha|H_i-\alpha)}P_{\beta\alpha} & \text{if } H_{i-1} = H_i - \alpha + \beta, \\ C^{-1}\binom{n_\alpha}{2}\dfrac{1}{\hat{\pi}(\alpha|H_i-\alpha)} & \text{if } H_{i-1} = H_i - \alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{25}$$

where n_α is the number of chromosomes of type α in H_i and $C = n(n-1+\theta)/2$ where n is the number of chromosomes in H_i.

Proposition 2. For any given H_i, the backward transition probabilities $\hat{q}_\theta(H_{i-1}|H_i)$ defined by equation (25) sum to 1 and may be sampled from efficiently as follows.

(a) Choose a chromosome uniformly at random from those in H_i. Denote the type of the chosen chromosome by α.
(b) For each type $\beta \in E$ for which $P_{\beta\alpha} > 0$, calculate $\hat{\pi}(\beta|H_i - \alpha)$ from equation (18).
(c) Sample H_{i-1} by setting

$$H_{i-1} = \begin{cases} H_i - \alpha + \beta \text{ with probability proportional to } \theta\hat{\pi}(\beta|H_i-\alpha)P_{\beta\alpha}, \\ H_i - \alpha \text{ with probability proportional to } n_\alpha - 1. \end{cases}$$

Remark 2. The proof of proposition 2 involves showing that, under the transition probabilities $\hat{q}_\theta(H_{i-1}|H_i)$, the probability that a randomly sampled transition involves either a mutation from a chromosome of type α or a coalescence of two chromosomes of type α is n_α/n. The sampling algorithm described makes use of this feature to reduce the number of pairs (α, β) for which $\hat{\pi}(\beta|H_i - \alpha)$ must be calculated, by first sampling the type α of a chromosome which is involved in the transition, and then sampling the exact nature of the transition (a mutation to type β, or a coalescence with another chromosome of type α), from those possible. This, combined with the fact that $\hat{\pi}(\beta|H_i - \alpha)$ is easy to calculate (see remark 1), makes Q_θ^{SD} very efficient to simulate from (in general much more efficient than the Griffiths–Tavaré proposal distribution, particularly where the number of possible types is large).

Proof. The results follow directly from the definition of the backward transition probabilities $\hat{q}_\theta(H_{i-1}|H_i)$ given in equation (25), provided that we can show that the probability that a randomly sampled transition involves either a mutation from a chromosome of type α or a coalescence of two chromosomes of type α equals n_α/n.

The probability that a transition randomly sampled from $\hat{q}_\theta(H_{i-1}|H_i)$ involves a mutation

from a chromosome of type α is given by

$$p_{\mathrm{m}}(\alpha) = C^{-1} \sum_{\beta \in E} \frac{\theta}{2} n_\alpha \frac{\hat{\pi}(\beta|H_i - \alpha)}{\hat{\pi}(\alpha|H_i - \alpha)} P_{\beta\alpha}, \tag{26}$$

and the probability that it involves a coalescence of two chromosomes of type α is given by

$$p_{\mathrm{c}}(\alpha) = C^{-1} \binom{n_\alpha}{2} \frac{1}{\hat{\pi}(\alpha|H_i - \alpha)}, \tag{27}$$

where n_α is the number of chromosomes of type α in H_i. We need to show that $p_{\mathrm{m}}(\alpha) + p_{\mathrm{c}}(\alpha) = n_\alpha/n$.

From equation (22) we have

$$1 = \sum_{\beta \in E} \frac{\hat{\pi}(\beta|H_i - \alpha)}{\hat{\pi}(\alpha|H_i - \alpha)} \left(\frac{\theta}{n - 1 + \theta} P_{\beta\alpha} + \frac{n_\alpha - 1}{n - 1 + \theta} \right).$$

Thus

$$\begin{aligned} \frac{n_\alpha}{n} &= \sum_{\beta \in E} \frac{\hat{\pi}(\beta|H_i - \alpha)}{\hat{\pi}(\alpha|H_i - \alpha)} \left(\frac{\theta}{n - 1 + \theta} \frac{n_\alpha}{n} P_{\beta\alpha} + \frac{n_\alpha}{n} \frac{n_\alpha - 1}{n - 1 + \theta} \right) \\ &= \sum_{\beta \in E} \frac{\hat{\pi}(\beta|H_i - \alpha)}{\hat{\pi}(\alpha|H_i - \alpha)} C^{-1} \frac{\theta}{2} n_\alpha P_{\beta\alpha} + \frac{1}{\hat{\pi}(\alpha|H_i - \alpha)} C^{-1} \binom{n_\alpha}{2} \\ &= p_{\mathrm{m}}(\alpha) + p_{\mathrm{c}}(\alpha). \end{aligned}$$

□

5. Applications

We now illustrate our method on some examples and compare its performance with the Griffiths–Tavaré scheme and MCMC schemes devised by Kuhner *et al.* (1995, 1998) and Wilson and Balding (1998). All the methods that we consider naturally provide estimators of likelihood or relative likelihood surfaces. From a classical perspective it would be more natural to consider estimating and plotting log-likelihood surfaces. However, because of the limited information in data, it is at best unclear whether the classical asymptotic theory relating to the interpretation of the log-likelihood applies in genetics settings of interest. (Further work on this question would be welcomed.) From a Bayesian viewpoint, the likelihood is of more natural interest than is the log-likelihood. In part for these reasons, and in part to facilitate a direct comparison with published estimates, we focus here on likelihood and relative likelihood estimation.

Our comparisons, with one exception, are made on relatively small problems, so that by using a large number of iterations of our method we believe that we can estimate likelihoods accurately. Methods may then be assessed on the basis of the accuracy of likelihood estimates obtained by using fewer iterations (where possible, these estimates were obtained either directly from the literature or by applying each method according to published general guidelines given by the authors). To allow a fair comparison of the relative efficiency of the methods used, we have tried where possible to give (in table and figure captions) an idea of the central processor unit time (on a Sun Ultra-Sparc 200 workstation) required to produce the data shown. Even for these small problems an accurate estimation of the likelihood can involve non-trivial amounts of computing, and modern computer power puts us at a distinct advantage over practitioners of earlier years.

Methods which give reasonably accurate estimates of the likelihood may be more stringently compared, by assessing the degree of uncertainty associated with the estimates due to sampling variance. For the IS schemes, assuming that the distribution of the weights has finite variance σ^2, then (by the central limit theorem) the estimator (8) is asymptotically normal with variance σ^2/M. A natural measure of the variability is then given by the standard error $\hat{\sigma}/\sqrt{M}$, where $\hat{\sigma}^2$ is the sample variance of the M weights. However, caution is necessary, as even assuming finite variance (which is not guaranteed) the distribution of the weights may be so highly skewed that, even for very large M, σ^2 is (with high probability) underestimated by $\hat{\sigma}^2$, and/or the normal asymptotic theory does not apply. Despite this important *caveat*, we quote standard errors in some of our examples to allow a direct comparison with published estimates. For both MCMC schemes and IS schemes, an alternative and more reliable (though computationally more expensive) assessment of sampling variability can be obtained by comparing the results of different runs using different seeds for the pseudorandom number generator. We plot results obtained from several different runs for some of our examples, to give an indication of sampling variability, though we have not performed the larger scale simulation studies that are necessary to obtain accurate estimates of Monte Carlo errors. Further discussion on methods of estimating sampling variability is deferred to Section 6.

5.1. Parent-independent mutation

For PIM models, the conditional distributions $\pi(\cdot|\cdot)$ are known exactly (see equation (16) above), and hence so also is the likelihood. This class of models thus provides a convenient check on the correctness of the implementation of any algorithm for estimating the likelihood. In our case, it follows from proposition 1 (property (a)) that for PIM our proposal distribution Q_θ^{SD} is exactly the optimal proposal distribution $Q_\theta^*(\mathcal{H}) = P_\theta(\mathcal{H}|A_n)$. Thus our IS estimator (9) should have zero variance and give the exact value for the likelihood, which in fact it does.

Our focus throughout the paper is inference for θ. Although this is trivial for PIM, questions of ancestral inference, i.e. inference for aspects of the ancestry of the sample, remain apparently non-trivial in this context. However, for PIM our IS is actually independent sampling from the full conditional distribution of the history of the sample and thus represents an extremely efficient solution to questions of ancestral inference for these kinds of data.

5.2. Simulated sequence data

We consider data given in Griffiths and Tavaré (1994a), who simulated sequences of length 10 from a model with only two possible nucleotides at each sequence position (so $E = \{1, 2\}^{10}$). Mutations were assumed to occur at total rate $\theta/2$, with the location of each mutation being chosen independently and uniformly along the sequence. The transition matrix governing mutations at each position was

$$P = \begin{pmatrix} 0.5 & 0.5 \\ 0.1 & 0.9 \end{pmatrix}. \tag{28}$$

Note that this model has 2^{10} different alleles, and so the calculation of the quantities $\hat{\pi}(\beta|A_n)$ using equations (18) and (19) appears to be computationally daunting. In fact the special structure of this model allows an efficient approximation of these quantities, as described in Appendix A.

Table 1. Comparison of estimated likelihoods (with standard errors in parentheses) obtained by using the IS functions Q_θ^{GT} and Q_θ^{SD}, for simulated data from Griffiths and Tavaré (1994a), described in Section 5.2†

θ	Q_θ^{GT} (20000 samples)	Q_θ^{SD} (20000 samples)	Q_θ^{SD} (10^7 samples)
2.0	7.34×10^{-6} (1.35×10^{-6})	7.33×10^{-6} (4.01×10^{-8})	7.29×10^{-6} (1.87×10^{-9})
10.0	6.96×10^{-9} (3.52×10^{-9})	3.10×10^{-9} (2.53×10^{-11})	3.09×10^{-9} (1.50×10^{-12})
15.0	2.41×10^{-19} (1.89×10^{-19})	4.74×10^{-12} (1.39×10^{-13})	5.37×10^{-12} (8.08×10^{-14})

†The second column comes from Griffiths and Tavaré (1994a), Table VI. The calculations required to produce the third column took less than 2 min per row on average.

Griffiths and Tavaré simulated three sets of 10 sequences from the model, each set being produced using a different value of θ ($\theta = 2.0, 10.0, 15.0$). In each case they used 20000 iterations of their algorithm (or equivalently 20000 samples from their IS function Q_θ^{GT}) to estimate the likelihood (together with a standard error). For comparison we obtained corresponding estimates by using 20000 and 10 million samples from the proposal distribution Q_θ^{SD}. The results are shown in Table 1.

We can assess the accuracy of the estimates obtained from the shorter runs by comparing them with those obtained from the longer run. For the data generated using $\theta = 2.0$, 20000 samples from either Q_θ^{GT} or Q_θ^{SD} appear to produce a reasonably accurate estimate of the likelihood. For $\theta = 10.0$, 20000 samples from Q_θ^{SD} are sufficient to estimate the likelihood accurately, but 20000 samples from Q_θ^{GT} are not. For $\theta = 15.0$, although neither method produces a very accurate estimate of the likelihood using 20000 samples, the samples from Q_θ^{GT} underestimate the likelihood by seven orders of magnitude. The IS interpretation of the Griffiths–Tavaré method helps to explain this rather startling observation. Griffiths and Tavaré (1994a) noted that, for the data generated with $\theta = 10.0$, effectively only nine of their sampled histories contribute to their estimate of the likelihood, and that many of their sampled histories contained huge numbers of mutations. We conclude from this that their IS function puts too much weight on histories which have a large number of mutations and are unlikely under the posterior distribution $P_\theta(\mathcal{H}|A_n)$, and correspondingly too little weight on histories with a smaller number of mutations. This results in a very skewed distribution for the importance weights, which are very small with high probability, but are occasionally very large, producing a highly variable estimator. For $\theta = 15.0$, the distribution of the importance weights is so skewed that only very small (effectively negligible) importance weights are observed in the 20000 iterations (none of their sampled histories had fewer than 20 mutations, and most had many more, when a minimum of six are required to produce the data). As a result the sample mean of the importance weights severely underestimates the true mean (the likelihood).

Although in principle the accuracy could be improved with more iterations, in our own implementation of the Griffiths–Tavaré proposal distribution, 1 million samples took 72 h and gave an estimate of the likelihood (dominated by a single large importance weight) of 4.93×10^{-13}. In contrast, 20000 iterations of a modified Griffiths–Tavaré scheme (implemented in the computer program SEQUENCE, kindly provided by R. C. Griffiths), which includes some of the computational tricks described in Griffiths and Tavaré (1994a), took 23 min and produced an estimate of 4.67×10^{-12}, with an estimated standard error of 8.72×10^{-13}. The lesson is clear: longer runs are not always a satisfactory replacement for better methods.

Examining the standard errors of the estimates in Table 1, it is notable that the standard error obtained from 20000 iterations of the Q_θ^{GT} with $\theta = 15.0$ does not adequately reflect the

uncertainty in the estimated likelihood. This is another effect of the severely skewed distribution of the importance weights in this case. A failure to observe rare, large, weights not only leads to severe underestimation of the likelihood but also means that calculated standard errors can seriously underestimate the standard deviation of the importance weights and hence give an extremely misleading impression of the accuracy of the algorithm.

We can use the standard errors from the long run of Q_θ^{SD} to check whether the standard errors for the shorter run accurately reflect the standard deviation of the importance weights. If they do then the longer run should result in standard errors which are smaller by a factor of $\sqrt{500} \approx 22$. For $\theta = 2.0$ and $\theta = 10.0$ the changes in the estimated standard error (by factors of about 21 and 17 respectively) between the long and short runs for Q_θ^{SD} suggest that these standard errors are being estimated reasonably accurately. In contrast, between the long and short runs for Q_θ^{SD} with $\theta = 15.0$, the change in the estimated standard error is less than a factor of 2, indicating that (at least for the short run) the standard error severely underestimates the standard deviation in this case.

5.3. Likelihood surfaces for sequence data

Griffiths and Tavaré (1994a) also used their method to estimate likelihood curves for simulated data, consisting of 50 sequences of length 20, using the same model as above, but with $\theta = 1.0$ and

$$P = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}. \tag{29}$$

Figs 2(a) and 2(b) show a comparison of the likelihood surface obtained by Griffiths and Tavaré (1994a), Table IX, using 10000 samples from $Q_{\theta_0=1.0}^{\mathrm{GT}}$, with the likelihood surface that we obtain using 10000 samples from the IS function $Q_{\theta_0=1.0}^{\mathrm{SD}}$. The standard errors for the

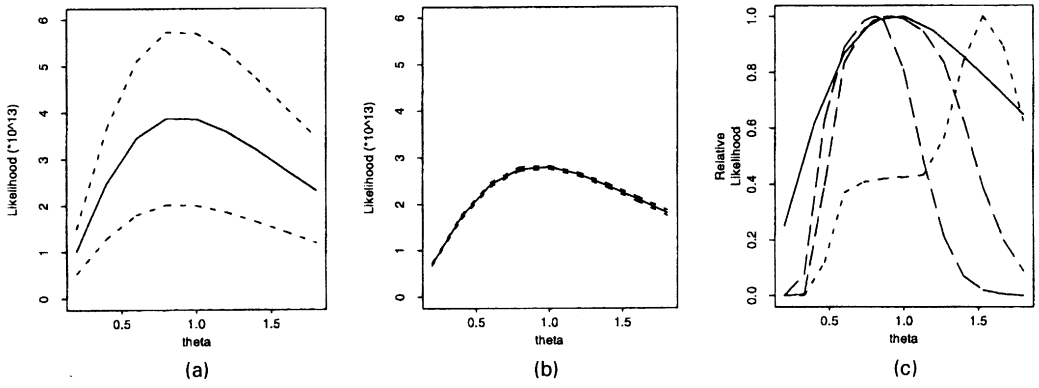

Fig. 2. Comparison of estimated likelihood and relative likelihood surfaces obtained for simulated sequence data from Griffiths and Tavaré (1994a), described in Section 5.3: (a) likelihood surface estimate (———) with ±2 standard deviations (- - - - -) obtained by using 10000 samples from IS function $Q_{\theta_0=1.0}^{\mathrm{GT}}$; (b) the same, by using 10000 samples from IS function $Q_{\theta_0=1.0}^{\mathrm{SD}}$; (c) relative likelihood surfaces (each scaled to have a maximum of 1.0) obtained by using 10000 samples from $Q_{\theta_0=1.0}^{\mathrm{SD}}$ (———) and the `Fluctuate` program of Kuhner *et al.* (1998) using long chains of length 50000 iterations (- - - - -) and 1 million iterations (— — —, corresponding to two different initial seeds, which led `Fluctuate` to select final driving values of θ_0 near 0.6 and 1.0 respectively) (as the data contained only two 'nucleotides', `Fluctuate` was used as in the analysis of purine data in Kuhner *et al.* (1995); all runs were started from the same starting tree obtained from the file `besttree` output by `Fluctuate` during a preliminary run (from a random starting tree) of five short chains of length 10000 and two long chains of length 50000)

surface obtained using $Q^{SD}_{\theta_0=1.0}$ are substantially smaller, suggesting that this is the more efficient of the two proposal distributions in this case. We confirmed this by comparing the estimated surfaces with a surface estimated using 10 million samples from $Q^{SD}_{\theta_0=1.0}$. The estimated likelihood values using this larger sample were very close to those in Fig. 2(b) (data not shown), and the standard errors were smaller by about a factor of 31, suggesting that the standard errors in the smaller sample are accurately reflecting the standard deviation of the weights (they should in theory be reduced by a factor of $\sqrt{1000} \approx 32$).

