Adv. Appl. Prob. **36,** 417–433 (2004)
Printed in Northern Ireland

IMPORTANCE SAMPLING ON COALESCENT HISTORIES. I

MARIA DE IORIO,[*] *Imperial College London*
ROBERT C. GRIFFITHS,[**] *University of Oxford*

Abstract

Stephens and Donnelly (2000) constructed an efficient sequential importance-sampling proposal distribution on coalescent histories of a sample of genes for computing the likelihood of a type configuration of genes in the sample. In the current paper a characterization of their importance-sampling proposal distribution is given in terms of the diffusion-process generator describing the distribution of the population gene frequencies. This characterization leads to a new technique for constructing importance-sampling algorithms in a much more general framework when the distribution of population gene frequencies follows a diffusion process, by approximating the generator of the process.

Keywords: Coalescent process; diffusion process; importance sampling

2000 Mathematics Subject Classification: Primary 60G40
Secondary 93E25; 92D25

1. Introduction

Molecular data have a sample distribution which is a mixture over possible ancestries. The state space of ancestries is huge and closed-form expressions for the probability distribution of the type configuration of genes in a sample are only known in the simplest case. An active research area in recent years has been in computing sample likelihoods by Markov chain Monte Carlo (MCMC), Bayesian or importance-sampling (IS) methods on the ancestral process back in time. A selection of research is: on IS, Griffiths and Tavaré (1994a), (1994b), (1994c), (1996), (1997), (1999), Griffiths and Marjoram (1996), Nath and Griffiths (1996), Nielsen (1997), Bahlo and Griffiths (2000), Stephens and Donnelly (2000), Slade (2000a), (2000b), Fearnhead and Donnelly (2001), Stephens (2001); MCMC by Felsenstein *et al.* (1999), Beerli and Felsenstein (1999), Kuhner *et al.* (1995), (1997); and other approaches, sometimes MCMC and of a Bayesian nature by Wilson and Balding (1998), Wilson *et al.* (2003), Beaumont (1999), Markovtsova *et al.* (2000a), (2000b), Beaumont (2001), Wakeley *et al.* (2001). Liu (2001) discussed many modern MCMC and IS techniques which improve simulation performance when the state space of the process is large. Parameter estimation and ancestral inference from the underlying stochastic model of ancestry is possible once the likelihood can be computed. The methods are extremely computer intensive because of the large state space of genealogies. Stephens and Donnelly (2000) delineated the way to use IS efficiently in a model where there is a general Markov mutation mechanism between gene types.

Received 6 November 2002; revision received 19 December 2003.
[*] Postal address: Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7 2BZ, UK.
[**] Postal address: Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK.
Email address: griff@stats.ox.ac.uk
Supported by BBSRC Bioinformatics grant 43/BIO14435.

In this paper a new method is presented for deriving IS proposal distributions based on the diffusion-process generator that describes the distribution of the population gene frequencies. An approximate sampling distribution is found which then leads to an IS proposal distribution. The technique is very general because it is common in population genetics to assume a diffusion-process model for population gene frequencies. The proposal distribution is identical to that of Stephens and Donnelly (2000) for the same model. There is also mathematical interest in the characterization of the Stephens–Donnelly proposal distribution in terms of an approximation to the diffusion-process generator.

A gene tree is a perfect phylogeny constructed from the configuration of mutations on a sample of DNA sequences when the infinitely-many-sites model of mutation holds. The IS proposal distribution can be extended to gene trees. This allows ancestral inference, such as ages of mutations and time to the most recent common ancestor, of sequences to be computed conditional on the topology of the gene tree (Griffiths and Tavaré (1999)).

A companion paper, De Iorio and Griffiths (2004), considers a general class of subdivided population models and develops an IS proposal distribution based on the techniques in this paper for computing the likelihood of a sample of genes from subpopulations. De Iorio *et al.* (2004) considered in detail IS for a stepwise-mutation model in subdivided populations.

2. Importance sampling

A coalescent history $\{H_k,\ k = 0, -1, \dots, -m\}$ is defined as the set of ancestral configurations at the embedded events in the Markov process where coalescence, mutation or other events take place. H_0 denotes the current state, and H_{-m} the state when a singleton ancestor is reached. In the original coalescent process (Kingman (1982)), coalescent events from k to $k-1$ ancestors take place at rate $\binom{k}{2}$, $k = n, \dots, 2$. Here a more general ancestral process is envisaged which might include many complex models, for example those with migration, recombination and selection. The Markov nature of the process implies that the probability $p(H_k)$ of a configuration H_k is given by

$$p(H_k) = \sum_{\{H_{k-1}\}} p(H_k \mid H_{k-1}) p(H_{k-1}). \tag{1}$$

While $p(H_k)$ and $\{p(H_{k-1})\}$ are unknown, the probabilities $p(H_k \mid H_{k-1})$ are straightforward to derive from the distribution of the coalescent tree. The Stephens–Donnelly IS representation is based on finding a good approximation to the reverse chain probabilities $\hat{p}(H_{k-1} \mid H_k)$. Their IS representation is then

$$p(H_0) = \mathrm{E}_{\hat{p}}\left[\frac{p(H_0 \mid H_{-1})}{\hat{p}(H_{-1} \mid H_0)} \cdots \frac{p(H_{-m+1} \mid H_{-m})}{\hat{p}(H_{-m} \mid H_{-m+1})} p(H_{-m})\right]. \tag{2}$$

Here, $\mathrm{E}_{\hat{p}}$ denotes expectation taken over histories in the reverse direction, $H_{-1}, \dots, H_{-m}$, with the reverse chain transition probabilities $\hat{p}(H_{k-1} \mid H_k)$. Thus, the likelihood of the data can be evaluated by repeated simulation of sample histories under $\hat{p}$, averaging the importance weights to approximate (2). If $\hat{p} = p$, (2) simplifies to $p(H_0)$, because then the argument in the expectation can be expressed as

$$\frac{p(\mathcal{H}_{\rightarrow})}{p(\mathcal{H}_{\leftarrow})/p(H_0)} = p(H_0)\ ,$$

where $p(\mathcal{H}_{\rightarrow})$ and $p(\mathcal{H}_{\leftarrow})$ are probabilities of a history sample path, evaluated as Markov chain probabilities in forward and reverse directions respectively.

3. Sampling distributions

Let $E = \{1, 2, \ldots, d\}$ be the type space for a collection of n genes. A type change occurs by mutation from a parent to an offspring according to a transition matrix P. There are two equivalent ways of viewing the distribution of a sample of n genes:

(i) directly from a coalescent model; and

(ii) as a sample from a population of genes with a random distribution of gene frequencies for the d types.

In this section we briefly review Stephens and Donnelly's method of constructing proposal distributions, pointing out the significance of sampling exchangeability and relating the proposal-distribution construction to an approximation deduced from the diffusion generator describing the distribution of population gene frequencies.