We also compared the results obtained by our IS method with those obtained by using the MCMC scheme developed by Kuhner *et al.* (1995, 1998), implemented in their program `Fluctuate` (available from

`http://www.evolution.genetics.washington.edu/lamarc.html`).

Instead of considering the typed ancestry of the sample as missing data, they considered the unknown *scaled genealogical tree* $\widetilde{\mathcal{G}}$, i.e. the genealogical tree $\mathcal{G}$ with branch lengths scaled by the mutation rate $\theta/2$, so that mutations may be assumed to occur at unit rate along the branches. Thus $P_\theta(\widetilde{\mathcal{G}})$ depends on θ and is easy to calculate from the coalescence prior, whereas $P_\theta(A_n|\widetilde{\mathcal{G}})$ is independent of θ and may be calculated efficiently by using the peeling algorithm (Felsenstein, 1981). Kuhner *et al.* (1995) used a Metropolis–Hastings algorithm to construct a Markov chain with stationary distribution $P_{\theta_0}(\widetilde{\mathcal{G}}|A_n)$ and estimated the relative likelihood surface $L(\theta)/L(\theta_0)$ by using equation (7).

Kuhner *et al.* (1998) suggested running several short MCMC runs (of a few thousand iterations each), with each run being used to estimate a relative likelihood surface, and the θ which maximizes this surface being used as the driving value θ_0 for the next chain. One or more long chains would then be run in the same way to obtain more accurate estimates of the likelihood surface. We found that using this method with chains of the kind of lengths suggested (specifically five short chains of 10000 iterations, followed by two long chains of 50000 iterations) could give very inaccurate estimates of the likelihood surface (Fig. 2(c)). Increasing the lengths of the long chains to 1 million iterations improved the accuracy, but the surfaces obtained still tended to underestimate the relative likelihood away from the driving values used, thus giving a false impression of the tightness and/or position of the peak of the likelihood surface, presumably because $P_{\theta_0}(\widetilde{\mathcal{G}}|A_n)$ is a poor approximation to $P_\theta(\widetilde{\mathcal{G}}|A_n)$ in this range. (In fact Stephens (1999) has recently shown that for $\theta > 2\theta_0$ the estimator has infinite variance.)

5.4. Microsatellites

At microsatellite loci, alleles consist of a number of repetitions of a short DNA motif. Alleles are conveniently defined by counting the number of repeats. A commonly used mutation model is the so-called *stepwise* model in which mutations occur at rate $\theta/2$, regardless of allele length, and mutation either increases or decreases by 1 the number of repeats, with both possibilities being equally likely. Under this mutation model the joint distribution of sample configurations is invariant under the addition of any fixed number of repeats to each sampled allele (Moran, 1975).

The implementation of our IS scheme is facilitated by centring the sample distribution near 10 repeats and truncating the type space E to $\{0, 1, \ldots, 19\}$ by insisting that all mutations to alleles of length 0 or 19 involve the gain or loss respectively of a single repeat. This truncation will make little difference to the likelihood of samples whose allele lengths are not too close to these boundaries. Nielsen (1997) implemented the Griffiths–Tavaré proposal distribution for

a similar model using more alleles. We consider estimating likelihood surfaces for a sample consisting of allele lengths {8, 11, 11, 11, 11, 12, 12, 12, 12, 13}. (This sample was obtained by adding five repeats to each allele of a random sample simulated under the stepwise model by Wilson and Balding (1998), shown in their Fig. 1.) Fig. 3 compares estimated likelihood surfaces obtained for these data by using our own implementation of the Griffiths–Tavaré proposal distribution $Q^{\mathrm{GT}}_{\theta_0=10.0}$ with those obtained by using our proposal distribution $Q^{\mathrm{SD}}_{\theta_0=10.0}$. It is clear from the variability exhibited in Fig. 3(a) that 10000 samples from $Q^{\mathrm{GT}}_{\theta_0=10.0}$ are not sufficient to obtain an accurate estimate of the likelihood surface. (Indeed, Fig. 3(c) indicates that 50000 samples are also insufficient.) In contrast, 10000 samples from $Q^{\mathrm{SD}}_{\theta_0=10.0}$ appear to suffice, demonstrating the increased efficiency of our proposal distribution.

Fig. 4 shows a comparison of estimated relative likelihood surfaces obtained by using our proposal distribution $Q^{\mathrm{SD}}_{\theta_0=10.0}$ and by using an MCMC scheme developed by Wilson and Balding (1998), implemented in their program `micsat` (available from

`http://www.maths.abdn.ac.uk/~ijw/`).

Their scheme is designed to perform a full Bayesian analysis, and `micsat` can be used to produce a sample from the posterior distribution of θ given the data. For an (improper) uniform prior on $\theta > 0$, this posterior distribution is proportional to the likelihood, and so a relative likelihood curve can be obtained by smoothing a sample from this posterior distribution. A comparison of the accuracy and variability of the estimated relative likelihood surfaces, together with a consideration of the computer time required to produce these results (see the caption of Fig. 4), suggests that for this problem our IS method is considerably more efficient than `micsat`, although there are many ways in which our use of the MCMC scheme could be improved (for example, the parameters of the MCMC scheme could be tuned to achieve better mixing over θ; we used the default values).

As a more challenging example we also applied our method to the so-called NSE data set considered by Wilson and Balding (1998) and supplied with `micsat`. The data are a subset of those given by Cooper *et al.* (1996) and consist of 60 males from Nigeria, Sardinia and

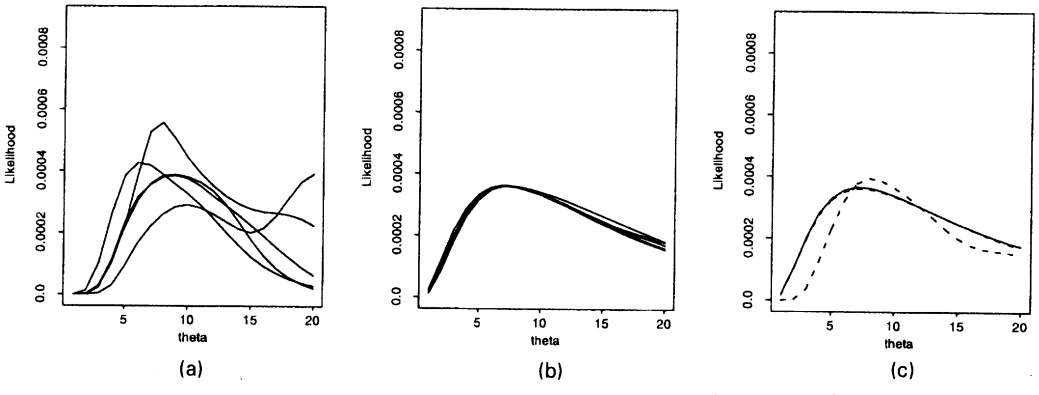

Fig. 3. Comparison of estimated likelihood surfaces obtained by using $Q^{\mathrm{GT}}_{\theta_0=10.0}$ and $Q^{\mathrm{SD}}_{\theta_0=10.0}$ for the simulated microsatellite data from Wilson and Balding (1998) described in Section 5.4: (a) five independent likelihood surface estimates, each obtained by using 10000 samples from the IS function $Q^{\mathrm{GT}}_{\theta_0=10.0}$ (each run took about 330 s); (b) the same, but using 10000 samples from the IS function $Q^{\mathrm{SD}}_{\theta_0=10.0}$ (each run took about 30 s); (c) an 'accurate' estimate of the likelihood surface obtained by using 10 million samples from the IS function $Q^{\mathrm{SD}}_{\theta_0=10.0}$ (———), and likelihood surface estimates obtained from the combined samples of size 50000 from $Q^{\mathrm{GT}}_{\theta_0=10.0}$ (- - - -) and $Q^{\mathrm{SD}}_{\theta_0=10.0}$ (·····, which is almost superimposed on ———)

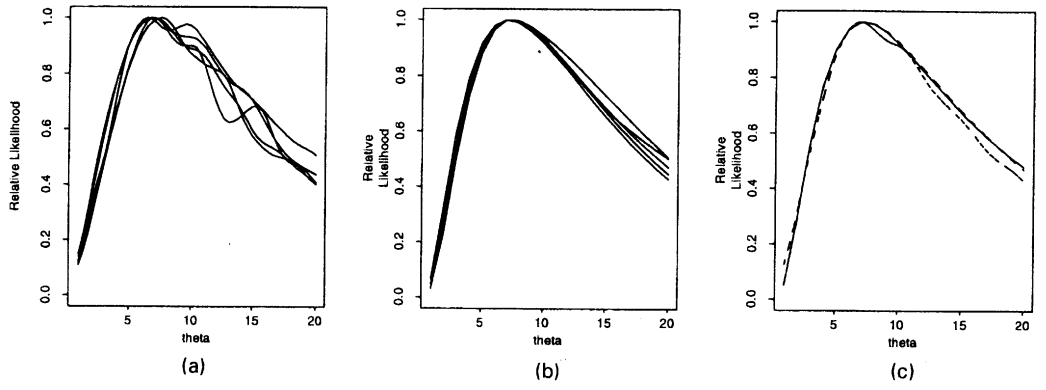

Fig. 4. Comparison of estimated relative likelihood surfaces obtained by using `micsat` and $Q^{SD}_{\theta_0=10.0}$ for the simulated microsatellite data from Wilson and Balding (1998) described in Section 5.4: (a) five independent relative likelihood surface estimates (each scaled to have a maximum of 1.0) obtained by smoothing a sample of size 10000 obtained from the posterior distribution of θ given an (improper) uniform prior, using the `micsat` program of Wilson and Balding (1998) (each run took about 1.5 h) (the density estimate obtained is sensitive to the smoothing method used; we used a kernel density smoother, with bandwidth chosen automatically according to a rule given by Sheather and Jones (1991) using the S routine `width.SJ` from Venables and Ripley (1997)); (b) the same, but using 10000 samples from the IS function $Q^{SD}_{\theta_0=10.0}$ (each run took about 30 s); (c) an 'accurate' estimate of the relative likelihood surface obtained by using 10 million samples from the IS function $Q^{SD}_{\theta_0=10.0}$ (———), and relative likelihood surface estimates obtained from the combined samples of size 50000 from `micsat` (- - - - -) and $Q^{SD}_{\theta_0=10.0}$ (·····, which is almost superimposed on ———)

East Anglia, each typed at five microsatellite loci on the Y-chromosome (so we assume no recombination). Following Wilson and Balding (1998), the loci are each assumed to mutate independently at the same rate, $\theta/2$, according to the stepwise model of mutation. The type space is large ($E = \{0, 1, \ldots, 19\}^5$) but the required backward transition probabilities may be efficiently approximated by the computational methods described in Appendix A. Fig. 5

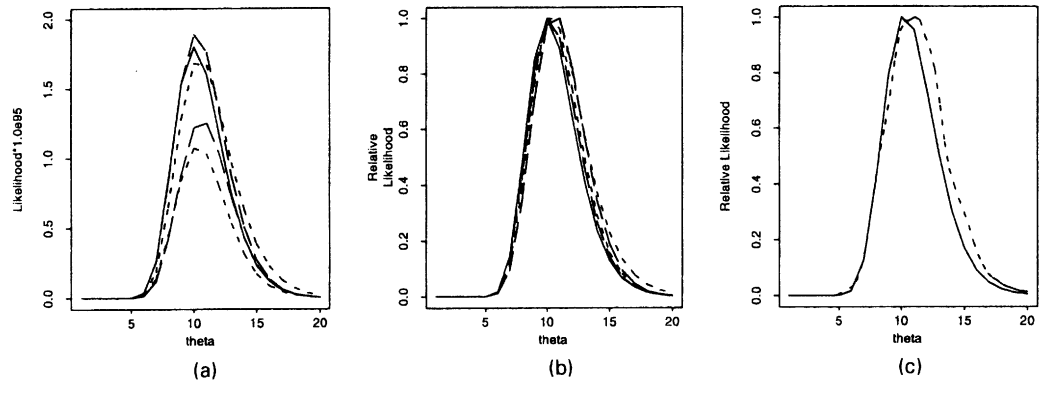

Fig. 5. Comparison of estimated likelihood and relative likelihood surfaces obtained for microsatellite data set NSE considered by Wilson and Balding (1998): (a) estimated likelihood surfaces obtained by using five independent samples of size 500000 from $Q^{SD}_{\theta_0=8.0}$ (each run took about 18 h); (b) relative likelihood surfaces (each scaled to have a maximum of 1.0) estimated by using the same five independent samples from $Q^{SD}_{\theta_0=8.0}$; (c) relative likelihood surfaces (each scaled to have a maximum of 1.0) obtained by using the combined sample of size 2.5 million from $Q^{SD}_{\theta_0=8.0}$(———), and by smoothing (using the method described in the caption to Fig. 4) a sample of size 10000 from the posterior distribution of θ given an (improper) uniform prior, obtained by using `micsat` (- - - - -), which took 4.5 h

shows the results obtained by using both our method and `micsat`. Five different estimated likelihood and relative likelihood curves were found by using our method, each based on half a million samples from $Q^{\mathrm{SD}}_{\theta_0=8.0}$. The variability of the estimated curves is notably larger for the absolute likelihoods than for the relative likelihoods, as might have been expected. There are small but noticeable differences in the relative likelihood curves obtained by using our method and `micsat`. Further investigation (more runs of each method) suggested that the curve obtained by using `micsat` is more accurate.

5.5. Infinite sites data

The *infinite sites* model of mutation is applicable to DNA sequence data and assumes that no two mutations occur at the same site. A rigorous formulation of the model is given by Ethier and Griffiths (1987). Briefly, the sequence is modelled by the unit interval [0, 1], and each mutation is assumed to occur at a position uniformly distributed along the sequence. Thus, with probability 1, all mutations occur at distinct sites. In many ways the infinite sites assumption simplifies modelling and analysis. Certain types of genetic data, notably samples of nuclear DNA sequences, are often consistent with this model.

For the infinite sites model the type space E becomes uncountably infinite and the conditional probabilities $\pi(\cdot|A_n)$ become densities. Technical challenges then arise when attempting to extend the theory of the previous sections to develop an efficient IS function for this model. Furthermore, the mutation process is not reversible. These problems are not insurmountable, but for simplicity we adapt our earlier approach to this context by analogy with proposition 2: recall that one method of simulating from our IS function Q^{SD} begins by choosing a chromosome uniformly at random from those present and assuming that this chromosome is involved in the most recent event backwards in time.

Under the infinite sites model any configuration of types H_i is equivalent to a unique 'unrooted gene tree' (Griffiths and Tavaré, 1995) which is a graph representing the relationships between the sequence types and mutations, as illustrated in Fig. 6. The chromosomes which may be involved in the most recent event backwards in time from H_i are limited:

(a) any chromosome which is not the only one of its type may coalesce with another of the same type and
(b) any chromosome which is the only one of its type and has only one neighbour on the unrooted gene tree corresponding to H_i may have arisen from a mutation to that neighbour.

For example, if H_i consists of the types a–e in Fig. 6, with two sequences of type a and one each of the other types, then the possible events are a coalescence of two of type a or one of the mutations 3, 4, 2 or 5. Since no chromosome can satisfy both these conditions, knowing the type of a chromosome which took part in the last event backwards in time is equivalent to knowing the event. Analogy with proposition 2 then suggests choosing the most recent event backwards in time by drawing a chromosome uniformly at random from those satisfying either (a) or (b). This procedure defines an IS function Q^{SD} which we note is independent of θ, removing the need to specify a driving value. The likelihood for the unrooted gene tree is estimated, as in equation (9), by the average of the importance weights.

The natural comparison for our method is with the Griffiths–Tavaré scheme for infinite sites data, described in Griffiths and Tavaré (1994b), and implemented in their program `genetree` available from

`http://www.stats.ox.ac.uk/mathgen/software.html`.

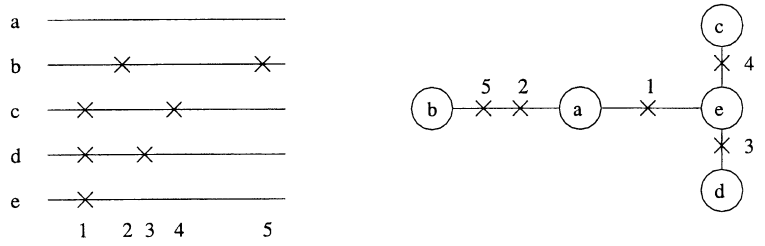

Fig. 6. Illustration of an unrooted gene tree (right) corresponding to sequence types a–e (left) assuming that they were generated under the infinite sites model; there are five segregating sites, labelled 1–5; the two bases at each segregating site are represented by a cross, and by no cross; in general it is not known which base is the original (*wild*) type, and which is the result of a mutation; note that the order of the mutations 2 and 5 is unknown, and their labelling in the gene tree is arbitrary

(We know of no published MCMC schemes for infinite sites data.) In fact they adopted a slightly different approach for infinite sites data, in that the support of their IS function is concentrated entirely on histories with a particular root (equivalently, a particular type for the MRCA). The mean of their importance weights thus estimates the joint probability of the data and a particular root, or in other words the probability of a particular rooted gene tree. In the absence of information on the position of the root (see below), the likelihood for the data is the probability of the unrooted gene tree, which is the sum of the probabilities associated with each possible rooted gene tree. (There are $S+1$ such rooted gene trees, where S is the number of mutations in the data.) We note that there is a natural efficiency gain in our method since it tends to concentrate sampling effort on the roots which contribute most to the likelihood, although similar gains could be achieved by using suitable adaptive strategies with methods which separately estimate the probability of each rooted gene tree.

If genetic data, O say, from related species ('outgroups') are available, then interspecies comparisons can be performed which typically provide a substantial amount of information on the type of the MRCA. Using IS as before gives

$$P_\theta(A_n, O) \approx \frac{1}{M} \sum_{i=1}^{M} \pi_\theta(A_n|\mathcal{H}^{(i)})\, P_\theta(O|\mathcal{H}^{(i)}) \frac{P_\theta(\mathcal{H}^{(i)})}{Q_\theta(\mathcal{H}^{(i)})}. \tag{30}$$

Typically it is reasonable to assume that $P_\theta(O|\mathcal{H})$ depends only on the root $r(\mathcal{H})$ of $\mathcal{H}$ and is otherwise independent of θ and $\mathcal{H}$. The optimal proposal distribution is then given by

$$Q^*_\theta(\mathcal{H}) \propto P_\theta(\mathcal{H}|A_n)\, P\{O|r(\mathcal{H})\}$$

and thus shifts towards histories with roots which are most consistent with the outgroup data. In extreme cases $P\{O|r(\mathcal{H})\}$ is non-zero for only one root, and so the optimal proposal distribution is concentrated entirely on histories with that particular root. It would be of interest to design efficient proposal distributions which take proper account of such information.

To facilitate a comparison with published estimates, we modified our IS function to analyse rooted trees, by adding to conditions (a) and (b) above a condition that no mutation can occur backwards in time from the type of the MRCA. Note that this does *not* take full account of the information contained in the position of the root, and the modified sampler is likely to be less efficient for roots which are relatively unlikely given the data. We believe that this property is shared by the Griffiths–Tavaré proposal distribution. It should be straightforward to design an importance sampler for rooted trees which makes more effective use

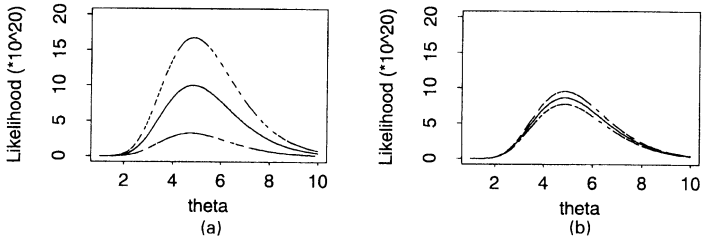

Fig. 7. Comparison of estimated likelihood surfaces obtained for infinite sites data considered by Griffiths and Tavaré (1994b) (their Table 2): (a) likelihood surface estimate (———) with ±2 standard deviations (- - - -) obtained by using 100 000 iterations of `genetree`, with $\theta_0 = 4.0$, which took 20 min; (b) the same, using 100 000 samples from the IS function Q^{SD}, which took 10 min

of information on the root position than does either of these samplers. Fig. 7 shows a comparison of likelihood surfaces obtained for data given in Griffiths and Tavaré (1994b), using 100 000 samples from Q^{SD}, and 100 000 samples from $Q^{\mathrm{GT}}_{\theta_0=4.0}$ (as implemented in `genetree`). The results suggest that our modified IS function is more efficient (with standard errors reduced by about an order of magnitude in this example). Similar results were obtained on a variety of other examples, with efficiency gains typically being greater for larger data sets (not shown.)

It should be noted that IS methods remain practicable for reasonably large infinite sites data sets, presumably because the space of possible histories is smaller in this context. Further, our experience is that, even when the estimate of the likelihood surface is poor, the corresponding estimate of the relative likelihood surface is often accurate. This might be explained by the observation (R. C. Griffiths, personal communication) that, as a function of θ, the shape of $P_\theta(\mathcal{H})$ normalized by its maximum seems to be similar for many $\mathcal{H}$. A more thorough investigation of this would be welcome.

6. Conclusions

6.1. General

The examples in the previous section demonstrate the substantial gains in efficiency which can be obtained by viewing the Griffiths–Tavaré scheme as IS and designing a more efficient proposal distribution. Placing the method in the context of IS can also be helpful in extending it to more complicated settings. For example, the various recurrence and integrorecurrence equations derived in Griffiths and Tavaré (1994a, b, c, 1999) for the coalescent, Bahlo and Griffiths (2000) for the structured coalescent, Griffiths and Marjoram (1996) for the ancestral recombination graph and Slade (2000) for the ancestral selection graph are all effectively replaced by the standard IS formula (8).

Our starting-point was the development of improved proposal distributions in the IS framework, by considering the time reversal of underlying stochastic processes. We note that, although developed for the IS framework, our proposal distribution leads naturally to a class of Metropolis–Hastings schemes. Most naïvely it could be used as the proposal distribution for an independence sampler. More generally a simple Metropolis–Hastings scheme could be created by using our proposal distribution to propose an update to a random amount of the upper (i.e. furthest back in time) portion of the existing tree. It is an open question as to which such schemes perform best, and whether they outperform existing MCMC algorithms, but there may be grounds for optimism, at least in cases where our IS method does well. The

idea of exploiting appropriate time reversals may also be fruitful in other settings involving inference for high dimensional stochastic processes.

Useful insights for further improving the transition rates in IS proposal distributions might be gained by examining those of the optimal proposal distribution in small problems in which the latter can be calculated using good estimates of the conditional distribution π in equations (15). Since the likelihood in question is the solution of a system of linear equations, it may also be fruitful to pursue other (non-stochastic) numerical methods.

Our examples illustrate some of the relative strengths and weaknesses of MCMC and IS methods in these applications, which we believe may be of general interest. We summarize our experience in the remainder of this section. We found that IS methods were generally easier to code, and they performed well in situations (such as the infinite sites model) where the space of trees consistent with the data is reasonably tightly constrained, both in the sense of having lower dimension and in the sense of making movement around tree space for a generic MCMC method more difficult. In the one larger and less constrained example that we considered here an MCMC scheme appeared to have an advantage, and we conjecture that this might be typical. Although the amount of human effort that is required to design an efficient IS is often (as here) substantial, relatively simple MCMC update proposal schemes may give a reasonable performance for a wide range of problems. However, the issues involved in the choice of missing data and proposal distributions for MCMC methods are not well understood in this context and remain an important open problem.

MCMC methods also provide a very natural framework for Bayesian inference. Although this is theoretically straightforward using IS (posterior distributions can be found by using the likelihood estimates for example), practical problems will arise as the dimension of the parameter space increases. Further, we believe that MCMC methods II (see Section 3) which move around the parameter space (i.e. that allow θ to vary in the cases that we consider here) will tend to mix better over tree space and are an efficient method of performing either Bayesian or likelihood inference for the parameters, concentrating computational effort on regions with reasonable support from the data (this advantage becoming more marked when the dimension of the parameter space is higher). MCMC methods which fix θ at a 'driving value' θ_0, and use IS to estimate the relative likelihood, appear to make things unnecessarily difficult for themselves (see Stephens (1999)). Since the IS function which these MCMC methods use is most efficient (in fact optimal) at θ_0, and tends to become less efficient for θ away from θ_0, the distribution of the importance weights tends to be more skewed for θ away from the θ_0. As a result such methods will tend to underestimate the relative likelihood away from θ_0, leading to an estimated curve which is artificially peaked about this driving value (which we believe helps to explain the overly peaked curves in Fig. 2(c) for example). In principle IS methods based on a driving value of θ will tend to share this undesirable property, as designing a single IS function Q_{θ_0} which is universally efficient for all θ may be extremely challenging. Although this did not appear to cause major problems for our method in the examples considered here, we note that methods which combine the results of more than one IS function, such as bridge sampling (see for example Gelman and Meng (1998)) would produce more reliable results.

For both IS and MCMC methods there is the tricky question of how long the algorithms should be run for. Although generalizations are difficult to make, IS appears to have a slight advantage here: although highly skewed distributions for the weights may cause problems, detecting this pitfall once we are aware of it appears to be an easier problem (in most contexts) than monitoring the convergence of a Markov chain in such a high dimensional space. (It is, however, not difficult to contrive situations where the distribution of the weights is such that

it would be impossible to detect any problems just by looking at the observed weights.) Where more than one MCMC and/or IS scheme is available, running the different schemes until they give similar results provides a powerful check of both convergence and coding fidelity. Some other sensible procedures for deciding how long to run the IS algorithm are discussed in Section 6.4 below.

6.2. Extensions

Although here we have concentrated on estimating the likelihood for θ, an extension of these methods to questions of ancestral inference is straightforward. Indeed, it follows from standard IS theory that the distribution with atoms of size $w_i/\Sigma_j w_j$ on histories $\mathcal{H}^{(i)}$ $(i = 1, \ldots, M)$ is an approximation to the conditional distribution of $\mathcal{H}$ given A_n. For the simple demographic models considered here, the conditional distribution of the full typed ancestry $\mathcal{A}$ of the sample given $\mathcal{H}$ is easy to find, and it is straightforward to perform inference for all aspects of the typed ancestry relating the sample, including the ages of particular mutations in the ancestry, and the time to the MRCA (see Stephens (2000) for more details).

Although we have considered here only constant-sized randomly mating populations, the Griffiths–Tavaré method and the MCMC method of Kuhner, Yamato and Felsenstein have been extended to more complex demographic scenarios, involving for example varying population size (Griffiths and Tavaré, 1994c; Kuhner *et al.*, 1998) and structured populations (Bahlo and Griffiths, 2000; Beerli and Felsenstein, 1999). An obvious extension of our work would be to devise suitable IS functions for these contexts. Indeed, features of the structured population case suggest that it should be possible to find much more efficient proposal distributions than those so far considered. Our work, particularly theorem 1 and propositions 1 and 2, provides a starting-point for this search. In the case of a varying population size, the simple Markov structure of the underlying processes is lost, and it seems necessary to include time information in the missing data (as in Griffiths and Tavaré (1994c)).

Similarly, the Griffiths–Tavaré and Kuhner–Yamato–Felsenstein methods have been extended to more general genetic scenarios involving for example recombination (Griffiths and Marjoram, 1996; Kuhner *et al.*, 1999) and selection (Slade, 2000). An extension of our ideas to the recombination context by Fearnhead (2000) has resulted in substantial, and practically important, improvements in accuracy for a range of examples. Similarly, an extension of our ideas to selection results in a huge improvement in efficiency (details will be published elsewhere).

Most realistic sequence data typically consist of sequences which are thousands of sites long, with only a few of these sites varying in type among different chromosomes in the sample. Unless the infinite sites model for mutation can be assumed, a naïve application of our method to such data would be extremely inefficient. Most of the information about the history $\mathcal{H}$ is in the bases which vary between chromosomes, and it would be much more efficient to define the IS distribution on the basis of those positions which are varying. The effect of the non-varying sites could then be taken into account by the factor $\pi_\theta(A_n|\mathcal{H}^{(i)})$ in estimator (9), which could be calculated by the peeling algorithm (to do this it would be necessary to apply IS to the full typed ancestry $\mathcal{A}$ of the sample). A similar approach might make the MCMC method of Wilson and Balding (1998) applicable to sequence data.

One important type of data that were not treated above arises from a common modern two-stage experimental procedure. In the first stage a subset of the chromosomes in the sample (the 'panel') is sequenced completely. In the second stage the remaining chromosomes are assayed only at selected sites, often called single-nucleotide polymorphisms (SNPs), which

were observed to vary in type among the panel. This ascertainment process complicates the analysis. (In some ways this is the simplest version of the ascertainment problem for SNP data.) Dealing with the particular ascertainment effect just described is straightforward in IS and MCMC schemes, though the details will depend on the choice of missing data $\mathcal{T}$ and the mutation mechanism that is assumed. For the IS schemes that we have considered, assuming the infinite sites mutation model, the ascertainment effect can be accommodated by labelling every lineage which leads to any chromosome in the panel as a *panel lineage* and adapting both $P_\theta(\mathcal{H})$ and $Q_\theta(\mathcal{H})$ so that mutations can occur only on panel lineages. Modifying an MCMC scheme typically involves simply ensuring that the evaluation of $P_\theta(A_n, \mathcal{T})$ takes appropriate account of the ascertainment process. For example, if $\mathcal{T}$ is the genealogical tree $\mathcal{G}$, this probability can be evaluated by peeling each site separately, and for each site including only those chromosomes assayed at that site.

6.3. Bells and whistles

Our search for an efficient IS function was based exclusively on our knowledge of the underlying stochastic processes. In fact, we tried a range of other possible IS schemes which arose from different approximations to the conditional probabilities $\pi(\cdot|A_n)$. Many of them were very much more computationally intensive than the scheme that we presented here (which benefits considerably from the computationally convenient properties discussed in remark 2) and none of them produced a consistent improvement in efficiency, even when efficiency was measured *per iteration* and took no account of the amount of computation required. We suggest that further efficiency improvements might be most easily made by employing general computational tricks which have been successful in other contexts. In particular the IS scheme that we described, which recreates the tree in a Markov way from the sample back to the MRCA, may be viewed as sequential imputation (Kong *et al.*, 1994; Irwin *et al.*, 1994) of the ancestral states. It may then be fruitful to apply the rejection control ideas of Liu *et al.* (1999), in which unpromising trees would be discarded before reaching the MRCA, with appropriate modifications of the weights of the undiscarded trees. (This is a more sophisticated version of the strategy of discarding trees with too many mutations which was used by Griffiths and Tavaré (1994a) and Nielsen (1997).)

6.4. Diagnostics

We return to the problem of deciding when the IS algorithm has been run for sufficiently long. Some sensible procedures include monitoring (graphically for example) the mean and variance of the importance weights, the effective sample size (Kong *et al.*, 1994) and the relative size of the maximum weight. However, all these methods will suffer if the sample variance of the importance weights substantially underestimates the variance of the underlying distribution, and it would be useful to have some way of correcting for this. Although analytical results may sometimes be available, this is often not the case. For example, we could not prove finiteness of the variance of our weights, except in the special case of the infinite sites model where the number of possible histories is finite. (In contrast the naïve estimator (5) is guaranteed to have a finite (though usually huge) variance, as the weights are bounded. We could use this fact as in Hesterberg (1995) to modify our IS function Q_θ^{SD} to guarantee finite variance of the weights for $L(\theta)$, but it seems unlikely to be a fruitful way to proceed in this case.) A possible procedure, which appears promising on initial investigation, is to model the distribution of the weights by using distributions developed in extreme value theory for highly skewed distributions (this idea has been suggested independently by Shephard

(2000)). In particular we propose fitting a generalized Pareto distribution (see Davison and Smith (1990) for example) to the weights above some threshold, and using a parametric bootstrap to estimate confidence intervals for the mean of the weights (i.e. the estimate of the likelihood). Although this procedure may be inexact, it should better represent the uncertainty in the estimate than any current method that we are aware of.

6.5. Future challenges

Full likelihood-based inference for modern population genetics data presents computational and statistical challenges. In this paper we have deliberately focused on a one-dimensional inference problem under the simplest evolutionary model and compared methods on 'small' data sets. Although the field is still at an early stage, inference in these contexts is already important in practice, and methods described in the paper are becoming routinely used by biologists for much larger data sets, and in some more complex settings. The imminent completion of many of the current genome projects will be followed by an explosion of data documenting molecular variation in natural populations. Indeed, many real data sets are at or beyond the computational limits of current algorithms. There is thus an urgent need for the continuing development of more efficient inference methods, for their extension to more complex genetic and demographic scenarios and for the incorporation of the kinds of ascertainment effects which are rife in modern experimental data.

In parallel with the development of full likelihood-based methods in more complicated settings, we should also look for practicable inference procedures which, although not fully efficient, do not sacrifice much of the information in molecular genetics data. The availability in simple settings of fully efficient methods provides a useful yardstick for comparison. A better understanding of both small sample and asymptotic properties of the likelihood for genetics models would also be valuable.

Acknowledgements

We would like to thank Robert Griffiths, Augustine Kong and Simon Tavaré for stimulating discussions on the material in this paper, David Balding and Ian Wilson for advice on the use of `micsat` and Neil Shephard, Joe Felsenstein and the referees for helpful comments on an earlier version. The first author was supported by a grant from the University of Oxford and a Wellcome Trust Research Fellowship (reference 057416); the second author received support from Engineering and Physical Sciences Research Council grant GR/M14197 and Biotechnology and Biological Sciences Research Council grant 43/MMI09788. Part of the paper was written while the authors were at the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK.

Appendix A: Calculation of $\hat{\pi}(\beta|\boldsymbol{A}_n)$ for sequence and multilocus models

Consider a model with l completely linked loci which mutate independently. If each locus has k possible alleles, then the number of possible types is k^l. This model includes models for sequence data (such as the model described in Section 5.2) as a special case, with l being the length of the sequence, and each site in the sequence being thought of as a separate 'locus'. The large number of types in such models makes a naïve application of formulae (18) and (19) computationally daunting. However, the simple structure of the mutation mechanism allows us to approximate $\hat{\pi}(\beta|A_n)$ efficiently for any given type $\beta = (\beta_1, \beta_2, \ldots, \beta_l)$, as we now describe. We assume for simplicity that each locus mutates independently at rate $\theta/2$ (so the total rate is $l\theta/2$), according to a $k \times k$ transition matrix P. The methods that we

describe are easily extended to the case where each locus has a different set of possible alleles, mutation rate and transition matrix.