In the coalescent model (i) the type of the ancestor is chosen from the stationary distribution of P, then changes of type occur along the edges of the coalescent tree at a rate $\theta/2$ according to a Markov chain with transition matrix $P = (P_{ij})$. The sample configuration of types is determined in the n leaves of the tree. Let $\boldsymbol{n} = (n_i)_{i \in E}$ denote the numbers of genes of different types in a sample of n and let $p(\boldsymbol{n})$ be the probability distribution of the unordered type configuration $\boldsymbol{n}$. An argument considering whether the first event back in time was a mutation with probability $\theta/(n - 1 + \theta)$ or a coalescence with probability $(n - 1)/(n - 1 + \theta)$, and then what the sample configuration prior to the event must have been leads to a recursive set of equations for $p(\boldsymbol{n})$ from (1),

$$p(\boldsymbol{n}) = \frac{\theta}{n + \theta - 1} \sum_{i, j \in E, n_j > 0} \frac{n_i + 1 - \delta_{ij}}{n} P_{ij} p(\boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i) + \frac{n-1}{n + \theta - 1} \sum_{j \in E, n_j > 0} \frac{n_j - 1}{n - 1} p(\boldsymbol{n} - \boldsymbol{e}_j). \tag{3}$$

This is the model in which IS techniques were developed by Griffiths and Tavaré (1994c) and Stephens and Donnelly (2000) to find $p(\boldsymbol{n})$ by simulation. In (3), $\{\boldsymbol{e}_j\}$ is a unit vector and δ_{ij} is 1 if $i = j$ and 0 if $i \neq j$. In the history process, if $H_k = \boldsymbol{n}$, then $H_{k-1} = \boldsymbol{n} - \boldsymbol{e}_j$ if a pair of type-j genes coalesce, or $H_{k-1} = \boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i$ if forward in time a mutation changed a type-i gene to a type-j gene. A coalescent tree illustrating the history process back in time is shown in Figure 1.

Let $\mathcal{A}_n = (a_1, \ldots, a_n)$ denote the ordered type configuration of n genes sequentially sampled from the population. An important *exchangeability condition* is that

$$p(a_{\tau(1)}, \ldots, a_{\tau(n)}) = p(a_1, \ldots, a_n) \tag{4}$$

for any permutation τ of $1, \ldots, n$. This exchangeability condition implies that

$$p(\boldsymbol{n}) = \binom{n}{\boldsymbol{n}} p(\mathcal{A}_n)$$

for an ordered configuration $\mathcal{A}_n$ corresponding to an unordered type configuration $\boldsymbol{n}$. The multinomial coefficient is denoted by $\binom{n}{\boldsymbol{n}}$. Define $\pi(i \mid \boldsymbol{n})$ as the probability that an additional type chosen from the population is of type i, given a sample configuration of $\boldsymbol{n}$. The exchangeability

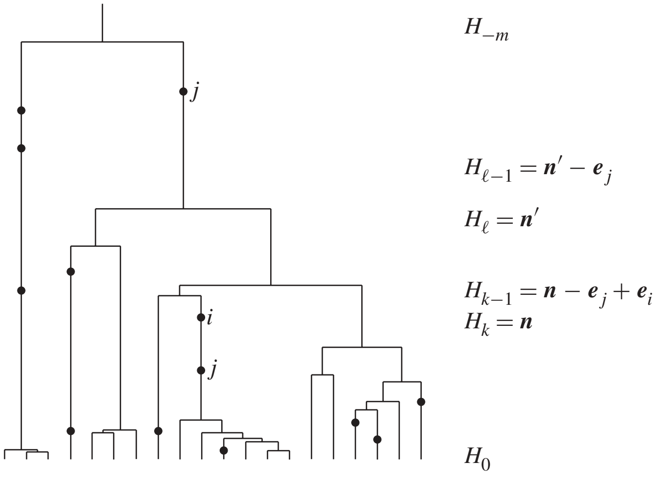
H_{-m}
j
$H_{\ell-1} = \boldsymbol{n}' - \boldsymbol{e}_j$
$H_\ell = \boldsymbol{n}'$
$H_{k-1} = \boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i$
i
$H_k = \boldsymbol{n}$
j
H_0

FIGURE 1: Coalescent tree with history process.

condition (4) implies the *symmetry condition*

$$\pi(i \mid \boldsymbol{n} - \boldsymbol{e}_j) p(\boldsymbol{n} - \boldsymbol{e}_j) = \frac{n_i + 1 - \delta_{ij}}{n} p(\boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i) \tag{5}$$

by noting that each side of (5) is equal to

$$\binom{n-1}{\boldsymbol{n} - \boldsymbol{e}_j} p(a_1, \dots, a_{\ell-1}, a_{\ell+1}, \dots, a_{n-1}, i)$$

for an ordered configuration of $n - 1$ genes $(a_1, \dots, a_{\ell-1}, a_{\ell+1}, \dots, a_{n-1})$ corresponding to an unordered configuration $\boldsymbol{n} - \boldsymbol{e}_j$. The type of the omitted ℓth gene is $a_\ell = j$. A condition equivalent to (5), replacing $\boldsymbol{n} - \boldsymbol{e}_j$ by $\boldsymbol{n}$, is

$$\pi(i \mid \boldsymbol{n}) p(\boldsymbol{n}) = \frac{n_i + 1}{n + 1} p(\boldsymbol{n} + \boldsymbol{e}_i).$$

The reverse chain distributions $p(H_{k-1} \mid H_k)$ can be expressed in terms of π by using Bayes's rule and (5):

$$p(H_{k-1} \mid H_k) = p(H_k \mid H_{k-1}) \frac{p(H_{k-1})}{p(H_k)}$$

and, if $H_k = \boldsymbol{n}$,

$$\begin{aligned} p(H_{k-1} \mid H_k) &= \begin{cases} \dfrac{n_j - 1}{n + \theta - 1} \dfrac{p(\boldsymbol{n} - \boldsymbol{e}_j)}{p(\boldsymbol{n})} & \text{if } H_{k-1} = \boldsymbol{n} - \boldsymbol{e}_j, \\ \dfrac{\theta(n_i + 1 - \delta_{ij})/n}{n + \theta - 1} \dfrac{p(\boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i)}{p(\boldsymbol{n})} & \text{if } H_{k-1} = \boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i \end{cases} \\ &= \begin{cases} \dfrac{n_j(n_j - 1)}{n(n + \theta - 1)} \dfrac{1}{\pi(j \mid \boldsymbol{n} - \boldsymbol{e}_j)} & \text{if } H_{k-1} = \boldsymbol{n} - \boldsymbol{e}_j, \\ \dfrac{\theta}{n + \theta - 1} \dfrac{n_j}{n} \dfrac{\pi(i \mid \boldsymbol{n} - \boldsymbol{e}_j)}{\pi(j \mid \boldsymbol{n} - \boldsymbol{e}_j)} & \text{if } H_{k-1} = \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_j. \end{cases} \end{aligned} \tag{6}$$

The sampling probability ratios in the case $H_{k-1} = \boldsymbol{n} - \boldsymbol{e}_j$ of (6) can be simplified by using the symmetry condition (5). Taking $i = j$ in (5) for the first ratio,

$$\frac{p(\boldsymbol{n} - \boldsymbol{e}_j)}{p(\boldsymbol{n})} = \frac{n_j}{n} \frac{1}{\pi(j \mid \boldsymbol{n} - \boldsymbol{e}_j)}$$

and, in the case $H_{k-1} = \boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i$ of (6),

$$\begin{aligned}\frac{p(\boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i)}{p(\boldsymbol{n})} &= \frac{p(\boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i)}{p(\boldsymbol{n} - \boldsymbol{e}_j)} \frac{p(\boldsymbol{n} - \boldsymbol{e}_j)}{p(\boldsymbol{n})} \\ &= \frac{\pi(i \mid \boldsymbol{n} - \boldsymbol{e}_j)}{(n_i + 1 - \delta_{ij})/n} \frac{n_j}{n} \frac{1}{\pi(j \mid \boldsymbol{n} - \boldsymbol{e}_j)}.\end{aligned}$$

IS proposal distributions can be found by substituting an approximation $\hat{\pi}$ for π in (6) to obtain $\hat{p}(H_{k-1} \mid H_k)$.

3.1. The Stephens–Donnelly sampling approximation

Stephens and Donnelly (2000) constructed an IS proposal distribution on coalescent histories approximating π by $\hat{\pi}$, the stationary distribution in a Markov chain with transition probability matrix

$$\frac{\theta P + N}{n + \theta},$$

where N is the $d \times d$ matrix with each row equal to $\boldsymbol{n} = (n_1, \ldots, n_d)$. That is, for $j \in E$,

$$\hat{\pi}(j \mid \boldsymbol{n}) = \sum_{i \in E} \hat{\pi}(i \mid \boldsymbol{n}) \frac{\theta P_{ij} + n_j}{n + \theta}. \tag{7}$$

The solution to (7) is that $\hat{\pi}(j \mid \boldsymbol{n})$ is the jth entry in the vector $(\boldsymbol{n}/n)(1 - \rho)(I - \rho P)^{-1}$, where $\rho = \theta/(n + \theta)$ and I is the identity matrix. Define the sampling approximation $\hat{p}(\boldsymbol{n})$ recursively in terms of $\hat{\pi}$ by

$$\hat{p}(\boldsymbol{n}) = \sum_{j \in E} \hat{\pi}(j \mid \boldsymbol{n} - \boldsymbol{e}_j) \hat{p}(\boldsymbol{n} - \boldsymbol{e}_j), \tag{8}$$

ending with the probability distribution of choosing a singleton sample gene as the stationary distribution of P. In the case of parent-independent mutation with $E = \{1, \ldots, d\}$ when $P_{ij} = P_j$, $\hat{\pi} = \pi$ and the approximate sampling distribution is the real sampling distribution, with

$$\pi(j \mid \boldsymbol{n}) = \frac{\theta}{\theta + n} P_j + \frac{n}{\theta + n} \frac{n_j}{n}$$

and

$$p(\boldsymbol{n}) = \binom{n}{\boldsymbol{n}} \frac{[\theta P_1]_{(n_1)} \cdots [\theta P_d]_{(n_d)}}{\theta_{(n)}},$$

using the notation $r_{(k)} = r(r + 1) \cdots (r + k - 1)$.

How close is the approximate sampling distribution $\hat{p}(\boldsymbol{n})$ to the real sampling distribution $p(\boldsymbol{n})$?

Proposition 1. *If $\hat{\pi}$ and $\hat{p}$ defined by (7) and (8) satisfy the symmetry condition (5), then $\hat{p} = p$.*

We expect that the symmetry condition will only hold in the parent-independent-mutation case; however, it shows how close p and $\hat{p}$ are and where exchangeability is needed.

Proof. We have

$$\begin{aligned}
\hat{p}(\boldsymbol{n}) &= \sum_{j\in E} \hat{\pi}(j \mid \boldsymbol{n} - \boldsymbol{e}_j)\hat{p}(\boldsymbol{n} - \boldsymbol{e}_j) \\
&= \frac{\theta}{n+\theta-1} \sum_{i,j\in E, n_j>0} \hat{\pi}(i \mid \boldsymbol{n} - \boldsymbol{e}_j)\hat{p}(\boldsymbol{n} - \boldsymbol{e}_j)P_{ij} \\
&\quad + \frac{1}{n+\theta-1} \sum_{j\in E, n_j>0} (n_j - 1)\hat{p}(\boldsymbol{n} - \boldsymbol{e}_j) \\
&= \frac{\theta}{n+\theta-1} \sum_{i,j\in E, n_j>0} \frac{n_i + 1 - \delta_{ij}}{n}\hat{p}(\boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i)P_{ij} \qquad (9) \\
&\quad + \frac{n-1}{n+\theta-1} \sum_{j\in E, n_j>0} \frac{n_j - 1}{n-1}\hat{p}(\boldsymbol{n} - \boldsymbol{e}_j). \qquad (10)
\end{aligned}$$

Simplification to obtain (9) uses (5) whereas (10) is true since $\sum_{i\in E} \hat{\pi}(i \mid \boldsymbol{n}) = 1$, irrespective of (5). Thus, $\hat{p}$ satisfies the same recursive set of equations as p with the type distribution of singleton samples being the same, implying that $\hat{p} = p$.

3.2. Population frequencies

The population frequencies of types of genes $\boldsymbol{X} = (X_i)_{i\in E}$ (with $E = \{1, \dots, d\}$) are distributed according to the stationary distribution in a diffusion process with state space

$$S = \left\{\boldsymbol{x} = (x_i)_{i\in E} : x_i \geq 0, \sum_{i\in E} x_i = 1\right\}$$

and generator

$$\mathcal{L} = \frac{1}{2}\sum_{i,j\in E} x_i(\delta_{ij} - x_j)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{j\in E}\left(\sum_{i\in E} x_i r_{ij}\right)\frac{\partial}{\partial x_j}\ , \qquad (11)$$

where the mutation-rate matrix $R = (r_{ij})$ is defined as

$$R = \frac{\theta}{2}(P - I).$$

The sample distribution is a multinomial mixture with respect to the stationary distribution of the population frequencies, that is,

$$p(\boldsymbol{n}) = \binom{n}{\boldsymbol{n}}\mathrm{E}\left(\prod_{i\in E} X_i^{n_i}\right),$$

where E denotes expectation in the stationary distribution of the diffusion process. The recursion (3) for $p(\boldsymbol{n})$ can be derived by simplifying the generator equation

$$\mathrm{E}\left\{\mathcal{L}\prod_{i\in E} X_i^{n_i}\right\} = 0. \tag{12}$$

This equation holds because $\mathrm{E}(\mathcal{L}g(\boldsymbol{X})) = 0$ for any bounded continuous function g whose second derivatives exist. Details of the derivation of (3) from a coalescent approach (i) and the diffusion generator approach (ii) using (12) are given in Griffiths and Tavaré (1994c).