According to definition 1, a draw from $\hat{\pi}(\cdot|A_n)$ for this model may be made by choosing a chromosome from A_n uniformly at random, and then applying m mutations to this chromosome (each of which involves choosing a locus uniformly and changing the type at that locus according to P), where m is geometrically distributed with parameter $l\theta/(n+l\theta)$. It follows from elementary properties of Poisson processes that this is equivalent to drawing a time t from an exponential distribution with rate parameter 1, and then applying m_i mutations to locus i ($i = 1, \ldots, l$), where the m_i are independent and Poisson distributed with mean $\theta t/n$, and the mutations at each locus are governed by transition matrix P. Thus, writing types as $\alpha = (\alpha_1, \ldots, \alpha_l)$ and $\beta = (\beta_1, \ldots, \beta_l)$ we have

$$\hat{\pi}(\beta|A_n) = \sum_{\alpha \in A_n} \frac{n_\alpha}{n} \int \exp(-t) F^{(\theta,t,n)}_{\alpha_1\beta_1} \ldots F^{(\theta,t,n)}_{\alpha_l\beta_l} \, \mathrm{d}t \tag{31}$$

where

$$F^{(\theta,t,n)}_{\alpha_i\beta_i} = \sum_{m=0}^{\infty} \frac{(\theta t/n)^m}{m!} \exp\left(-\frac{\theta t}{n}\right) (P^m)_{\alpha_i\beta_i}. \tag{32}$$

The integral in equation (31) may be approximated by using Gaussian quadrature (see for example Evans (1993)):

$$\hat{\pi}(\beta|A_n) = \sum_{\alpha \in A_n} \sum_{i=1}^{s} \frac{n_\alpha}{n} w_i F^{(\theta,t_i,n)}_{\alpha_1\beta_1} \ldots F^{(\theta,t_i,n)}_{\alpha_l\beta_l} \tag{33}$$

where $t_1, \ldots, t_s$ are the quadrature points, and $w_1, \ldots, w_s$ are the corresponding quadrature weights. The matrices $F^{(\theta,t_i,n)}$ given by equation (32) may each be approximated by a finite sum with a large number of terms (these matrices need only be found once for any particular problem). We used $s = 4$ quadrature points. Although in some cases the approximation to $\hat{\pi}(\cdot|\cdot)$ obtained through this procedure is rather rough, we note that in any case the IS function defined by this approximation to $\hat{\pi}(\cdot|\cdot)$ is a valid IS function in its own right, and so leads to an estimator (9) which is consistent.

References

Bahlo, M. and Griffiths, R. C. (2000) Inference from gene trees in a subdivided population. *Theor. Popln Biol.*, **57**, 79–95.

Beaumont, M. (1999) Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013–2029.

Beerli, P. and Felsenstein, J. (1999) Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763–773.

Chen, M. H. (1994) Importance-weighted marginal Bayesian posterior density-estimation. *J. Am. Statist. Ass.*, **89**, 818–824.

Cooper, G., Amos, W., Hoffman, D. and Rubinsztein, D. C. (1996) Network analysis of human Y microsatellite haplotypes. *Hum. Molec. Genet.*, **5**, 1759–1766.

Davison, A. C. and Smith, R. L. (1990) Models for exceedances over high thresholds (with discussion). *J. R. Statist. Soc.* B, **52**, 393–442.

Donnelly, P. (1986) Dual processes in population genetics. *Lect. Notes Math.*, **1212**, 94–105.

Donnelly, P. and Kurtz, T. G. (1996a) A countable representation of the Fleming–Viot measure-valued diffusion. *Ann. Probab.*, **24**, 698–742.

——— (1996b) The asymptotic behaviour of an urn model arising in population genetics. *Stochast. Process. Applic.*, **64**, 1–16.

——— (1999) Particle representations for measure-valued population models. *Ann. Probab.*, **27**, 166–205.

Donnelly, P. and Tavaré, S. (1995) Coalescents and genealogical structure under neutrality. *A. Rev. Genet.*, **29**, 401–421.

Edwards, A. W. F. (1970) Estimation of the branch points of a branching diffusion process (with discussion). *J. R. Statist. Soc.* B, **32**, 155–174.

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

Ethier, S. N. and Kurtz, T. (1993) Fleming-viot processes in population genetics. *SIAM J. Control Optimsn*, **31**, 345–386.

Evans, G. (1993) *Practical Numerical Integration*. New York: Wiley.

Ewens, W. J. (1979) *Mathematical Population Genetics*. Berlin: Springer.

Fearnhead, P. (2000) Estimating recombination rates by importance sampling. To be published.
Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Molec. Evoln*, **17**, 368–376.
Felsenstein, J., Kuhner, M. K., Yamato, J. and Beerli, P. (1999) Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. *IMS Lect. Notes Monogr. Ser.*, **33**, 163–185.
Gelman, A. and Meng, X. L. (1998) Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, **13**, 163–185.
Geyer, C. J. (1996) Estimation and optimization of functions. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall.
Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc.* B, **54**, 657–699.
Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.
Griffiths, R. C. and Marjoram, P. (1996) Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, **3**, 479–502.
Griffiths, R. C. and Tavaré, S. (1994a) Simulating probability distributions in the coalescent. *Theor. Popln Biol.*, **46**, 131–159.
———(1994b) Ancestral inference in population genetics. *Statist. Sci.*, **9**, 307–319.
———(1994c) Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. Lond.* B, **344**, 403–410.
———(1995) Unrooted tree probabilities in the infinitely-many-sites model. *Math. Biosci.*, **127**, 77–98.
———(1997) Computational methods for the coalescent. *IMA Vol. Math. Applic.*, **87**, 165–182.
———(1999) The ages of mutations in gene trees. *Ann. Appl. Probab.*, **9**, 567–590.
Hesterberg, T. C. (1995) Weighted average importance sampling and defensive mixture distributions. *Technometrics*, **37**, 185–194.
Hoppe, F. M. (1984) Polya-like urns and the Ewens sampling formula. *J. Math. Biol.*, **20**, 91–94.
Irwin, M., Cox, N. and Kong, A. (1994) Sequential imputation for multilocus linkage analysis. *Proc. Natn Acad. Sci. USA*, **91**, 11684–11688.
Kingman, J. F. C. (1982a) On the genealogy of large populations. *J. Appl. Probab.* A, **19**, 27–43.
———(1982b) Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics* (eds G. Koch and F. Spizzichino), pp. 97–112. Amsterdam: North-Holland.
———(1982c) The coalescent. *Stochast. Process. Applic.*, **13**, 235–248.
Kong, A., Liu, J. S. and Wong, W. H. (1994) Sequential imputation and Bayesian missing data problems. *J. Am. Statist. Ass.*, **89**, 278–288.
Kuhner, M. K., Yamato, J. and Felsenstein, J. (1995) Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics*, **140**, 1421–1430.
———(1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429–434.
———(1999) Recombine. *Computer Program.* University of Washington, Seattle. (Available from `http://www.evolution.genetics.washington/lamarc.html`.)
Liu, J. S., Chen, R. and Wong, W. H. (1999) Rejection control and sequential importance sampling. *J. Am. Statist. Ass.*, **93**, 1022–1031.
Moran, P. A. P. (1975) Wandering distributions and the electrophoretic profile. *Theor. Popln Biol.*, **8**, 318–330.
Nielsen, R. (1997) A likelihood approach to population samples of microsatellite alleles. *Genetics*, **146**, 711–716.
Raftery, A. E. (1996) Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall.
Ripley, B. D. (1987) *Stochastic Simulation*. New York: Wiley.
Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Statist. Soc.* B, **53**, 683–690.
Shephard, N. (2000) Discussion on 'Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives' (by J. Durbin and S. J. Koopman). *J. R. Statist. Soc.*, B, **62**, 30–32.
Slade, P. (2000) Simulation of selected genealogies. *Theor. Popln Biol.*, to be published.
Stephens, M. (1999) Problems with computational methods in population genetics. *Bull. 52nd Sess. Int. Statist. Inst.*, book 1, 273–276.
———(2000) Times on trees and the age of an allele. *Theor. Popln Biol.*, **57**, 109–119.
Tanner, M. (1993) *Tools for Statistical Inference*, 2nd edn. New York: Springer.
Venables, W. N. and Ripley, B. D. (1997) *Modern Applied Statistics with S-Plus*, 2nd edn. New York: Springer.
Whittle, P. (1970) Discussion on 'Estimation of the branch points of a branching diffusion process' (by A. W. F. Edwards). *J. R. Statist. Soc.* B, **32**, 167–169.
Wilson, I. J. and Balding, D. J. (1998) Genealogical inference from microsatellite data. *Genetics*, **150**, 499–510.
Wright, S. (1969) *Evolution and the Genetics of Populations*, vol. 2, *The Theory of Gene Frequencies*. Chicago: University of Chicago Press.

Discussion on the paper by Stephens and Donnelly

Ian Wilson (*University of Aberdeen*)

It is a great pleasure for me to propose the vote of thanks for this paper. For most of the last century, despite — or perhaps because of — the lack of data there was a huge production of mathematical models in population genetics. Now that we have data, can we develop powerful inferential techniques? Recent work is answering this question.

Inference for molecular genetics is primarily inference about evolution, a historical process. This history is written into patterns in deoxyribonucleic acid (DNA) at all levels, from different genes within a single genome to comparisons between genes for different species. Patterns are created by the duplication of genes and subsequent modification on the different lines of descent and can generally be described by tree structures. When we have recombination between different sequences the patterns can be described as series of trees linked along DNA sequences (Wuif and Hein, 1999). Inference from sequence variation should involve modelling what happened in the past. Methodologies that do not take account of this history — such as those based on summary statistics — risk confounding information about the evolutionary process with the history. Furthermore, aspects of the history, such as the time of the most recent common ancestor for humans, are sometimes of interest in themselves.

The key advance in the paper presented is the development of new importance sampling (IS) methodologies for the analysis of DNA sequence data that explicitly model the underlying mutational history of the sample. These methods are improvements to those that are currently available for many types of sequence variation and are extendable to refinements of the standard coalescent model such as those for varying population size and population subdivision. The general methodology may also be developed for other modelling frameworks, such as birth–death processes for phylogenetic inference (Rannala and Zhang, 1997) or the evolution of transposable elements (Brookfield, 1986).

The authors tested `micsat` (Wilson and Balding, 1998) against their method and obtain substantive agreement for the analysis of the samples from the Nigeria, Sardinia and East Anglia (NSE) data set, which is pleasing to me. This replication is important, as it is difficult to show that Markov chain Monte Carlo (MCMC) methods have converged and are mixing. Conversely the drawbacks of IS have been covered in the paper.

I would like to touch on two areas of concern to me. One area where I feel that statistics should try to take a more proactive role is in the design of surveys, particularly the design and analysis of surveys from many different human populations; how do we sample from the world population? Sampling individuals proportional to current population sizes may overrepresent ethnic groups whose populations have rapidly expanded over historical rather than biological timescales. One approach, taken with `BATWING`, 'Bayesian analysis of trees with internal node generation', available from

`http://www.maths.abdn.ac.uk/~ijw`,

is to model jointly a population 'supertree' — a tree of subpopulations — with the gene genealogy. This allows us to infer the relative size of different populations but may not reflect the complex history of human migration. Further developments of inferential techniques may allow more informed survey designs.

Another problem is the increasing volume of data. `micsat` and its successor `BATWING` have been successful at analysing variation in human Y-chromosomes, yet there is still the danger that the data sets expand quicker than computer speeds increase and our techniques advance. The NSE data set analysed here and in Wilson and Balding (1998) consisted of 60 Y-chromosomes with five microsatellite loci and a single unique event polymorphism (UEP) scored for each chromosome. A recent data set, that stretched `BATWING` to or beyond its limits, from the Centre for Genetic Anthropology at University College London consisted of 1700 Y-chromosomes with six microsatellite loci and seven UEP sites, from 10 populations. This is typical of the data that this centre is producing. Methods that can jointly analyse microsatellite and UEP data are at a real advantage with problems such as these as UEPs constrain the tree structure reducing the dimension of the tree space. The advantage of the IS methodology developed by Stephens and Donnelly when we have constrained data and the relatively better performance of MCMC methods for larger sample sizes with fewer constraints suggest that hybrid IS and MCMC methods may be a productive direction for research for these large problems.

To conclude, the authors have provided an important tool for the statistical toolbox which we shall need in the future. This should contain a diverse assortment of methods for accurate and efficient inference. Stephens and Donnelly are to be congratulated on producing a paper of much interest and substance that extends the possible directions of research in the analysis of the causes of genetic variation. I have great pleasure in proposing the vote of thanks.

D. A. Stephens (*Imperial College of Science, Technology and Medicine, London*)

I would like to concur with the proposer of the vote of thanks by saying that this paper, through the technical developments introduced in theorem 1 and their practical implementation, and the more general aspects of the inference problems that it contains, provides an excellent introduction to the area and thus represents a major contribution.

I begin with some minor requests for clarification. Using the notation of the paper, taking a fully Bayesian view, we have

$$p(\theta|A_n) = \int p(\theta, \mathcal{H}|A_n)\,\mathrm{d}\mathcal{H} = \int \frac{p(A_n|\theta, \mathcal{H})\,p(\mathcal{H}|\theta)p(\theta)}{p(A_n)}\,\mathrm{d}\mathcal{H} = \frac{\int p(A_n|\theta, \mathcal{H})\,p(\mathcal{H}|\theta)\,\mathrm{d}\mathcal{H}}{\int\int p(A_n|\theta, \mathcal{H})\,p(\mathcal{H}|\theta)\,\mathrm{d}\mathcal{H}\,p(\theta)\,\mathrm{d}\theta}p(\theta)$$

$$= \frac{L(\theta)\,p(\theta)}{\int L(\theta)\,p(\theta)\,\mathrm{d}\theta} = c\,L(\theta)\,p(\theta),$$

say, where $p(\theta)$ is the prior distribution for θ. Clearly, $p(\theta|A_n)$ is available from $L(\theta)$ by numerical integration, but obtaining $L(\theta)$ or c from $p(\theta, \mathcal{H}|A_n)$ is much more challenging. However, any converged Markov chain Monte Carlo (MCMC) scheme for $p(\theta, \mathcal{H}|A_n)$ has, theoretically, traversed the support of the posterior and, in particular, visited all $\mathcal{H}$ such that

$$p(A_n|\theta, \mathcal{H})\,p(\mathcal{H}|\theta)\,p(\theta) > 0.$$

Thus an approximation to $L(\theta)$ that utilizes the summation over (distinct) $\mathcal{H}$s visited by the Markov chain is available; such an approximation would be valid for all θ, and thus gives a complete summary of $L(\theta)$. Do the authors have any feeling for how good or poor such an approximation might be?

Secondly, I have similar doubts to the authors' concerning the use in general of a 'driving value', θ_0 say, in any of the importance sampling (IS) schemes described in the paper; for example, different values of θ encourage potentially radically different values of coalescence time m, thus compromising the accuracy of the approximation of P_θ by Q_θ. Do the computational 'efficiency' gains outweigh the loss in accuracy? Can the authors recommend a resolution to this problem? Is it valid to state that an MCMC scheme operating on the full posterior avoids this difficulty in any case?

Finally, in many statistical applications, inference based on *relative* rather than *absolute* likelihood is sufficient; in Gelman and Meng (1998), for example, emphasis is pointedly on the former. Can the authors describe more specifically the importance of evaluating $L(\theta)$ in absolute terms in the context of the genetic inference problem?

As the authors acknowledge in their discussion (page 630), a well-behaved MCMC scheme that facilitates inference via the full joint posterior (i.e. for $(\theta, \mathcal{H})$ jointly) is probably preferable to an IS scheme that holds θ fixed. However, theorem 1 may have a useful role to play in the construction of an MCMC scheme; it is possible to use Q_θ as a Metropolis–Hastings proposal density for updating $\mathcal{H}$ in its full conditional, or jointly with θ; the desirability of Q_θ as an IS proposal automatically recommends it as a Metropolis–Hastings proposal. It appears that such an approach may compete with the Wilson and Balding (1998) algorithm; have the authors any comments?

Throughout the analysis described, inference is carried out conditionally on the transition matrix P. In general, this matrix comprises fundamental evolutionary parameters that are unknown (an exception being the case of microsatellite data). In a system with a large number of alleles, the specification of P may be difficult, and any uncertainty in the specification of P should perhaps be recognized. Are inferences about θ necessarily robust to the specification of P? In a full Bayesian analysis, the uncertainty about P can be incorporated by including it as another parameter, although it may not be possible to learn about both θ and P from a single data sample. The incorporation of the extra parameter may not introduce many extra difficulties, as by construction Q_θ is 'optimal' for any P. To construct a prior distribution for P (or for specification of P itself), it may be possible to utilize specification ideas about point mutations developed for biological sequence analysis (see, for example, the discussion of PAM and BLOSUM matrices in Durbin *et al.* (1998)).

Finally, much recent attention has centred on other tree-based inference problems in genetics, specifically the probabilistic construction of phylogenetic trees (e.g. Mau *et al.* (1999), Newton *et al.* (1999) and Li *et al.* (2000)) where there is an emphasis on constructing MCMC schemes to solve a

similar missing data problem, albeit where the likelihood and inference goals are somewhat different from those in this paper. Are there any useful connections to techniques used in this parallel inference problem? Can any advantages be gained by developing a different 'co-ordinate system' (Diaconis and Holmes, 1998) for the missing data tree? Also, some standard bioinformatics texts (e.g. Baldi and Brunak (1998)) describe solutions of phylogenetic inference problems by using standard machine learning or Bayesian network algorithms; are such approaches feasible for the ancestral inference problem that is described in this paper?