3.3. Diffusion-generator approximation

This section describes a general technique to find an approximation for the sampling distribution $p(\boldsymbol{n})$ when the distribution of population gene frequencies is described by a generator of the form

$$\mathcal{L} = \sum_{j\in E} L_j \frac{\partial}{\partial x_j}, \quad \text{where } L_j = \frac{1}{2}\sum_{i\in E} x_i(\delta_{ij} - x_j)\frac{\partial}{\partial x_i} + \mu_j(\boldsymbol{x}), \tag{13}$$

with $\{\mu_j(\boldsymbol{x})\}$ and $\{x_i(\delta_{ij} - x_j)\}$ being the infinitesimal means and covariances. For example, in the model described by (11), $\mu_j(\boldsymbol{x}) = \sum_{i\in E} x_i r_{ij}$. The technique is to suppose that there is a distribution with expectation operator $\hat{\mathrm{E}}$ such that, for each $j \in E$,

$$\hat{\mathrm{E}}\left\{L_j\left[\frac{\partial}{\partial X_j}\prod_{i\in E} X_i^{n_i}\right]\right\} = 0. \tag{14}$$

The assumption (14) is equivalent to

$$\hat{\mathrm{E}}\left\{L_j\prod_{i\in E} X_i^{n_i}\right\} = 0, \tag{15}$$

by increasing the degree n_j by 1 in (14). In the equations obtained below, we make the approximation

$$\hat{p}(\boldsymbol{n}) = \binom{n}{\boldsymbol{n}}\hat{\mathrm{E}}\left\{\prod_{i\in E} X_i^{n_i}\right\},$$

then assume that $\mathrm{E} = \hat{\mathrm{E}}$. Further simplification is then obtained by assuming that the symmetry condition (5) holds. This is a device to simplify the equations and express them in terms of $\hat{\pi}$; solutions to (14) or (15) over different j will in general not be consistent or satisfy exchangeability. A solution to the set of equations from (15) depends on the infinitesimal means being polynomials in $\boldsymbol{x}$, or on making further approximations to obtain a solution. The aim is to find suitable approximations to use for IS, rather than a numerical analysis objective of finding a tight approximation. An interesting point is that this technique leads to sampling approximations and an IS proposal distribution that can be explained entirely from the diffusion generator. Of course the coalescent history is in the background, but it does not have to be explicitly invoked as it is here, in that the IS technique could be seen just as a way to solve the recursive system of equations for the sampling distribution.

The functions $\{\prod_{i\in E} x_i^{n_i}\}$ form a basis of all functions g with continuous bounded second derivatives and domain S. Thus, for such a function g and for each $j \in E$,

$$\mathrm{E}(L_j g(\boldsymbol{X})) = 0. \tag{16}$$

Note that

$$0 = \mathrm{E}(\mathcal{L}X_j g(\boldsymbol{X})) = \mathrm{E}(X_j \mathcal{L} g(\boldsymbol{X})) + \mathrm{E}(L_j g(\boldsymbol{X})) + \mathrm{E}\biggl(X_j(1 - X_j)\frac{\partial g(\boldsymbol{X})}{\partial X_j}\biggr),$$

showing that the assumption (16) is equivalent to

$$\mathrm{E}(X_j \mathcal{L} g(\boldsymbol{X})) = -\,\mathrm{E}\biggl(X_j(1 - X_j)\frac{\partial g(\boldsymbol{X})}{\partial X_j}\biggr). \tag{17}$$

In the sampling equation (15), $g = \boldsymbol{x}^{\boldsymbol{n}} := \prod_{i\in E} x_i^{n_i}$. The Stephens–Donnelly approximation (7) for $\hat{\pi}$ can be obtained directly from (14) as follows. Note that

$$\begin{aligned} L_j \frac{\partial}{\partial x_j}\boldsymbol{x}^{\boldsymbol{n}} &= \frac{1}{2}\sum_{i\in E} x_j(\delta_{ij} - x_i) n_j (n_i - \delta_{ij}) \boldsymbol{x}^{\boldsymbol{n}-\boldsymbol{e}_i-\boldsymbol{e}_j} + \frac{\theta}{2}\sum_{i\in E} x_i (P_{ij} - \delta_{ij}) n_j \boldsymbol{x}^{\boldsymbol{n}-\boldsymbol{e}_j} \\ &= \frac{1}{2} n_j \biggl\{-(n + \theta - 1)\boldsymbol{x}^{\boldsymbol{n}} + (n_j - 1)\boldsymbol{x}^{\boldsymbol{n}-\boldsymbol{e}_j} + \theta \sum_{i\in E} P_{ij} \boldsymbol{x}^{\boldsymbol{n}-\boldsymbol{e}_j+\boldsymbol{e}_i}\biggr\}. \end{aligned}$$

Thus,

$$\begin{aligned} L_j \frac{\partial}{\partial x_j}\binom{n}{\boldsymbol{n}}\boldsymbol{x}^{\boldsymbol{n}} = \frac{1}{2}\biggl\{ &-n_j(n + \theta - 1)\binom{n}{\boldsymbol{n}}\boldsymbol{x}^{\boldsymbol{n}} + n(n_j - 1)\binom{n-1}{\boldsymbol{n} - \boldsymbol{e}_j}\boldsymbol{x}^{\boldsymbol{n}-\boldsymbol{e}_j} \\ &+ \theta \sum_{i\in E} P_{ij}(n_i + 1 - \delta_{ij})\binom{n}{\boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i}\boldsymbol{x}^{\boldsymbol{n}-\boldsymbol{e}_j+\boldsymbol{e}_i}\biggr\}. \end{aligned}$$

Taking the expectation with respect to the stationary distribution of gene frequencies with the assumption (14),

$$n_j(n - 1 + \theta)\hat{p}(\boldsymbol{n}) = n(n_j - 1)\hat{p}(\boldsymbol{n} - \boldsymbol{e}_j) + \sum_{i\in E}\theta P_{ij}(n_i + 1 - \delta_{ij})\hat{p}(\boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i). \tag{18}$$

Assuming that the symmetry condition (5) holds for $\hat{\pi}$ and $\hat{p}$, then substituting into (18) yields that

$$n(n+\theta-1)\hat{\pi}(j \mid \boldsymbol{n}-\boldsymbol{e}_j)\hat{p}(\boldsymbol{n}-\boldsymbol{e}_j) = n(n_j-1)\hat{p}(\boldsymbol{n}-\boldsymbol{e}_j)+n\sum_{i\in E}\theta P_{ij}\hat{\pi}(i \mid \boldsymbol{n}-\boldsymbol{e}_j)\hat{p}(\boldsymbol{n}-\boldsymbol{e}_j). \tag{19}$$

Dividing this equation by $n(n + \theta - 1)\hat{p}(\boldsymbol{n} - \boldsymbol{e}_j)$ and replacing $\boldsymbol{n}$ with $\boldsymbol{n} + \boldsymbol{e}_j$ produces the stationary distribution equation (7) for $\hat{\pi}$. Conversely, retracing steps back from (19) shows that assuming the symmetry condition and the stationary distribution equation for $\hat{\pi}$ holding implies (14). If it were true that $\hat{\pi} = \pi$, then the symmetry condition automatically holds because, by Proposition 1, $\hat{p} = p$. A characterization of the Stephens–Donnelly assumption in terms of the diffusion-process generator is thus the following.

Proposition 2. *The Stephens–Donnelly assumption that π satisfies the stationary distribution equation (7) is equivalent to (16) or (17).*

There are two equivalent methods of obtaining a system of equations for $\hat{\pi}$: (i) apply the diffusion-generator technique and then use a symmetry argument, and (ii) equate terms in the sampling recursion inside the summation for each $j \in E$ and then use a symmetry argument. The form of (3) for the method (ii) is

$$\sum_{j \in E} \biggl(n_j(n+\theta-1)p(\boldsymbol{n}) - n(n_j-1)p(\boldsymbol{n}-\boldsymbol{e}_j) \\ - \theta \sum_{i \in E} (n_i + 1 - \delta_{ij}) P_{ij} p(\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j) \biggr) = 0. \tag{20}$$

Equating each term inside the summation with $j \in E$ to zero in (20) gives (18), and thus (19).

3.4. Justification of the coalescent and diffusion approximation

3.4.1. *Coalescent justification.* The following is a justification of the approximation (18) based on the coalescent. Let B_j be the event that a gene of type $j \in E$ is the first to be involved in either a coalescent or a mutation event back in time, and let $\boldsymbol{Y}$ be a random vector describing the configuration of types so that $\mathrm{P}(\boldsymbol{Y}=\boldsymbol{n}) = p(\boldsymbol{n})$. Then

$$\begin{aligned} \mathrm{P}(\{\boldsymbol{Y}=\boldsymbol{n}\} \cap B_j) &= p(\boldsymbol{n})\mathrm{P}(B_j \mid \boldsymbol{Y}=\boldsymbol{n}) \\ &= \frac{n-1}{n+\theta-1}\frac{n_j-1}{n-1} p(\boldsymbol{n}-\boldsymbol{e}_j) \\ &\quad + \frac{\theta}{n+\theta-1}\sum_{i \in E} \frac{n_i+1-\delta_{ij}}{n} P_{ij} p(\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j). \end{aligned} \tag{21}$$

The equation (21) is exact from a probabilistic argument, rather than approximate. Approximating $\mathrm{P}(B_j \mid \boldsymbol{Y}=\boldsymbol{n})$ by

$$\hat{\mathrm{P}}(B_j \mid \boldsymbol{Y}=\boldsymbol{n}) = \frac{n_j}{n} \tag{22}$$

yields the approximation (19). A heuristic argument for the approximation (22) is the following. Suppose that the matrix P has a stationary distribution $(P_j)_{j \in E}$. Then the unconditional probability that, at a given time in a sample's history, the first event prior to the time occurred to a gene of type j is

$$\mathrm{P}(B_j) = P_j. \tag{23}$$

This is because of symmetry as to which gene is involved in the event, and because P_j is the unconditional expected proportion of type-j genes at a given time in a sample's ancestry. Thus, the approximation (22) is a sample approximation to (23). Even when the matrix P may not have a stationary distribution, (22) is a sensible local approximation in the coalescent process. The condition (22) is exact in the parent-independent-mutation model when $P_{ij} = P_j$ for all $i, j \in E$.

3.4.2. *Moran model and diffusion-limit justification.* The justification argument is now extended from the sample to the population level via a Moran model which approximates the diffusion model. The Moran model is a multitype birth-and-death process with a fixed population size of N genes. Letting $z_i(t)$ be the number of genes of type $i \in E$ at time t and $\boldsymbol{z}(t) = (z_i(t))_{i \in E}$ for $t \geq 0$, the state space of the process is $\{(z_i)_{i \in E}, \sum_{i \in E} z_i = N\}$. Let $\lambda_{ij}^{(N)}(\boldsymbol{z})$ denote the rate corresponding to a death of type $i \in E$ and a birth of type $j \in E$. A process whose limit as $N \to \infty$ is the diffusion process with generator (13) has

$$\lambda_{ij}^{(N)}(z) = x_i \left(\frac{N^2}{2} x_j + N \mu_j(\boldsymbol{x}) \right), \qquad i, j \in E,$$

with $x_i = z_i / N$. The functions $\{\mu_j(\boldsymbol{x})\}$ are infinitesimal means from the diffusion process. The choice of a gene to die is uniform in the model. The diffusion-process limit for $\boldsymbol{X}(t) = \boldsymbol{Z}(t)/N$ follows by a calculation of the infinitesimal mean and covariance matrix of the increments $\Delta \boldsymbol{X}(t) = \boldsymbol{X}(t + \mathrm{d}t) - \boldsymbol{X}(t)$. The total death rate for type-i genes is

$$\lambda_{i\cdot}^{(N)}(z) = \sum_{j \in E, j \neq i} \lambda_{ij}^{(N)}(z) = x_i \left(\frac{N^2}{2}(1 - x_i) - N \mu_i(\boldsymbol{x}) \right), \tag{24}$$

and the total birth rate for type-j genes is

$$\lambda_{\cdot j}^{(N)}(z) = \sum_{i \in E, i \neq j} \lambda_{ij}^{(N)}(z) = (1 - x_j) \left(\frac{N^2}{2} x_j + N \mu_j(\boldsymbol{x}) \right). \tag{25}$$

The identity $\sum_{j \in E} \mu_j(\boldsymbol{x}) = 0$, which is true because $\sum_{j \in E} X_j(t) = 1$, is used in simplifying (24) and (25). The mean and covariance matrix of the increments are easily found by noting that, conditional on $\boldsymbol{Z}(t) = z$, $\boldsymbol{Z}(t + \mathrm{d}t) - \boldsymbol{Z}(t)$ is equal to $\boldsymbol{e}_j - \boldsymbol{e}_i$ with probability $\lambda_{ij}^{(N)}(z)\,\mathrm{d}t + o(\mathrm{d}t)$. Thus, when $i \neq j$, as $N \to \infty$,

$$\begin{aligned}
\lim_{\mathrm{d}t \to 0} (\mathrm{d}t)^{-1}\, \mathrm{E}(\Delta X_j(t) \mid \boldsymbol{X}(t) = \boldsymbol{x}) &= \frac{1}{N}(\lambda_{\cdot j}^{(N)}(z) - \lambda_{j\cdot}^{(N)}(z)) \\
&= \mu_j(\boldsymbol{x}), \\
\lim_{\mathrm{d}t \to 0} (\mathrm{d}t)^{-1}\, \mathrm{E}(\Delta X_j^2(t) \mid \boldsymbol{X}(t) = \boldsymbol{x}) &= \frac{1}{N^2}(\lambda_{j\cdot}^{(N)}(z) + \lambda_{\cdot j}^{(N)}(z)) \\
&\approx x_j(1 - x_j), \\
\lim_{\mathrm{d}t \to 0} (\mathrm{d}t)^{-1}\, \mathrm{E}(\Delta X_i(t) \Delta X_j(t) \mid \boldsymbol{X}(t) = \boldsymbol{x}) &= -\frac{1}{N^2}(\lambda_{ij}^{(N)}(z) + \lambda_{ji}^{(N)}(z)) \\
&\approx -x_i x_j, \\
\lim_{\mathrm{d}t \to 0} (\mathrm{d}t)^{-1}\, \mathrm{E}(|\Delta X_j^r(t)| \mid \boldsymbol{X}(t) = \boldsymbol{x}) &= \frac{1}{N^r}(\lambda_{j\cdot}^{(N)}(z) + \lambda_{\cdot j}^{(N)}(z)) \\
&\approx \frac{x_j(1 - x_j)}{N^{r-2}}, \qquad r \geq 2.
\end{aligned} \tag{26}$$

The conditions (26) imply weak convergence of the Moran process to the diffusion process (see Ethier and Kurtz (1986, Theorem 3.5, p. 428)).