The authors are to be congratulated on the methodological developments introduced, their clear elucidation of more general aspects of inference in such genetics problems and their attempt to address the practical issues involved. I feel that the algorithm proposed will play an important role in the simulation-based approach to inference in both IS and MCMC contexts. It gives me great pleasure to second the vote of thanks to the authors.

The vote of thanks was passed by acclamation.

Rosalind M. Harding (*University of Oxford*)

There are two points on which I would like to congratulate the authors; first, for their synthesis. This synthesis makes it easier to appraise both the strengths and the limitations of these new methods, and to see a little way ahead into the manner of their improvement. Secondly, I congratulate them for using this analysis of the role of importance sampling to find a way of improving the accuracy and efficiency, but especially the efficiency, of these methods. In population genetics there is usually little reason for confidence that an estimate is correct even to within an order of magnitude, but reaching it faster is definitely progress. This is significant because, in applications to real rather than simulated data, hundreds of computing hours are needed to explore the likelihood surfaces of models of interest.

I also have two questions. This paper suggests to me that there is *still* a long way to go before well-designed tools will be available for working on the particular problems that geneticists currently choose to study. The authors develop a better method for estimating θ but estimating θ is a means to an end. It is not the end in itself. Among the reasons for estimating θ is to ask questions about the population size N. Has the population in the past fluctuated in size and been squeezed through bottle-necks? Is the population currently in an expansion phase, and, if so, how far back does this phase date? If N has been small in the past, then how small? When it was small, did it become more structured, or did it shrink into a single founding group? The right way to try to answer these questions is with likelihood-based inference applied to population genetic data. So, how long is it to be before the new and improved methods that you are developing will be available for application to the sorts of problem that motivate data collection? My other question concerns the kind of problem that makes the estimation of θ 'difficult' for some algorithms, such as the example in Table 1. Is the difficulty caused by data that are unlikely given the generating value of θ? Or, are problems difficult for a more complicated set of reasons?

Bob Griffiths (*University of Oxford*)

Briefly here is some chronological history about the Griffiths–Tavaré method and the new importance sampling method of the authors. In the original Griffiths–Tavaré method, we started from equations for the probability $p(\mathbf{n})$ of a sample configuration of $\mathbf{n} = (n_1, \ldots, n_d)$ in a model with d possible types. d could be very large, perhaps 4^k for sequences of length k. The set of equations obtained by considering the first event back in the coalescent process is

$$p(\mathbf{n}) = \frac{\theta}{n+\theta-1}\left\{\sum_{\alpha=1}^{d}\sum_{\beta=1}^{d}\frac{n_\beta + 1 - \delta_{\alpha\beta}}{n}p_{\beta\alpha}\,p(\mathbf{n}+\mathbf{e}_\beta-\mathbf{e}_\alpha)\right\} + \frac{n-1}{n+\theta-1}\sum_{\{\alpha:n_\alpha>0\}}\frac{n_\alpha-1}{n-1}p(\mathbf{n}-\mathbf{e}_\alpha). \qquad (34)$$

In Stephens–Donnelly notation let H_j be the configuration of types at step j of the history, $j = 0, -1, -2 \ldots$; then equation (34) has the form

$$p(H_j) = \sum P(H_j|H'_{j-1})\,p(H'_{j-1}),$$

with summation over possible configurations H'_{j-1}. This corresponds to equation (34) with $H_j = \mathbf{n}$ and $\{H'_{j-1}\} = \{\{\mathbf{n}+\mathbf{e}_\beta-\mathbf{e}_\alpha\} \cup \{\mathbf{n}-\mathbf{e}_\alpha\}\}$. $p(H_j)$ and $\{p(H'_{j-1}\}$ are unknown, and $p(H_j|H'_{j-1})$ are the known coefficients in equation (34).

In the Griffiths–Tavaré method, view the problem of finding $\{p(\mathbf{n})\}$ as solving the system of linear

equations (34). Importance sampling techniques were used to obtain a solution of the equations. Rescale $P(H_j|H'_{j-1})$ and interpret as a probability distribution

$$q(H'_{j-1}|H_j) = p(H_j|H'_{j-1})/f(H_j),$$

where $f(H_j) = \Sigma\, p(H_j|H'_{j-1})$. Then

$$p(H_j) = f(H_j) \sum q(H'_{j-1}|H_j)\, p(H'_{j-1}),$$

leading to an importance sampling representation

$$p(H_0) = E_q\{f(H_0)f(H_{-1}). \ . \ . \ f(H_{-m})\},$$

with absorption at $-m$, where there first is a single ancestor of the sample. The proposal distribution for importance sampling is $q(\cdot|\cdot)$.

In the Stephens–Donnelly method,

$$p(H_j) = \sum \frac{p(H_j|H'_{j-1})}{\hat{p}(H'_{j-1}|H_j)} \hat{p}(H'_{j-1}|H_j)\, p(H'_{j-1}).$$

The importance sampling representation is

$$\begin{aligned} p(H_0) &= E_{\hat{p}}\left\{ \frac{p(H_0|H_{-1})}{\hat{p}(H_{-1}|H_0)} . \ . \ . \frac{p(H_{-m+1}|H_{-m})}{\hat{p}(H_{-m}|H_{-m+1})} p(H_{-m}) \right\} \\ &= E_{\hat{p}}\left\{ \frac{p(H_0|H_{-1}). \ . \ . p(H_{-m+1}|H_{-m})\, p(H_{-m})}{\hat{p}(H_{-m}|H_{-m+1}). \ . \ . \hat{p}(H_{-1}|H_0)} \right\} \\ &= E_{\hat{p}}\left\{ \frac{p(\mathcal{H}_{\rightarrow})}{\hat{p}(\mathcal{H}_{\leftarrow})} \right\}. \end{aligned} \tag{35}$$

The proposal distribution for importance sampling is $\hat{p}(\cdot|\cdot)$ constructed from the authors' $\{\hat{\pi}(\cdot|\cdot)\}$. In the numerator of equation (35) the history is evaluated in the direction from the ancestor type to the current time, and in the denominator in the reverse direction.

A point is that all the information about $\{p(\mathbf{n})\}$ is contained in equations (34), even though explicit knowledge of the coalescent process is used to construct $\{\hat{p}(\cdot|\cdot)\}$.

The Stephens–Donnelly approach suggests that there may be a nice probabilistic way to solve large systems of equations by an analogous method, possibly combined with existing numerical methods.

Problem

Suppose that we have a set of N sparse linear equations in $\mathbf{a} = (\alpha_1, \ . \ . \ ., \alpha_N)^{\mathrm{T}}$ with N a large dimension with a non-negative coefficient matrix B,

$$\mathbf{a} = B\mathbf{a}, \tag{36}$$

for which we know a solution exists, and we know precisely a set $\{\alpha_j: j \in \mathcal{A} \subset \{1, \ . \ . \ ., N\}\}$.

Find an efficient stochastic importance sampling representation related to the Stephens–Donnelly approach that allows a simulated solution of $\mathbf{a}$. Equations (36) are related to those obtained for the probability of absorption in a Markov chain with absorbing states $\mathcal{A}$, but the solution will be much more sophisticated than a simple simulation to absorption.

Paul Joyce (*University of Idaho, Moscow*)

This paper makes an important contribution to the analysis of complex genetics data. Central to their importance sampling method is the construction of a 'good' proposal distribution. For this, they make two key observations.

(a) The structure of the history of the process conditional on the data is characterized by $\pi(\cdot|A_n)$, the probability distribution of an $(n+1)$th sampled chromosome given the types A_n of the first n sampled chromosomes. Under parent-independent mutation the distribution is well known.
(b) Constructing a probability distribution $\hat{\pi}(\cdot|A_n)$ that is 'close' to $\pi(\cdot|A_n)$ forms the basis of a proposal distribution for the history of the data given the sample. Simulated histories, according to the proposal distribution, form the bases for inference using importance sampling.

The authors are to be congratulated for their accomplishment. The insightful observation of Felsenstein, who noted that the Griffiths–Tavaré likelihood method is in fact a use of importance sampling, is the starting-point for this clever paper.

The example given after the proof of theorem 1 is very illuminating, demonstrating that the Griffiths–Tavaré method will (in this example) generate many unlikely histories that contribute little to the likelihood of the data. It certainly helped me to gain insight into their methods.

Monte Carlo methods are more akin to stochastic numerical analysis than to statistics. The missing data problem forces the investigator to tackle a nasty integral or summation over a space of large dimension. More attention is paid to the integration problem than to the inferences one draws from the data. Some statistical issues of concern are as follows.

Reporting errors in the estimates
Since one cannot rely on standard asymptotic theory, it is unclear whether or not the curvature of the likelihood surface is a valid estimate of the standard error. What is the standard error of the maximum likelihood estimate?

Robustness
In complex genetics data, how well do inferences hold up under mild violations of the model assumptions?

Model selection
Can one develop criteria for selecting between coalescent models? As we add complexity to the model, such as selection, recombination, populations structure and expansion, we begin to face the problem, which is common to many statistical scenarios, between adding parameters and losing degrees of freedom.

I am not so much concerned that these items are addressed in the present paper as I am that these issues should come to the forefront as we begin to overtake the technical issues of generating the likelihood.

Again, I would like to thank the authors for a very insightful and useful paper.

A. W. F. Edwards (*University of Cambridge*)
I am glad to note the authors' reference to Professor Whittle's contribution to the discussion of my paper which was read to the Royal Statistical Society in 1969 (Edwards, 1970). A combination of the theory of the coalescent and the invention of powerful methods for the computation of likelihoods has since completely revolutionized the field, but I should like to think that in the early years we were on the right track.

In particular, we had had the good fortune to have learnt statistics at the feet of R. A. Fisher himself, twice a Professor of Genetics but never a Professor of Statistics. So we knew about likelihood, but what we did not at first appreciate was that statisticians on the whole did not. The text-books of the day only ever mentioned it in connection with the method of maximum likelihood. In my reply to Professor Whittle's comment I expressed the hope that I would have the opportunity to do some calculations along the lines he suggested, but in the event I answered instead a compelling vocation to write a book, *Likelihood* (Edwards, 1972, 1992), intended to persuade people to take the concept more seriously.

In their Section 5, 'Applications', Stephens and Donnelly imply that the main reason for using log-likelihoods rather than likelihoods is the classical, i.e. repeated sampling, perspective. They go on to claim that 'From a Bayesian viewpoint, the likelihood is of more natural interest than is the log-likelihood'. I differ from both these opinions. The second reveals no more than a preference for multiplication over addition, but the first is more serious, for it ignores the role of log-likelihood as itself a measure of *support*, requiring neither repeated sampling nor Bayesian justification. That was the main message of my book. The additive nature of support is natural because it corresponds to the addition of information, so although I applaud the authors' decision to 'focus here on likelihood and relative likelihood estimation', I should like to see all the authors' diagrams on a logarithmic scale.

Two things became clear in the early years of inference studies on trees and genealogies: first, the need to work with likelihoods, and secondly that whereas we could see how to solve in principle all our probability problems, by complex calculation on trees and genealogies or even by forward simulation, we could equally clearly see that finding likelihoods was an inverse problem of greater theoretical difficulty, and that it was apparently intractable by simulation. Forward simulation conditional on hitting the data precisely, which is what computing the likelihood seemed to require, was orders of magnitude too time consuming. Professor Whittle's 'time machine' suggestion was the first mathematical attempt to analyse how the system might be run backwards.

Paul Fearnhead (*University of Oxford*)

I should like to congratulate the authors on an interesting and stimulating paper. My comment is based on their assumption of no recombination. In particular, I should like to explain how, using similar ideas, importance sampling can be used to estimate the likelihood surface for models with recombination efficiently.

Modelling recombination is important not only because the assumption of no recombination is unrealistic for most genetic data but also because an estimation of the amount of recombination is very important. Population data enable recombination to be estimated over small scales, and knowledge of the small scale variation in recombination is important for understanding genetic effects involved with diseases.

For population genetic models with recombination, the genealogy of a sample can no longer be represented by a tree. Instead it is represented by a graph. This graph contains bifurcations (which in some sense represent recombination events) as well as coalescences. Thus, the history of any sample can now be represented by such a graph, and the mutations on it. Although the ancestral history is more complicated, it is still straightforward to calculate the probability of any history.

As a result, the calculation of the likelihood can still be viewed as a missing data problem, with the missing data being the history of the sample. Therefore, we can approximate the likelihood by using importance sampling, with a proposal density which simulates ancestral histories for our sample.

As in the case of no recombination, we can characterize the optimal proposal density by considering the time reversal of the underlying stochastic process (as in theorem 1). The optimal proposal density depends on one-dimensional sampling distributions. By approximating these sampling distributions (which is considerably more complicated in this setting, owing to the effects of recombination) we obtain an approximation to the optimal proposal density. We use this approximation as our proposal density in the importance sampling scheme.

The results for this importance sampling scheme are encouraging. There are two obvious comparisons. The first is with the importance sampling scheme of Griffiths and Marjoram (1996) (which is an extension of the Griffiths–Tavaré approach), and the new importance sampling scheme can be up to three orders of magnitude more efficient. It also appears more efficient than the Markov chain Monte Carlo scheme of Kuhner *et al.* (2000) (see Fearnhead (2000) for more details).

W. J. Ewens (*University of Pennsylvania, Philadelphia*)

I congratulate Dr Stephens and Professor Donnelly on their excellent paper. Several discussants have already raised some of the points which I had intended to make, so I shall restrict myself to three or four fairly independent comments.

So far as historical matters are concerned, this paper comes full circle in a pleasing way. In the 1970s the importance of θ as an evolutionary parameter was appreciated and the problem of its estimation from data already considered. However, the models used were very simple, often one or other of the parent–independent mutation models referred to by Stephens and Donnelly. For these models a likelihood was often explicitly available so the question of the estimation of the likelihood did not arise. However, the results obtained were sufficiently unexpected that Kingman was led to explain them through the introduction of the coalescent process. This concept has revolutionized population genetics theory and it is interesting to see that it is used in a central way by Stephens and Donnelly in their likelihood estimation procedure, thus answering questions that could not have been attacked in the 1970s.

My next point relates to the estimation of θ rather than of its likelihood function. It seems to be unavoidable that the variance of any reasonable estimator is of the order of $1/\log(n)$, where n is the sample size. Although an accurate estimation of the likelihood function is clearly an important first step in the estimation of θ, the value of an accurate estimation is diminished if an accurate estimation of θ remains elusive.

Finally, I remark that, although the focus in their paper is on the parameter θ, it is possible that the methods that they propose would be useful for other parameters. My own interest is in the genetics of diseases and here the parameter of interest is the linkage parameter between a purported disease locus and a marker locus. There is a natural coalescent process from the disease genes of affected individuals in a sample back to an originating disease mutation. However, we do not observe the genes at the disease loci but the genes at marker loci, so these undergo a 'quasi-coalescent' process associated in some way with that of the disease genes. Do Dr Stephens and Professor Donnelly see any way of using their importance sampling methods in conjunction with the coalescent process to assist in estimating linkage parameters?

Mark A. Beaumont (*University of Reading*)

One of the many points raised by this important paper is that the computational techniques involved in solving these problems are complex, and independent corroboration is useful. I present some results that are relevant to one of the applications that the authors discuss.

The Markov chain Monte Carlo (MCMC) approach described in Beaumont (1999) is similar to that of Wilson and Balding (1998) but draws samples from $\pi(\theta, \mathcal{H}|A_n)$. Using a uniform (improper) prior, relative likelihoods for θ can be obtained by standard density estimation on the sampled values. Stephens and Donnelly note that the results depend on the smoothing method used (Fig. 4) and suggest (Section 3.3) that more efficient approaches are available.

A feature of using an MCMC approach that samples $\mathcal{H}$ is that it is possible to use Rao–Blackwellization (Gelfand and Smith, 1990) to estimate the marginal posterior distribution of θ. It is straightforward to calculate

$$\pi(\theta|\mathcal{H}^i, A_n) = \frac{T^i}{2} \frac{(T^i\theta/2)^{m^i} \exp(-T^i\theta/2)}{m^i!},$$

where m^i and T^i are respectively the total number of mutations and the total branch length in the ith sampled genealogical history $\mathcal{H}^i$. We then estimate

$$\hat{\pi}(\theta|A_n) = \frac{1}{n} \sum^{n} \pi(\theta|\mathcal{H}^i, A_n).$$

The aim of this contribution is to obtain posterior densities for θ by using Rao–Blackwellization, and to compare the relative likelihoods with those obtained by Stephens and Donnelly (Figs 3 and 4) using the test data of Wilson and Balding (1998). In addition, likelihoods are estimated using rejection sampling as described by Beaumont (1999). The advantage of the latter approach is that sampling is independent with known error.

Five independent simulations of 10^7 iterations were run. Trial values of θ were proposed with probability 0.05 ($\mathcal{H}$ was updated otherwise). Every 500 iterations, values of $\mathcal{H}$ and θ were sampled. Each simulation took about 5 min on a 500 MHz Pentium computer.