Let $\boldsymbol{Z}$ be a random vector describing the configuration of types and $\mathcal{B}_j$ be the event that the first birth back in time was of type $j \in E$. Then in the stationary distribution, an exact equation is

$$\left(\sum_{k, \ell \in E} \lambda_{k\ell}^{(N)}(z) \right) \mathrm{P}(\{\boldsymbol{Z} = z\} \cap \mathcal{B}_j) = \sum_{i \in E} \lambda_{ij}^{(N)}(z - \boldsymbol{e}_j + \boldsymbol{e}_i) p(z - \boldsymbol{e}_j + \boldsymbol{e}_i),$$

that is,

$$\begin{aligned} p(z)\,\mathrm{P}(\mathscr{B}_j \mid \boldsymbol{Z} = z) &= \mathrm{P}(\{\boldsymbol{Z} = z\} \cap \mathscr{B}_j) \\ &= \frac{2}{N^2}\sum_{i\in E}\lambda_{ij}^{(N)}(z - \boldsymbol{e}_j + \boldsymbol{e}_i)p(z - \boldsymbol{e}_j + \boldsymbol{e}_i) \end{aligned} \tag{27}$$

since

$$\sum_{k,\ell\in E}\lambda_{k\ell}^{(N)}(z) = \frac{N^2}{2}.$$

The sample analogue of (27) is obtained by multiplying (27) by the hypergeometric probability of obtaining a sample of size $\boldsymbol{n}$ without replacement,

$$h^{(N)}(z;\boldsymbol{n}) = \frac{\prod_{k\in E}\binom{z_k}{n_k}}{\binom{N}{n}},$$

then taking expectation in the stationary distribution of $\boldsymbol{Z}$. The system of equations obtained for $j \in E$ is

$$\mathrm{E}(h^{(N)}(\boldsymbol{Z};\boldsymbol{n})p(\mathscr{B}_j \mid \boldsymbol{Z})) = \frac{2}{N^2}\sum_{i\in E}\mathrm{E}(h^{(N)}(\boldsymbol{Z} + \boldsymbol{e}_j - \boldsymbol{e}_i;\boldsymbol{n})\lambda_{ij}^{(N)}(\boldsymbol{Z})) \tag{28}$$

$$= \mathrm{E}\bigg(\bigg(X_j + \frac{2}{N}\mu_j(\boldsymbol{X})\bigg)\sum_{i\in E}\frac{Z_i}{N}h^{(N)}(\boldsymbol{Z} + \boldsymbol{e}_j - \boldsymbol{e}_i;\boldsymbol{n})\bigg). \tag{29}$$

The sum on the right-hand side of (29) is

$$\begin{aligned} &\sum_{i\in E}\frac{z_i}{N}h^{(N)}(z;\boldsymbol{n} + \boldsymbol{e}_j - \boldsymbol{e}_i) \\ &\quad= \sum_{i\in E}\frac{z_i + \delta_{ij}}{N}h^{(N)}(z + \boldsymbol{e}_j - \boldsymbol{e}_i;\boldsymbol{n}) - \frac{1}{N}h^{(N)}(z;\boldsymbol{n}) \\ &\quad= \frac{n+1}{N}\frac{\binom{N+1}{n+1}}{\binom{N}{n}}\sum_{i\in E}\frac{n_i + 1}{n+1}h^{(N+1)}(z + \boldsymbol{e}_j;\boldsymbol{n} + \boldsymbol{e}_i) - \frac{1}{N}h^{(N)}(z;\boldsymbol{n}) \\ &\quad= \frac{n+1}{N}\frac{\binom{N+1}{n+1}}{\binom{N}{n}}h^{(N+1)}(z + \boldsymbol{e}_j;\boldsymbol{n}) - \frac{1}{N}h^{(N)}(z;\boldsymbol{n}) \\ &\quad= \bigg(\frac{N+1-n}{N}\frac{z_j + 1}{z_j - n_j + 1} - \frac{1}{N}\bigg)h^{(N)}(z;\boldsymbol{n}) \\ &\quad= \bigg(1 - \frac{n}{N} + \frac{n_j}{Nx_j}\bigg)h^{(N)}(z;\boldsymbol{n}) + o(N^{-1}), \end{aligned} \tag{30}$$

where $No(N^{-1}) \to 0$ as $N \to \infty$. The identity

$$h^{(N+1)}(z + \boldsymbol{e}_j;\boldsymbol{n}) = \sum_{i\in E}\frac{n_i + 1}{n+1}h^{(N+1)}(z + \boldsymbol{e}_j;\boldsymbol{n} + \boldsymbol{e}_i)$$

used in (30) is immediate by an interpretation that a sample of n from $\boldsymbol{z} + \boldsymbol{e}_j$ can be obtained by choosing a sample of $n+1$ then randomly omitting one of the sample members. Thus, from (29) and (30),

$$\begin{aligned}
&\mathrm{E}(h^{(N)}(\boldsymbol{Z}; \boldsymbol{n})\,\mathrm{P}(\mathcal{B}_j \mid \boldsymbol{Z})) \\
&\quad = \mathrm{E}\biggl(X_j\biggl(1 + \frac{2\mu_j(\boldsymbol{X})}{NX_j}\biggr)\biggl(1 - \frac{n}{N} + \frac{n_j}{NX_j}\biggr)h^{(N)}(\boldsymbol{Z}; \boldsymbol{n})\biggr) + o(N^{-1}) \\
&\quad = \mathrm{E}\biggl(X_j\biggl(1 + \frac{2\mu_j(\boldsymbol{X})}{NX_j} - \frac{n}{N} + \frac{n_j}{NX_j}\biggr)h^{(N)}(\boldsymbol{Z}; \boldsymbol{n})\biggr) + o(N^{-1}).
\end{aligned} \tag{31}$$

Rearranging (31),

$$\begin{aligned}
&\mathrm{E}(h^{(N)}(\boldsymbol{Z}; \boldsymbol{n})N(\mathrm{P}(\mathcal{B}_j \mid \boldsymbol{Z}) - x_j)) \\
&\quad = \mathrm{E}((2\mu_j(\boldsymbol{X}) - nX_j + n_j)h^{(N)}(\boldsymbol{Z}; \boldsymbol{n})) + No(N^{-1}).
\end{aligned} \tag{32}$$

The equation (32) is an exact equation to $o(N^{-1})$, rather than an approximation. We make the local approximation that

$$\hat{p}(\mathcal{B}_j \mid \boldsymbol{Z} = z) = x_j, \tag{33}$$

or at least that $N(\hat{\mathrm{P}}(\mathcal{B}_j \mid \boldsymbol{Z} = z) - x_j) \to 0$ as $N \to \infty$. The heuristic argument for (33) is a similar population analogue of the argument used to obtain (22). Assuming (33), the left-hand side of (32) is zero and, taking the limit as $N \to \infty$,

$$\mathrm{E}\biggl((n_j - nX_j + 2\mu_j(\boldsymbol{X}))\binom{n}{\boldsymbol{n}}\prod_{i\in E} X_i^{n_i}\biggr) = 0. \tag{34}$$

It is straightforward to check that

$$2L_j \prod_{i\in E} x_i^{n_i} = (n_j - nx_j + 2\mu_j(x))\prod_{i\in E} x_i^{n_i},$$

so (34) and (15) are equivalent.

As a summary then, the main approximation assumption (22) that $\hat{\mathrm{P}}(B_j \mid \boldsymbol{Y} = \boldsymbol{n}) = n_j/n$ is equivalent to $\mathrm{E}\{L_j \prod_{i\in E} X_i^{n_i}\} = 0$; and the approximation that $\hat{\mathrm{P}}(\mathcal{B}_j \mid \boldsymbol{Z} = z) = x_j$ in the Moran model is asymptotically equivalent as $N \to \infty$ to the two mentioned sampling conditions in the limit diffusion process.

4. Gene trees

A model widely used in biological applications is the infinitely-many-sites model, where mutations occur on DNA sequences. New mutations are assumed to occur at sites never previously mutant. Each mutation produces a new gene type agreeing with the infinitely-many-alleles model, but the detail of where mutations occur is recorded. The configuration of mutations on the sequences is equivalent to a gene tree $\mathcal{T}$, a tree whose vertices are labelled mutations. Each gene is labelled by its path of mutations to the root of the tree, $j = (z_0, z_1, \dots)$, with $\{z_\ell\}$ denoting mutation labels. A sample configuration is thus $(\mathcal{T}, \boldsymbol{n})$, where $\mathcal{T} = \{j_1, \dots, j_n\}$ describes mutation paths and $\boldsymbol{n}$ denotes multiplicities of the sequences. Types in a gene tree are mutation paths from the leaves to the root of the tree (labelled 0). Figure 2 illustrates a gene tree of 50 sequences. Vertices in the tree are labelled mutations. Numbers

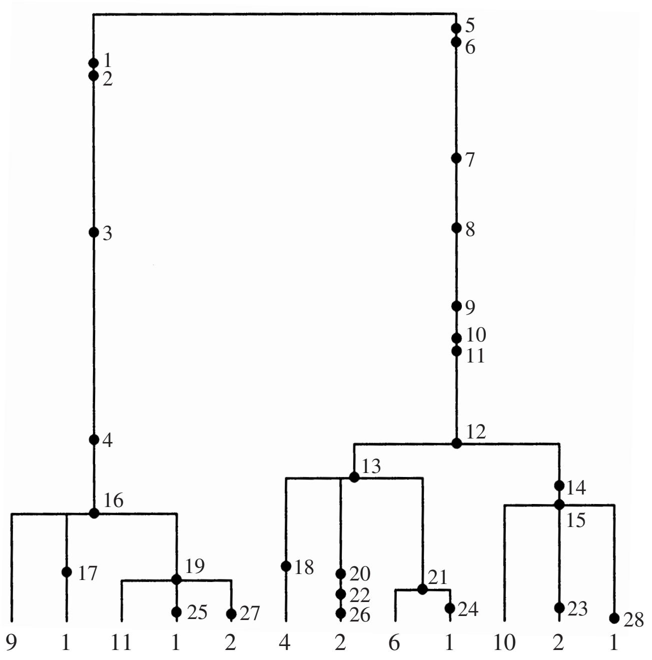

FIGURE 2: A gene tree.

under the leaves of the tree are multiplicities of the sequences in $\boldsymbol{n}$. For example, the type of the nine genes at the extreme left of the tree is $(16, 4, 3, 2, 1, 0)$. Detail about the distribution of gene trees is given in Ethier and Griffiths (1987), Griffiths (1989), (2001) and Griffiths and Tavaré (1994a), (1999). There is much evolutionary interest in regarding a sample of DNA sequences as a gene tree; a classic example is the β-globin study of Harding *et al.* (1997), a gene tree from fungus DNA is given in Carbone and Kohn (2001) and a review article is by Griffiths (2001). The analogue of (3) for a gene tree is

$$\begin{aligned} p(\mathcal{T}, \boldsymbol{n}) = {} & \frac{n-1}{n-1+\theta} \sum_{k:n_k \geq 2} \frac{n_k - 1}{n-1} p(\mathcal{T}, \boldsymbol{n} - \boldsymbol{e}_k) \\ & + \frac{\theta}{n-1+\theta} \sum_{k:n_k=1} \frac{1}{n} p(\mathcal{T}'_{k-}, \boldsymbol{n}) \\ & + \frac{\theta}{n-1+\theta} \sum_{\substack{k:n_k=1 \\ k \to j}} \frac{n_j + 1}{n} p(\mathcal{T}''_{k-}, \boldsymbol{n} + \boldsymbol{e}_j). \end{aligned} \tag{35}$$

In the second term on the right-hand side of (35), removing the last mutation from lineage k leaves the lineage as a singleton in the data and a tree $(\mathcal{T}'_{k-}, \boldsymbol{n})$; in the third term, the lineage k with the mutation z_0 removed is identical to lineage j in the sample and removing the mutation leaves a tree $(\mathcal{T}''_{k-}, \boldsymbol{n} + \boldsymbol{e}_j)$, where the new multiplicity is adjusted. See Figure 3 for an illustration. An optimal proposal distribution for the gene-tree case can be determined by applying the method (ii) described in Section 3.3. Given the tree $(\mathcal{T}, \boldsymbol{n})$, the number of possible

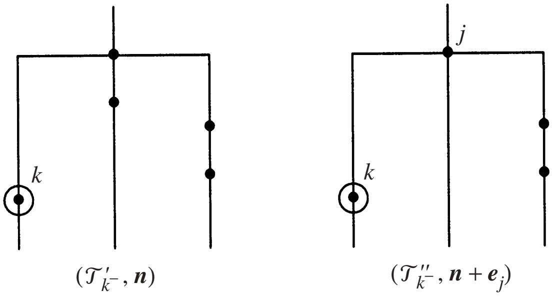
k
$(\mathcal{T}'_{k^-}, \boldsymbol{n})$
j
k
$(\mathcal{T}''_{k^-}, \boldsymbol{n}+\boldsymbol{e}_j)$

FIGURE 3: Removing a mutation.

TABLE 1: Proposal distribution and importance weights for gene trees.