The estimated posterior distributions are scaled to enclose the same volume over the range $\theta = 0$–20 as the estimates from rejection sampling. There is good concordance between all the methods (Fig. 8), although there is appreciable variability between independent runs of the MCMC simulation. Estimates of the density using the program `Locfit` showed a similar degree of variability between independent runs but there was substantial oversmoothing of the lower tail. A better fit would be obtained by modifying the estimation procedure, but the Rao–Blackwellization avoids an *ad hoc* treatment. In more complex models, the conditional distributions for the parameters will be more difficult to estimate and approaches such as that suggested by Chen (1994) will be required.

Mary K. Kuhner and Peter Beerli (*University of Washington, Seattle*)

In their Fig. 2 Stephens and Donnelly show our Markov chain Monte Carlo (MCMC) algorithm `Fluctuate` producing curves which vary greatly between runs and are narrower than their importance sampling (IS) curve. We have independently repeated these simulations and confirm that on these data our MCMC algorithm produces unstable results even in lengthy searches. We shall discuss a possible reason for this behaviour and suggest a way to improve MCMC performance.

Stephens and Donnelly mention, but do not emphasize, a fundamental difference between current MCMC and IS approaches. (Actually both sets of methods use IS, so the term 'IS' is somewhat misleading). In MCMC sampling, the missing data of the genealogy is represented as a topology with branch lengths. The size of the search space depends on the number of branches. Short sequences lead to many possible combinations of branch lengths and make a stable estimate difficult.

In contrast, IS represents the missing data as a topology with mutations. The size of the search space is determined by the number of mutations, which depends mainly on the number of sequence positions. For short sequences the program needs to assign only a few mutations.

The data set used in Fig. 2 contains 50 individuals and only 20 sites, making it well suited to the IS approach. As Stephens and Donnelly note, MCMC sampling and IS have distinct and complementary strengths. Additional data would stabilize the MCMC estimate at little cost in speed.

Stephens and Donnelly suggest that MCMC sampling with a fixed 'driving value' may produce a likelihood curve that is too narrow. Our simulations suggest that this does not happen with estimation

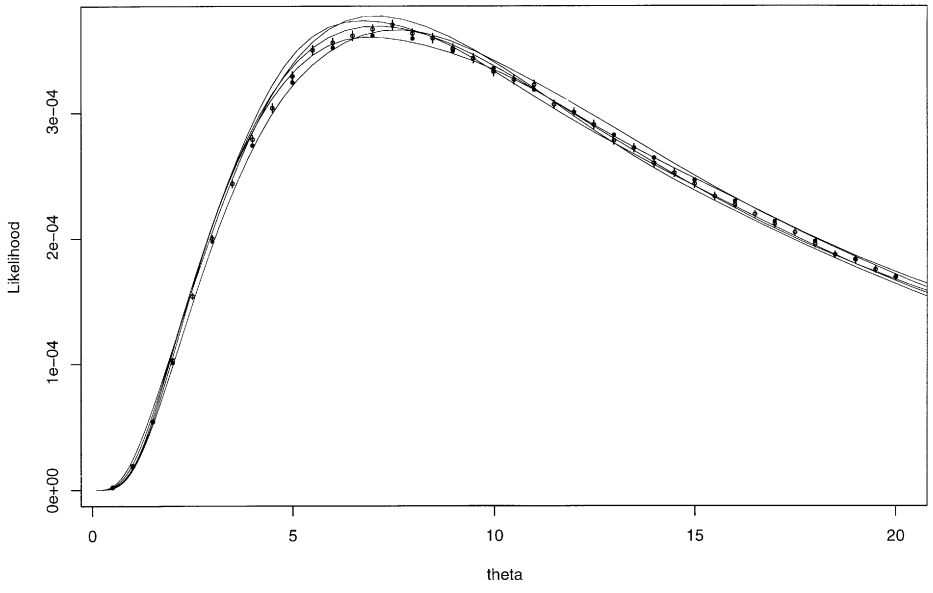

Fig. 8. Plot of the likelihood for different values of θ: ———, scaled posterior density estimates from the MCMC simulations; ●, likelihood estimates from the method of Stephens and Donnelly; ⌽, estimates from rejection sampling, with approximate 95% confidence intervals

of θ on larger data sets (Kuhner *et al*., 1995) but may be a problem for the co-estimation of θ and the growth rate (Kuhner *et al*., 1998). A possible solution is to make several runs with different driving values and to combine the resulting samples by using the method of Geyer (1991). Shown in Fig. 9 is an MCMC curve (generated by the `MIGRATE` program of Beerli and Felsenstein (1999), without migration) for the data of Fig. 2 combining 10 independent runs at different driving values of θ. The curve is now similar to the result of Stephens and Donnelly.

The following contributions were received in writing after the meeting.

Stephen Brooks (*University of Surrey, Guildford*) **and Andrew Gelman** (*Columbia University, New York*)
First, we congratulate the authors on a stimulating paper. Our attention was drawn in particular to Section 6.4 where we see some overlap with our own work.

One approach to detecting a lack of convergence is to estimate, using simulation, quantities that have known values under the target distribution. If θ denotes the parameter vector sampled via iterative simulation, then we can use simulation draws to estimate $E\{U(\theta)\}$ for any computable function U. Many diagnostic techniques are based on monitoring functions that converge to some specific value. However, in general this value is not known and so the resulting diagnostic is rather

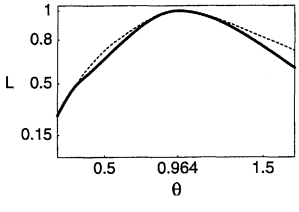

Fig. 9. Results from MIGRATE (———) superimposed on the Fig. 2 results of Stephens and Donnelly (·····): the MIGRATE estimate combines samples from the final long chains (100 000 steps each) of 10 independent runs

difficult to interpret in that it may have settled to some value, but it is unclear whether it is the *true* value (e.g. Gelman and Rubin (1992a)). With Markov chain Monte Carlo (MCMC) algorithms, it is often possible to diagnose convergence with multiple overdispersed sequences, but this approach does not work with algorithms such as importance sampling that do not have local dependence (Brooks and Gelman, 1998). This is one reason why we have found convergence monitoring to be easier for MCMC than for importance sampling. We therefore welcome the suggestion in Section 6.1 that the new importance sampling methods be used within an expanded MCMC framework.

The difficulties in monitoring convergence of functions $E\{U(\theta)\}$ would be removed if we knew the true expectation of U under the stationary distribution, and there are some functions U for which this is the case. One is the score function. If $\theta \in R^K$, and we let $\pi(\theta)$ denote the target distribution for the simulations, then we might take

$$U_k(\theta) = \frac{\partial[\log\{\pi(\theta)\}]}{\partial\theta_k}, \qquad k = 1, \ldots, K.$$

Under fairly general conditions on the density π, $E_\pi\{U_k(\theta)\} = 0$ for all $k = 1, \ldots, K$.

Thus, one might monitor the sample mean of each of these U_k functions as the simulations proceed, until they appear to settle to around 0. One can estimate the standard error of the $U_k(\theta)$ from parallel independent runs of the importance sampling or MCMC procedure, to determine whether or not observed values are 'significantly' different from 0.

It is not necessarily true that, as claimed in the paragraph before Section 5.1, parallel simulation runs are computationally more expensive. The results of the parallel runs can themselves be averaged, so inference from several runs is more efficient than from any single run. More importantly, parallel runs can give confidence about the accuracy of simulation results so the simulations may be stopped far earlier than might be done under the condition of insecurity arising from using a single simulation run (see Gelman and Rubin (1992b), and accompanying discussion).

Yuguo Chen and Jun S. Liu (*Stanford University*)

Stephens and Donnelly present a comprehensive account of an important problem in molecular evolution and a new sequential importance sampling (SIS) method for computation with coalescence models. For several decades, SIS has attracted attention from researchers in fields ranging from molecular simulation to statistics (Liu and Chen, 1998; Liu *et al*., 2000). A technique proven essential in many SIS applications but not covered by Stephens and Donnelly is *resampling*, also known as 'pruning and enrichment' in molecular simulations (Grassberger, 1997). Here we show how resampling improves the SIS computation for coalescence models.

As with many SIS applications, both Stephens and Donnelly, and Griffiths and Tavaré (1994) propose trial densities $q_0(\mathcal{H})$ of the form

$$q_\theta(\mathcal{H}) = \prod_{i=0}^{-(m-1)} q_\theta(H_{i-1}|H_i).$$

For such constructions we define the *current weight* (for $t \leqslant m$)

$$w_{-t} = \frac{p_\theta(H_{-(t-1)}|H_{-t}) \ldots p_\theta(H_0|H_{-1})}{q_0(H_{-t}|H_{-(t-1)}) \ldots q_0(H_{-1}|H_0)} \equiv w_{-(t-1)} \frac{p_\theta(H_{-(t-1)}|H_{-1})}{q_0(H_{-t}|H_{-(t-1)})}.$$

The final weight is then $w = w_{-m}\, p_\theta(H_{-m})\, p_\theta(|H_1| = n + 1)$.

In a parallel implementation of SIS, we first generate M samples from $q_0(H_{-1}|H_0)$ and then recursively generate $\{H_{-t}^{(1)}, \ldots, H_{-t}^{(M)}\}$, called the *current sample*, for $t = 2, 3, \ldots$ until coalescence in all M processes. Along with producing the current sample, we also monitor the current weight. A resampling step is incurred at time $-t$ when the coefficient of variation in $\{w_{-t}^{(1)}, \ldots, w_{-t}^{(M)}\}$ exceeds a threshold B. In resampling, one produces a new current sample by drawing with replacement from $\{H_{-t}^{(1)}, \ldots, H_{-t}^{(M)}\}$ according to probability proportional to $\{w_{-t}^{(1)}, \ldots, w_{-t}^{(M)}\}$. The weight for each new sample is equal to the *sample average* of the $w_{-t}^{(j)}$. Resampling helps to prune the H_{-t} with small weights and to enrich those with large weights.

We note, however, that resampling among $\{H_{-t}^{(1)}, \ldots, H_{-t}^{(M)}\}$ is inefficient because these samples differ greatly in their coalescence speeds. Those $H_{-t}^{(j)}$ with small sizes (fast coalescence) often have small current

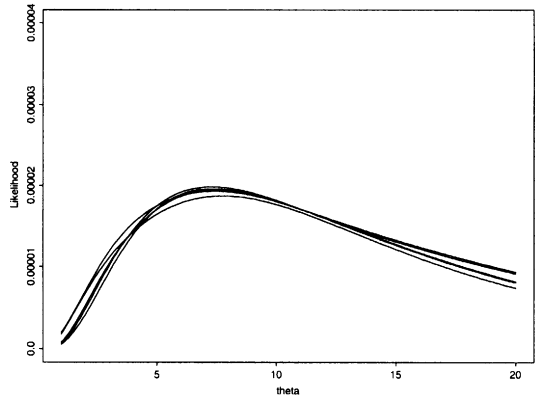

Fig. 10. Estimated likelihood curves for five independent runs of SIS with resampling: the scale difference between these curves and those in Fig. 3 is due to the fact that the authors omit the term $p_\theta(H_{-m})$, the stationary distribution of the mutation transition matrix

weights, but large final weights. A modification that we propose is to resample among the *same size* samples $\{H_{-i_1}^{(1)}, \ldots, H_{-i_M}^{(M)}\}$, where $i_j = \min\{t: |H_{-t}^{(j)}| = i\}$.

We applied our method with Griffiths and Tavaré's $q_0(\mathcal{H})$ to the example shown in Fig. 3. With sample size $M = 10000$ and bound $B = 4$, we incurred two resampling steps in each of the five independent runs. The extra computational cost was negligible. Fig. 10 displays the likelihood curves estimated from five runs of our method. Fig. 10 is almost indistinguishable from Fig. 3(b) resulting from the use of Stephens and Donnelly's new q-function. Lastly, we note that resampling can be combined with any SIS method including Stephens and Donnelly's to improve efficiency.

Mary Emond, Adrian E. Raftery and Russell Steele (*University of Washington, Seattle*)

We congratulate the authors on their paper, which helps us to understand the comparative performance of importance sampling and Markov chain Monte Carlo (MCMC) sampling in a complex setting. It is interesting that the generally simpler method of importance sampling compares favourably with MCMC sampling in the problems described. Could efficiency be gained by creating a more adaptive proposal distribution? The authors' proposal distribution depends on A_n in a fairly rigid manner. Could the dependence instead be parameterized by a small number of parameters that would then be estimated from an initial sample from the authors' proposal distribution? The study of histories with high weights from the authors' proposal distribution might suggest how to do this and how valuable it would be.

We have found this strategy to be useful in our study of importance sampling methods for computing integrated likelihoods for mixture models. These are an essential component of Bayes factors and posterior model probabilities used, for example, for choosing the number of components in the mixture (e.g. Raftery (1996)). We write the mixture model likelihood in its usual 'complete-data' form, the complete data consisting of the observed data X plus the missing group membership indicators Z. Integration of the complete-data likelihood over the parameters of the component distributions, τ, can often be done analytically. We use importance sampling to integrate over the finite space of group membership indicators.

Information from the maximum likelihood estimate for τ is used to create efficient importance sampling functions, or proposal distributions. In one such sampling scheme, the proposal distribution is itself a mixture of the form

$$h(Z) = \delta\, P(Z) + (1 - \delta)\, P(Z|\tau = \hat{\tau}, X).$$

This is an example of the 'defensive mixture' proposal distribution, which has the advantage of yielding importance sampling weights that are bounded above, thus ensuring a stable performance of the importance sampling estimator (Hesterberg, 1995).

The parameter δ is chosen to obtain good performance of the importance sampling scheme, and we have found the following adaptive scheme for choosing δ to work well. δ is initially taken to be 0.5 to

obtain an initial estimate of the integrated likelihood, $\hat{I}$. $\hat{I}$ is then used to obtain $\hat{P}(\hat{Z}_M|X) = P(X, \hat{Z}_M)/\hat{I}$, where $\hat{Z}_M$ is an estimate of $\arg\max_Z\{P(Z|X)\}$. An adaptive estimate of δ is then obtained by solving $h(\hat{Z}_M) = \hat{P}(\hat{Z}_M|X)$. This amounts to setting $h(Z)$ equal to the optimal distribution $P(Z|X)$ at its mode. $h(Z)$ may contain more than one component of the form $P(Z|\tau = \hat{\tau}_j, X)$, with the mixing proportions estimated adaptively by equating $h(Z)$ to $\hat{P}(Z|X)$ at more than one point. This adaptive approach not only decreases the variance of $\hat{I}$ but also increases the accuracy of its estimated standard error.

Joe Felsenstein (*University of Washington, Seattle*)

Stephens and Donnelly have made an excellent improvement on the Griffiths–Tavaré method which samples histories of mutation and coalescence events. The histories are sampled independently, which is desirable, but the improvement was needed to prevent the importance sampling from wasting a large fraction of its time. Stephens and Donnelly's improvement of it makes it more competitive with our Markov chain Monte Carlo (MCMC) approach (Kuhner *et al.*, 1995). It may be worthwhile to point out how our MCMC method differs from the Griffiths–Tavaré and Stephens–Donnelly sampler. The missing information $\mathcal{T}$ in equation (7) is not a history of discrete events but a genealogical tree connecting the samples, with times of coalescence specified but with no representation of mutations. Thus $\pi_\theta(A_n|\mathcal{T})$ must sum over all possible mutational events; fortunately this can be done using existing algorithms for computing likelihoods for phylogenies (Felsenstein, 1981). We propose changes in $\mathcal{T}$ by relocating branches on $\mathcal{T}$.

Our MCMC method samples from a more focused distribution than the Griffiths–Tavaré or Stephens–Donnelly independent sampling methods but, not being independent and identically distributed, runs the risk of failing to explore the space of genealogies adequately. However, it is comparatively straightforward to add new evolutionary forces; we have done so in separate programs, and a unified program is nearing completion.

Stephens and Donnelly raise some doubts about our methods. In this discussion, Kuhner and Beerli consider the poor performance of our methods in Stephens and Donnelly's test case. With more sites varying, we find an adequate performance in our simulations. Stephens and Donnelly also point out that the variance of our importance sampling weights can become infinite when the trial value of θ used in the sample is smaller than the true value by more than a factor of 2. It is important to note that this rule is for $n = 2$. For more reasonable sample sizes the conditions for misbehaviour may be more stringent. For example, Fig. 11 shows the progress of runs carried out by Peter Beerli on 10 cases with a sample size of 50 and 1000 sites. As we adjust the value of θ_0 in successive chains, an initial value of $\theta_0 = 0.0001$ successfully converges on estimates near the true value of $\theta = 0.01$ in all 10 runs. Most of the variation in the results is presumably natural rather than due to a failure of convergence.

Bret Larget (*Duquesne University, Pittsburgh*)

My comments pertain to the connections between the inference problems of molecular genetics described in the paper and the related problem of reconstructing evolutionary trees (phylogenies) from genetics data. In both problems the observable genetics data may be modelled as the result of a continuous time

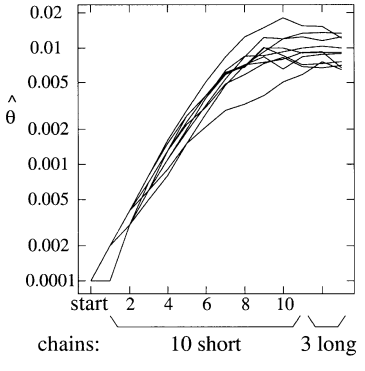

Fig. 11. 10 runs with initial starting-points of $\theta = 0.0001$ with 50 copies of one locus and 1000 sites: an MCMC sampler with 10 short chains and three long chains was run for each

stochastic process superimposed on a tree. In the problems addressed in the paper, the parameters of the stochastic process are of primary interest whereas the underlying tree is a nuisance parameter. For evolutionary biologists these roles are reversed. The authors make a convincing case that, in some situations, importance sampling (IS) is competitive with or superior to Markov chain Monte Carlo (MCMC) sampling as a computational tool for inference. As a developer of computational methods for phylogenetic inference based on MCMC methods (see Mau *et al*. (1999), Larget and Simon (1999) and Simon and Larget (2000)), I am intrigued to compare the techniques of this paper with our own.

Phylogenetic inferences produced by our methods are based on the post burn-in portion of a dependent sample of trees. As determined by comparing results from independent long runs, the amount of phylogenetic information in dependent samples can be similar to that from independent samples hundreds or thousands of times smaller. This paper indicates that IS may be a computationally tractable alternative.

I welcome a further expansion by the authors on the general characteristics for which an IS sampling scheme may outperform analysis by MCMC sampling. The discussion in Section 6 indicates that MCMC sampling may have an advantage when there are few constraints (meaning many trees have similar likelihoods for producing the observed data) whereas IS may have an advantage in more constrained situations.

The MCMC methods that we use propose new trees without regard to the genetics data and interact with the data only through the acceptance ratio. This has advantages and disadvantages. A sampler based on the methods of this paper may be much more efficient in some situations. However, changes to the form of the likelihood models by which genetics information evolves must entail substantial recoding for sampling methods based in part on the data whereas only minimal changes may be necessary in methods that ignore the data in proposing new trees. Direct applications of the ideas of this paper to phylogenetic inference and comparisons of the computational and statistical efficiencies would be most interesting.

Lada Markovtsova, Paul Marjoram and Simon Tavaré (*University of Southern California, Los Angeles*)

The authors have presented a most inspirational paper on computational methods for the coalescent. Their suggestion that one might combine Markov chain Monte Carlo (MCMC) and importance sampling (IS) approaches is particularly intriguing. As an example they suggest using the IS proposal distribution to update a random amount of the upper part of the tree. We have previously experimented with a pure MCMC approach in which the proposal distribution worked in a similar manner, replacing a random amount of the top of the tree with a *random* topology. Perhaps not surprisingly, such changes may have a very low acceptance rate. Updates that replaced the tree from a relatively high point were accepted frequently, but when the update involved a large part of the topology the acceptance rate was very low. One can improve this naïve approach by alternating such proposals with updates that replace a random amount of the lower part of the tree (again we generated a random topology), but it is not clear how this would be accomplished in the IS framework that the authors suggest. In our approach such a scheme was very inefficient, particularly for large sample sizes, but there is reasonable hope that the improved efficiency of the proposal distribution given in this paper might circumvent the problems that we experienced.

Our experience with implementations of a fully Bayesian approach for deoxyribonucleic acid sequence data (e.g. Markovtsova *et al*. (2000a)) supports the authors' observation that Bayesian methods provide useful computational tools even when one's interest is in maximum likelihood estimation. Checking the adequacy of the estimated likelihood near a maximum can be accomplished by using different priors. The apparent simplicity of estimating relative likelihoods from marginals of the posterior distribution and the prior seems difficult to pass up. Do the authors have any thoughts on when this approach is likely to be misleading? We also note that posterior trees and rates can be used in a bootstrap approach for checking model adequacy (e.g. Markovtsova *et al*. (2000b)).

Testing IS and MCMC implementations is notoriously difficult; the development of test examples seems worthwhile. Another useful approach is to compare results with those generated by simpler schemes like the rejection methods. For instance, we have found this helpful in checking tree topology updates.

Bob Mau (*University of Wisconsin, Madison*)

Problems in population genetics are often intractable with analytical methods once the sample size approaches 10. The reason is simple: the form of the likelihood, conditioned on the genealogy, is

determined by its tree shape, a nuisance parameter. To calculate the full likelihood, we are obliged to sum over all possible tree shapes — an impractical undertaking.

The computer-intensive approaches presented in this paper each recognize that a relatively minuscule number of tree shapes contribute almost all of the absolute likelihood mass. Locating and weighting tree shapes that support large chunks of mass is their common feature. Markov chain Monte Carlo sampling finds such shapes by running a Markov chain on the space of trees. It uses a proposal distribution that is independent of the underlying evolutionary model and the data, but subject to both in the Metropolis–Hastings acceptance step. Shape weights are proportional to the frequency of visits. By contrast, importance sampling employs a proposal distribution that incorporates a model and data. Every proposed sample is included in the estimate, weighted by a ratio of conditional likelihoods.

The authors cleverly sample families of tree shapes, each member having the same conditional likelihood, courtesy of the exchangeability assumption. Their concept of the history $\mathcal{H}$ yields a partial likelihood, summed over all genealogies in $\mathcal{H}$, in one pass. The efficacy of their algorithm derives from choosing the 'right' summary of the typed ancestry, and embedding the evolutionary process in the particle representation of Donnelly and Kurtz. Discovering representations that admit natural proposal mechanisms is the art to this science.

Perhaps my comments are obvious, but I had not made critical connections until I read this paper, and for that I am grateful to the authors. They give importance sampling a framework that motivates their proposal distribution. Too often, we are at a loss to understand *why* certain sampling algorithms work well, even if we know *how* they work.

One criticism concerns the absence of any discussion on how to choose a starting value θ_0. The authors acknowledge that their sample selection depends on θ_0, and that simulating with a θ_0 that is distant from the true maximum may yield a sample that excludes important histories. That it did not seem to matter in their reported examples is not a persuasive argument, especially since *a posteriori* one observes that they guessed quite well. Some direction is needed here.

Xiao-Li Meng (*University of Chicago*)

The Griffiths–Tavaré and the Geyer–Thompson methods respectively belong to what I shall label *normalized* importance sampling (NIS) and *unnormalized* IS (UIS), depending on whether we can *evaluate* the normalized proposal density or only an unnormalized one. The NIS category covers nearly all traditional applications of IS. The UIS category is largely a by-product of the Markov chain Monte Carlo revolution, which has made it routine to draw from densities computable up to a normalizing constant. However, these normalizing constants are the quantity of interest in the authors' applications. Let $Y_{\text{obs}} = A_n$, $Y_{\text{mis}} = \mathcal{H}$ (or $\mathcal{T}$) and $p(Y_{\text{mis}}|Y_{\text{obs}}, \theta) = P_\theta(\mathcal{H}|A_n)$. Then the almost tautological identity

$$L(\theta|Y_{\text{obs}}) = \int p(Y_{\text{mis}},\ Y_{\text{obs}}|\theta)\,\mu(\mathrm{d}Y_{\text{mis}}) = \int p(Y_{\text{obs}}|\theta)\,p(Y_{\text{mis}}|Y_{\text{obs}}, \theta)\,\mu(\mathrm{d}Y_{\text{mis}})$$

tells us two trivial but important facts:

(a) the optimal proposal density for NIS is $p(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$, the target density;
(b) $L(\theta|Y_{\text{obs}})$ is the *normalizing constant* of $p(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$, with the complete-data density $p(Y_{\text{mis}}, Y_{\text{obs}}|\theta)$ being the unnormalized density, viewed as a function of Y_{mis} only.

As a case of NIS, the authors' method uses (a) to choose an efficient proposal density guided by the unachievable target density. In contrast, Geyer–Thompson-like methods cash in on (b), turning the computation of likelihood ratios into computing ratios of normalizing constants of densities from which we can make draws. Interestingly, for these methods, the proposal density (class) is the target or optimal density (class) as specified in (a). The key design question is then at which points of θ shall we make these 'optimal' draws?

Although this question is not easy to answer in general, for univariate θ, with comparable computational load, a reasonably fine equal-spaced grid design can substantially outperform the Geyer–Thompson method as applied in the paper, as long as the estimation is done via (efficient) bridge sampling (see Meng and Wong (1996)), which is a multiproposal extension of UIS. Incidentally, an alternative way of utilizing the output from the Wilson–Balding Bayesian method is to use *path sampling*, which is an infinitely many proposal extension of UIS (Gelman and Meng, 1998). With enough draws $\{(Y_{\text{mis}}^{(i)}, \theta^{(i)}),\ i = 1, \ldots, M\}$ from a Wilson–Balding chain, we can estimate $\lambda(\theta) = \log\{L(\theta|Y_{\text{obs}})\} - \log\{L(\theta_0|Y_{\text{obs}})\}$ (assuming $\theta_0 < \theta$) by $\hat{\lambda}(\theta_0, \theta)$ from formula (15) of Gelman and

Meng (1998), page 167, which involves only finding the order statistics of $\{\theta^{(i)}, i = 1, \ldots, M\}$ and evaluations of $\{U_i = S(\theta^{(i)}|Y_{\text{mis}}^{(i)}, Y_{\text{obs}}), i = 1, \ldots, M\}$ where $S(\theta|Y_{\text{mis}}, Y_{\text{obs}})$ is the *complete-data* score function. These more advanced UIS methods require comparable or even less implementation effort (for example, no smoothing is used for $\hat{\lambda}(\theta_0, \theta)$), and several comparative studies (e.g. Meng and Schilling (1996), DiCiccio *et al*. (1997) and Jensen and Kong (1999)) give me reasons to speculate that they can offer non-trivial improvements over the methods compared in this timely paper.

E. A. Thompson (*University of Washington, Seattle*)

The proposed new class of importance sampling (IS) methods provides a breakthrough for Monte Carlo likelihood methods and greatly clarifies issues of effective Monte Carlo realization of latent variables in structured systems. Another such system arising in genetics analysis is the pedigree, more complex than coalescent trees in that diploid individuals have two parents. In inference from multilocus genetics data on members of a pedigree structure, there are two dimensions to the conditional independence structure of the latent variables which specify the descent of genes in pedigrees: $S_{i,j} = 0$ or $S_{i,j} = 1$ ($i = 1, \ldots, M$; $j = 1, \ldots, L$) as the maternal or paternal gene is transmitted in meiosis i at locus j. One is the structure imposed by Mendelian segregation, which provides that the vectors $S_{i,\cdot} = (S_{i,1}, \ldots, S_{i,L})$ are *a priori* independent. The other is the linear structure of the chromosome, which, in the absence of genetic interference, provides that the vectors $S_{\cdot,j} = (S_{1,j}, \ldots, S_{M,j})$ are first order Markov in j.

Both Markov chain Monte Carlo (MCMC) sampling and IS have been used in Monte Carlo likelihood analyses on pedigrees. Where exact single-locus computation is feasible, IS along the chromosome may be used to obtain an estimate of the likelihood (Kong *et al*., 1994; Irwin *et al*., 1994). Alternatively, an MCMC sampler updating jointly the components of $S_{\cdot,j}$ is quite effective (Heath, 1997). The latter provides only relative likelihoods but in a Bayesian framework allows for more complex models of trait determination. No thorough comparison of Monte Carlo performance of IS and MCMC sampling in this context has been made: examples can be constructed to favour either. Where exact single-locus computation is infeasible, an MCMC approach which updates jointly the components of $S_{i,\cdot}$ provides another alternative. This sampler has the advantage that it can be extended to more general meiosis models (Thompson, 2000). Although not in general irreducible, irreducibility can be guaranteed by combining the two MCMC samplers.

What do we know about the fourth possibility — IS rather than MCMC sampling over the meioses of the pedigree? As with coalescents, the data are at the bottom; gene ancestry must be realized within the pedigree structure. IS has been attempted on a very large and complex pedigree by C. J. Geyer (Geyer and Thompson, 1995), but without success. In the light of this paper, the reason is clear: we must consider backward rather than forward transition probabilities. How this is to be accomplished on a large and complex pedigree structure remains unclear, but the idea raises new hopes for Monte Carlo analyses of data on very complex pedigrees. In conjunction with MCMC sampling of $S_{i,\cdot}$, it would provide for irreducibility and an effective method for Metropolis-rejected restarts.

Valérie Ventura (*Carnegie Mellon University, Pittsburgh*)

A perfect importance sampling (IS) estimate of equation (4) is obtained using the sampler $Q_\theta^*(\mathcal{H}) \propto \pi_\theta(A_n|\mathcal{H})\, P_\theta(\mathcal{H})$ (Kahn and Marshall, 1953), i.e. $Q_\theta^*(\mathcal{H}) = P_\theta(\mathcal{H}|A_n)$. This paper provides a clever and effective approximation to $Q_\theta^*(\mathcal{H})$.

Because $P_\theta(\mathcal{H})$ is typically more spread than $\pi_\theta(A_n|\mathcal{H})$, sampling from densities such as P_θ or Q_θ^{GT} might yield pseudobiased estimates as in Section 5.2, even if M is huge (Ventura, 1997, 1998). Generally this means that there is a subset A of the sample space of $\mathcal{H}$ such that

$$\Pr\left(\bigcap_{i=1}^{M} \{\mathcal{H}^{(i)} \in A\}\right) > 1 - \epsilon$$

with ϵ small, and

$$E\{l_\theta(\mathcal{H})|\mathcal{H} \in A\} = a,$$

$$\text{var}\{l_\theta(\mathcal{H})|\mathcal{H} \in A\} = b,$$

while unconditionally

$$E\{l_\theta(\mathcal{H})\} = a + d,$$
$$\text{var}\{l_\theta(\mathcal{H})\} = b + c$$

with $c \gg b$. We estimate $E\{l_\theta(\mathcal{H})\}$ by $M^{-1}\,\Sigma_i\, l_\theta(\mathcal{H}^{(i)})$, which is close to a with high probability, and with estimated variance b/M, much smaller than the unconditional variance $(b+c)/M$ — so much smaller as to indicate that $E\{l_\theta(\mathcal{H})\}$ is significantly different from $a+d$. Here

$$l_\theta(\mathcal{H}) = \pi_\theta(A_n|\mathcal{H})\, w_\theta(\mathcal{H}),$$
$$w_\theta(\mathcal{H}) = P_\theta(\mathcal{H})/Q^{\text{SD}}_{\theta_0}(\mathcal{H})$$

and $A = \{\mathcal{H}\colon \omega_\theta(\mathcal{H}) \doteq 0\}$.

Use of the sampler $Q^*_{\theta_0}$ prevents pseudobias at θ_0 (incidentally, I am puzzled about the estimate of $L(15)$ being pseudobiased when the almost optimal $Q^{\text{SD}}_{\theta=15}$ is used) but does not guarantee success at θ_1, if $\pi_{\theta_0}P_{\theta_0}$ and $\pi_{\theta_1}P_{\theta_1}$ have different importance regions. The safe approach is the use of a mixture

$$Q(\mathcal{H}) = \sum_k \gamma_k\, Q_k(\mathcal{H}) \tag{37}$$

that provides coverage for all P_θ, and preferably consistent with A_n, e.g. with $Q_k = Q^{\text{SD}}_{\theta_k}$. Expression (37) is not optimal for any one $L(\theta)$. But efficiency may be recovered at all θs by using the control variate $w'_\theta = Q^{\text{SD}}_\theta(\mathcal{H})/Q(\mathcal{H})$ to form

$$M^{-1}\sum_{i=1}^{M} \pi_\theta(A_n|\mathcal{H}^{(i)})\frac{P_\theta(\mathcal{H}^{(i)})}{Q_{\theta_0}(\mathcal{H}^{(i)})} - \hat{\beta}'_\theta\left(M^{-1}\sum_i w'^{(i)}_\theta - 1\right), \tag{38}$$

with $\hat{\beta}'_\theta$ the least square regression slope of $l_\theta(\mathcal{H}^{(i)})$ on $w'^{(i)}_\theta$. Its asymptotic variance is $(1-\rho^2_\theta)^{-1}$ times smaller than that of expression (12), where typically $\rho^2_\theta = \text{corr}^2\{l_\theta(\mathcal{H}^{(i)}), w'^{(i)}_\theta\}$ is large since $\hat{Q}^*_\theta \doteq \hat{Q}^{\text{SD}}_\theta$ implies $l_\theta(\mathcal{H}^{(i)}) \propto w'^{(i)}_\theta$. This also works well if equation (37) is not everywhere consistent with A_n; see Ventura (1999).

Estimator (38) requires the evaluation of Q^{SD}_θ for all θ at all $\mathcal{H}^{(i)}$, but

$$M^{-1}\sum_{i=1}^{M} \pi_\theta(A_n|\mathcal{H}^{(i)})\frac{P_\theta(\mathcal{H}^{(i)})}{Q_{\theta_0}(\mathcal{H}^{(i)})} - \sum_k \hat{\beta}_{\theta k}\left\{M^{-1}\sum_i w_k(\mathcal{H}^{(i)}) - 1\right\}, \tag{39}$$

with $w_k(\mathcal{H}) = Q_k(\mathcal{H})/Q(\mathcal{H})$, and $\hat{\beta}_{\theta k}$ the estimates of $\beta_{\theta k}$ in the regression

$$l_\theta(\mathcal{H}) = \beta_0 + \sum_k \beta_{\theta k}\, w_k(\mathcal{H}) + \epsilon,$$

implicitly approximates expression (38), yet does not require additional computation other than the $\hat{\beta}_{\theta k}$. Owen and Zhou (2000) derived expression (39) independently and further proved that, although not optimal, it is never worse than the regression estimator from an importance sample of size $M\gamma_k$ from Q_k. Therefore equation (37) in conjunction with expression (38) or (39) makes IS safe and efficient.

The performances of estimators (12), (38) and (39) hinge on the quality of the mixture. But, as noted by the authors, more formal procedures are still needed to assess this, as well as the variability of the estimates.

The **authors** replied later, in writing, as follows.

We thank all the discussants for their interesting comments. For brevity, our response will focus on several common themes raised.

Several discussants (Mau, Meng, Stephens and Ventura) raise the issue of improving efficiency when estimating likelihood curves by combining estimates for different driving values θ_0. Although our method did not seem to experience difficulties in the simple examples considered in the paper — a fact that we speculate is in part due to the small size of the problems, and in part because our missing data $\mathcal{H}$ do not include times explicitly — this will be an important consideration in more complex problems. Much of the discussion of Geyer and Thompson (1992) focused on potential problems caused by failing

to combine estimates for different driving values, and led to the development by Geyer (1991) of the 'reverse logistic regression' method implemented by Kuhner and Beerli. Within our group the use of bridge sampling (as proposed in the paper, and by Meng) has also proved fruitful (Fearnhead, 2000). We welcome the additional methods suggested by Meng (path sampling) and Ventura. We also agree with Markovtsova and colleagues that in the Markov chain Monte Carlo (MCMC) framework the use of Bayesian machinery to find likelihood surfaces is appealing in its simplicity, particularly when the parameter of interest sits in a space of dimension 1 or 2, making estimation of the posterior density on a fine grid of parameter values relatively straightforward.

The improvements to the GT algorithm achieved by Chen and Liu's use of resampling (which has a subtle connection with the rejection control method suggested in Section 6.3) are very impressive. We look forward to assessing the improvements that its use offers to our method, especially in complex problems.

The suggestion of Emond and colleagues of an adaptive proposal distribution is one which we had considered. We have not yet been able to find any promising implementations. As noted in Section 6.4, we believe that the use of a 'defensive mixture' (Hesterberg, 1995) is unlikely to be fruitful in this context, though our view is tempered slightly by its success in another complex setting.

The question of diagnostics for importance sampling (IS)—assessing whether enough iterations have been performed for accurate estimation—is clearly important. Fig. 12 shows plots of the estimated mean and standard deviation of the importance weights, against the number of iterations, for 20000 samples from $Q^{SD}_{\theta_0=15.0}$ in the case $\theta = 15$ in Section 5.2. The discontinuities are striking and evidence that

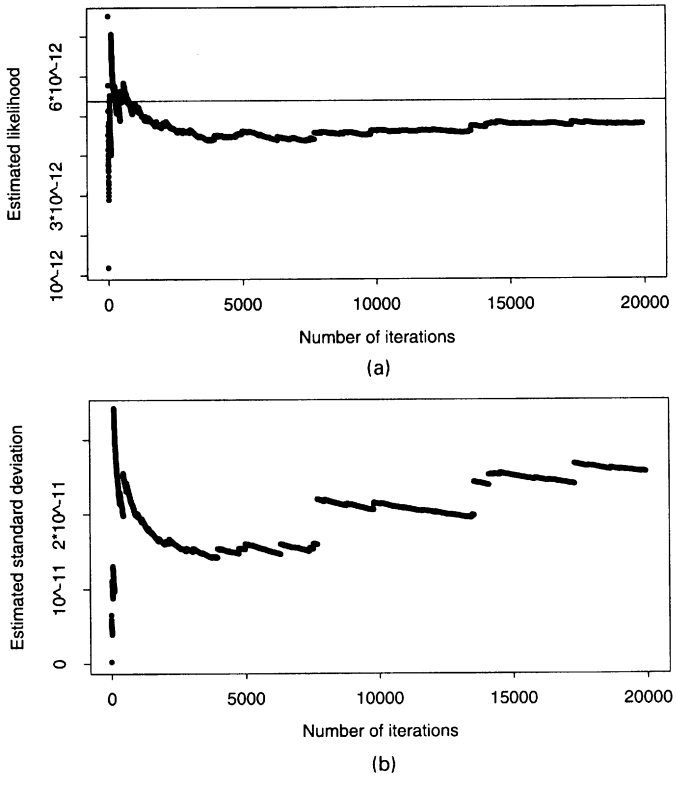

Fig. 12. (a) Estimated likelihood (mean of importance weights) and (b) standard deviation of the importance weights, against number of iterations, for 20000 samples from $Q^{SD}_{\theta_0=15.0}$, in the case $\theta = 15$ in Section 5.2: the horizontal line in (a) shows the estimate of the likelihood obtained from the 10^7 samples used in Table 1, illustrating the tendency for IS methods to underestimate the likelihood if too few samples are used; the standard deviation is even more substantially underestimated (using 10^7 samples we obtained an estimate of 2.6×10^{-10}, which is off the top of the scale of the graph shown)

insufficient iterations have been performed to overcome the extreme skewness of the distribution of the importance weights. Monitoring of the effective sample size, another diagnostic which we have found helpful, gives the same message. We welcome the suggestion by Brooks and Gelman, of a more formal approach, though it is not clear to us how their diagnostic could be implemented in our case. Our target distribution, which they denote π, consists of atoms in the space of histories and so cannot be differentiated, at least in the usual sense. Incidentally, we are in complete agreement with Brooks and Gelman on the use of parallel runs.

We are grateful to Beaumont, Fearnhead, Felsenstein, Griffiths, Kuhner and Beerli, Markovtsova and colleagues, and Wilson for further descriptions of the methods that they have developed. They, and other discussants, also highlight the importance of checking the accuracy of new, complex algorithms. The growing availability of different methods for the same kinds of data will facilitate this, and we welcome the suggestion by Markovtsova and colleagues of establishing standard test data sets. In the light of some of the comments made in the discussion, it is perhaps worth reiterating that not all MCMC methods involve IS, and that different MCMC methods can, and do, and IS methods could, involve different choices of the missing data. The efficiency of different such choices is not yet well understood, but Kuhner and Beerli's suggestions about the way in which these choices may affect the efficiency for different sorts of data is intriguing.

The result of Stephens (1999) can be extended to show that the variance of the IS weights for the `Fluctuate` scheme (Kuhner *et al.*, 1998) is infinite for $\theta > 2\theta_0$, *regardless* of the value of the sample size n. We thus do not share Felsenstein's optimism on this matter, and we note that Fig. 11 does not bear directly on the problem, which relates to underestimating the likelihood away from its peak. (Incidentally, the suggestion in Fig. 11 of differences by more than a factor of 2 in the estimated maximum likelihood estimates from different runs is somewhat disquieting.)

In his discussion, Stephens raises the important issue of robustness, explicitly in the context of a possible misspecification of the mutation matrix P. Analogous questions could be posed regarding other genetic and demographic modelling assumptions. For reasons given in the paper we considered a somewhat stylized inference problem. For real genetics problems little is known about the robustness of likelihood procedures, but we suspect that it is likely to be a serious problem. In fact, as Harding noted, in many practical problems the interesting inference questions concern these other aspects of the model, rather than the value of θ.

Ewens and Joyce echo our hope that theoretical results concerning the sampling properties of maximum likelihood estimators in these models will soon be available, and Edwards reminds us of the alternative possibility of regarding the log-likelihood as an intrinsic measure of support. The imminent availability of data from multiple, independent chromosomal regions across the genome will have a major effect on accuracy for many estimation problems. With n sampled chromosomes sampled from r regions, the information will grow as r or $r \log(n)$, in problems for which information can be combined across regions. Indeed, these developments promise to revolutionize many genetic inference problems (e.g. Di Rienzo *et al.* (1998), Reich *et al.* (1999), Pritchard and Rosenberg (1999), Devlin and Roeder (1999), Nielsen (2000), Pritchard *et al.* (2000a, b) and Wall and Przeworski (2000)) and the resulting methods at least partially address some of the important practical applications raised by Harding.

The theory developed in the paper, and its straightforward extension to more general demographic and genetics models, offers considerable guidance in the choice of proposal distributions, in either the IS or MCMC frameworks. None-the-less, there is a non-trivial human cost in the development of such proposals, at least some of which must be incurred anew in each different setting. A careful design of the proposal distribution, with the associated cost, seems essential to the development of practicable IS schemes, for all except the simplest problems. In contrast, as Felsenstein, Larget and Mau note, it is possible to design MCMC proposals which are independent of many of the details of the underlying model. Our IS proposal distribution depends on each of the data, the demographic model and the genetics model. Proposal distributions of the MCMC methods described in the paper and the discussion depend on various subsets of these three factors. To what extent there are 'plug-and-play' MCMC proposals that work well for a wide variety of models is not yet clear, but the difficulties described by Markovtsova and colleagues in using one such proposal provide a salutary warning. We remain convinced that insights gained through a careful consideration of the underlying stochastic models have a role to play in the development of computationally intensive inference procedures, in this and related settings. (Similar considerations with regard to the trade-offs between generic and tailored methods will apply in connection with Griffiths's suggestions concerning solutions of linear equations.)

Several discussants point to the connections between inference in population genetics and in other areas

of genetics. We are grateful to Larget for his description of the phylogenetic inference problem in which he and others have successfully used MCMC methods. We also enjoyed Thompson's insightful comments on the challenges of, and possibilities in, computationally intensive inference within pedigrees.

Ewens's challenge to adapt these methods to the important problem of mapping susceptibility genes for complex diseases from population data is timely. The insights gained from coalescent models, e.g. into the correlation structure of such data, seem destined to be important. Whether or not full coalescent-based likelihood procedures represent the best way to address the practical problem is not clear to us.

Finally, we appreciate the historical perspectives offered by Edwards, Ewens and Griffiths, and we join Harding in her implicit hope that the future will see ever more rapid progress towards tackling the challenging real world inference problems in molecular population genetics.

References in the discussion

Baldi, P. and Brunak, S. (1998) *Bioinformatics, the Machine Learning Approach*. Cambridge: MIT Press.

Beaumont, M. A. (1999) Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013–2029.

Beerli, P. and Felsenstein, J. (1999) Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763–773.

Brookfield, J. F. Y. (1986) A model for DNA sequence evolution within transposable element families. *Genetics*, **112**, 393–407.

Brooks, S. and Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.*, **7**, 434–455.

Chen, M. H. (1994) Importance-weighted marginal Bayesian posterior density-estimation. *J. Am. Statist. Ass.*, **89**, 818–824.

Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Diaconis, P. and Holmes, S. P. (1998) Matchings and phylogenetic trees. *Proc. Natn. Acad. Sci. USA*, **95**, 14600–14602.

DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997) Computing Bayes factors by combining simulation and asymptotic approximations. *J. Am. Statist. Ass.*, **92**, 903–915.

Di Rienzo, A., Donnelly, P., Toomajian, C., Sisk, B., Hill, A., Petzl-Erler, M. L., Haines, G. and Barch, D. (1998) Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. *Genetics*, **148**, 1269–1284.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological Sequence Analysis*. Cambridge: Cambridge University Press.

Edwards, A. W. F. (1970) Estimation of the branch points of a branching diffusion process (with discussion). *J. R. Statist. Soc.* B, **32**, 155–174.

——— (1972) *Likelihood*. Cambridge: Cambridge University Press.

——— (1992) *Likelihood*, expanded edn. Baltimore: Johns Hopkins University Press.

Fearnhead, P. N. (2000) Estimating recombination rates from population genetic data. To be published.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Molec. Evoln*, **17**, 368–376.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Gelman, A. and Meng, X.-L. (1998) Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, **13**, 163–185.

Gelman, A. and Rubin, D. B. (1992a) A single sequence from the Gibbs sampler gives a false sense of security. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 625–631. New York: Oxford University Press.

——— (1992b) Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.*, **7**, 457–511.

Geyer, C. (1991) Reweighting Monte Carlo mixtures. *Technical Report 568*. School of Statistics, University of Minnesota, Minneapolis. (Available from `http://stat.umn.edu/PAPERS/tr568r.html`.)

Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc.* B, **54**, 657–699.

——— (1995) Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Statist. Ass.*, **90**, 909–920.

Grassberger, P. (1997) Pruned-enriched Rosenbluth method: simulations of θ polymers of chain length up to 1000000. *Phys. Rev.* E, **56**, 3682–3693.

Griffiths, R. C. and Marjoram, P. (1996) Ancestral inference from samples of dna sequences with recombination. *J. Comput. Biol.*, **3**, 479–502.

Griffiths, R. C. and Tavaré, S. (1994) Simulating probability distributions in the coalescent. *Theory Popln Biol.*, **46**, 131–159.

Heath, S. C. (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am. J. Hum. Genet.*, **61**, 748–760.

Hesterberg, T. C. (1995) Weighted average importance sampling and defensive mixture distributions. *Technometrics*, **37**, 185–194.

Irwin, M., Cox, N. and Kong, A. (1994) Sequential imputation for multilocus linkage analysis. *Proc. Natn Acad. Sci. USA*, **91**, 11684–11688.

Jensen, C. S. and Kong, A. (1999) Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops. *Am. J. Hum. Genet.*, **65**, 885–901.

Kahn, H. and Marshall, A. W. (1953) Methods of reducing sample size in Monte Carlo computations. *J. Ops Res. Soc. Am.*, **1**, 263–278.

Kong, A., Liu, J. S. and Wong, W. H. (1994) Sequential imputation and Bayesian missing data problems. *J. Am. Statist. Ass.*, **89**, 278–288.

Kuhner, M. K., Yamato, J. and Felsenstein, J. (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421–1430.

———(1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429–434.

———(2000) Maximum likelihood estimation of recombination rates from population data. Submitted to *Genetics*.

Larget, B. and Simon, D. (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molec. Biol. Evoln*, **16**, 750–759.

Li, S., Pearl, D. K. and Doss, H. (2000) Phylogenetic tree construction using Markov chain Monte Carlo. *J. Am. Statist. Ass.*, to be published.

Liu, J. S. and Chen, R. (1998) Sequential Monte Carlo methods for dynamic systems. *J. Am. Statist. Ass.*, **93**, 1032–1044.

Liu, J. S., Chen, R. and Logvinenko, T. (2000) A theoretical framework for sequential importance sampling and resampling. In *Sequential Monte Carlo Methods in Practice* (eds A. Doucet, J. F. G. de Freitas and N. J. Gordon). New York: Springer.

Markovtsova, L., Marjoram, P. and Tavaré, S. (2000a) The age of a unique event polymorphism. *Genetics*, **156**, in the press.

———(2000b) The effects of rate variation on ancestral inference in the coalescent. *Genetics*, to be published.

Mau, B., Newton, M. A. and Larget, B. (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, **55**, 1–12.

Meng, X.-L. and Schilling, S. (1996) Fitting full-information item factor models and an empirical investigation of bridge sampling. *J. Am. Statist. Ass.*, **91**, 1254–1267.

Meng, X.-L. and Wong, W. H. (1996) Simulating ratios of normalizing constants via a simple identity: a theoretical explanation. *Statist. Sin.*, **6**, 831–860.

Newton, M. A., Mau, B. and Larget, B. (1999) Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. In *Proc. American Mathematical Society–Institute of Mathematical Statistics–Society for Industrial and Applied Mathematics Joint Summer Research Conf. Statistics and Molecular Biology* (eds F. Seillier-Moiseiwitsch *et al.*)

Nielsen, R. (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**, 931–942.

Owen, A. and Zhou, Y. (2000) Safe and effective importance sampling. *J. Am. Statist. Ass.*, **95**, 135–143.

Pritchard, J. K. and Rosenberg, N. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.*, **65**, 220–228.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000a) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pritchard, J. K., Stephens, M., Rosenberg, N. A. and Donnelly, P. (2000b) Association mapping in structured populations. *Am. J. Hum. Genet.*, **67**, 170–181.

Raftery, A. E. (1996) Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, D. J. Spiegelhalter and S. Richardson), pp. 163–188. London: Chapman and Hall.

Rannala, B. and Zhang, Z. H. (1997) Probability distributions of molecular evolutionary trees—a new method of phylogenetic inference. *J. Molec. Biol.*, **43**, 304–311.

Reich, D., Feldman, M. and Goldstein, D. (1999) Statistical properties of two tests that use multilocus data sets to detect population expansions. *Molec. Biol. Evoln*, **16**, 453–466.

Simon, D. and Larget, B. (2000) Bayesian analysis in molecular biology and evolution (BAMBE), version 2.02 beta. Department of Mathematics and Computer Science, Duquesne University, Pittsburgh.

Stephens, M. (1999) Problems with computational methods in population genetics. *Bull. 52nd Sess. Int. Statist. Inst.*, book 1, 273–276.

Thompson, E. A. (2000) MCMC estimation of multi-locus genome sharing and multipoint gene location scores. *Int. Statist. Rev.*, **68**, 53–73.

Ventura, V. (1997) Likelihood inference by Monte Carlo methods and efficient nested bootstrapping. *DPhil Thesis*. University of Oxford, Oxford.

———(1998) Nonparametric bootstrap recycling. *Technical Report 673*. Statistics Department, Carnegie Mellon University, Pittsburgh.
———(1999) Double importance sampling. *Technical Report 694*. Statistics Department, Carnegie Mellon University, Pittsburgh.
Wall, J. D. and Przeworski, M. (2000) When did the human population size start increasing? *Genetics*, to be published.
Wilson, I. J. and Balding, D. J. (1998) Genealogical inference from microsatellite data. *Genetics*, **150**, 499–510.
Wuif, C. and Hein, J. (1999) Recombination as a point process along sequences. *Theor. Popln Biol.*, **55**, 248–259.