H_{i-1}	Proposal distribution	Importance weights
$(\mathcal{T}, \boldsymbol{n}-\boldsymbol{e}_k)$	$\dfrac{n_k}{n^\circ}$	$\dfrac{n^\circ}{n_k}\dfrac{n_k-1}{n-1+\theta}$
$(\mathcal{T}_{k^-}, \boldsymbol{n})$	$\dfrac{1}{n^\circ}$	$\dfrac{n^\circ}{n}\dfrac{\theta}{n-1+\theta}$
$(\mathcal{T}''_{k^-}, \boldsymbol{n}+\boldsymbol{e}_j)$	$\dfrac{1}{n^\circ}$	$\dfrac{n^\circ}{n}\dfrac{(n_j+1)\theta}{n-1+\theta}$

events for the first step back in time is

$$n^\circ = \sum_{k:n_k\geq 2} n_k + \sum_{k:n_k=1,k^-} 1 + \sum_{\substack{k:n_k=1\\k\to j}} 1,$$

and (35) can be rewritten as

$$\begin{aligned}\frac{1}{n^\circ}\bigg(\sum_{k:n_k\geq 2} n_k + \sum_{k:n_k=1,k^-} 1 + \sum_{\substack{k:n_k=1\\k\to j}} 1\bigg)p(\mathcal{T}, \boldsymbol{n}) \\ = \frac{n-1}{n-1+\theta}\sum_{k:n_k\geq 2}\frac{n_k-1}{n-1}p(\mathcal{T}, \boldsymbol{n}-\boldsymbol{e}_k) + \frac{\theta}{n-1+\theta}\sum_{k:n_k=1}\frac{1}{n}p(\mathcal{T}'_{k^-}, \boldsymbol{n}) \\ + \frac{\theta}{n-1+\theta}\sum_{\substack{k:n_k=1\\k\to j}}\frac{n_j+1}{n}p(\mathcal{T}''_{k^-}, \boldsymbol{n}+\boldsymbol{e}_j).\end{aligned} \tag{36}$$

The history process $\{H_i\}$ is the set of subtree states that the ancestral configuration passes through until the most recent common ancestor is reached.

Proposition 3. *The IS proposal distribution $\hat{p}(H_{i-1} \mid H_i)$ with the respective importance weights is given in Table 1, where $H_i = (\mathcal{T}, \boldsymbol{n})$.*

Proof. Let $H_i = (\mathcal{T}, \boldsymbol{n})$. Consider the three following cases:

(i) $H_{i-1} = (\mathcal{T}, \boldsymbol{n}-\boldsymbol{e}_k)$;

(ii) $H_{i-1} = (\mathcal{T}'_{k^-}, \boldsymbol{n})$;

(iii) $H_{i-1} = (\mathcal{T}''_{k^-}, \boldsymbol{n} + \boldsymbol{e}_j)$.

To obtain an approximation $\hat{p}(H_{i-1} \mid H_i)$, equate, for each k, each term on the left-hand side of (36) with the corresponding term on the right-hand side. Recall that the proposal distribution is constructed from $\hat{p}(H_{i-1} \mid H_i) = p(H_i \mid H_{i-1})\hat{p}(H_{i-1})/\hat{p}(H_i)$, with importance weights $\hat{p}(H_i)/\hat{p}(H_{i-1})$. The three cases of H_{i-1} are treated separately.

(i) $H_{i-1} = (\mathcal{T}, \boldsymbol{n} - \boldsymbol{e}_k)$. For each k with $n_k > 1$, set

$$\frac{n_k}{n^\circ}\hat{p}(\mathcal{T}, \boldsymbol{n}) = \frac{n_k - 1}{n - 1 + \theta}\hat{p}(\mathcal{T}, \boldsymbol{n} - \boldsymbol{e}_k),$$

which implies that

$$\frac{\hat{p}(\mathcal{T}, \boldsymbol{n} - \boldsymbol{e}_k)}{\hat{p}(\mathcal{T}, \boldsymbol{n})} = \frac{n_k}{n^\circ}\frac{n - 1 + \theta}{n_k - 1}.$$

Then the proposal probability is

$$\hat{p}(H_{i-1} \mid H_i) = \frac{n_k - 1}{n - 1 + \theta}\frac{n_k}{n^\circ}\frac{n - 1 + \theta}{n_k - 1} = \frac{n_k}{n^\circ}.$$

(ii) $H_{i-1} = (\mathcal{T}_{k^-}, \boldsymbol{n})$. For each appropriate k with $n_k = 1$, set

$$\frac{1}{n^\circ}\hat{p}(\mathcal{T}, \boldsymbol{n}) = \frac{\theta}{n(n - 1 + \theta)}\hat{p}(\mathcal{T}'_{k^-}, \boldsymbol{n}),$$

which implies that

$$\frac{\hat{p}(\mathcal{T}'_{k^-}, \boldsymbol{n})}{\hat{p}(\mathcal{T}, \boldsymbol{n})} = \frac{1}{n^\circ}\frac{n(n - 1 + \theta)}{\theta}.$$

Then the proposal probability is

$$\hat{P}(H_{i-1} \mid H_i) = \frac{\theta}{n - 1 + \theta}\frac{1}{n}\frac{1}{n^\circ}\frac{n(n - 1 + \theta)}{\theta} = \frac{1}{n^\circ}.$$

(iii) $H_{i-1} = (\mathcal{T}''_{k^-}, \boldsymbol{n} + \boldsymbol{e}_j)$. For each appropriate pair k and j with $n_k = 1$, set

$$\frac{1}{n^\circ}\hat{p}(\mathcal{T}, \boldsymbol{n}) = \frac{\theta}{n - 1 + \theta}\frac{n_j + 1}{n}\hat{p}(\mathcal{T}''_{k^-}, \boldsymbol{n} + \boldsymbol{e}_j),$$

which implies that

$$\frac{\hat{p}(\mathcal{T}''_{k^-}, \boldsymbol{n} + \boldsymbol{e}_j)}{\hat{p}(\mathcal{T}, \boldsymbol{n})} = \frac{1}{n^\circ}\frac{n}{n_j + 1}\frac{n - 1 + \theta}{\theta}.$$

Then the proposal probability is

$$\hat{p}(H_{i-1} \mid H_i) = \frac{\theta}{n - 1 + \theta}\frac{n_j + 1}{n}\frac{1}{n^\circ}\frac{n}{n_j + 1}\frac{n - 1 + \theta}{\theta} = \frac{1}{n^\circ}.$$

The proposal distribution given in Table 1 coincides with the one suggested by Stephens and Donnelly (2000). It is uniform on possible history changes for genes back in time. That is, choose a gene at random from those which may have been involved in a history change, then either coalesce or remove a mutation. Once a gene is chosen, the event choice is unique.

5. Discussion

This paper sets up a framework for constructing IS proposal distributions for likelihood calculation of sample probabilities in population-genetics models where the distribution of gene frequencies in the population is described by a diffusion-process generator. The method is of interest because it gives a general way of constructing proposal distributions from a diffusion-process generator. The proposal distributions reduce to those of Stephens and Donnelly (2000) in the case of a single population, but the technique is much more general.

Two companion papers are De Iorio and Griffiths (2004), which applies the methods of this paper to subdivided population models, and De Iorio *et al.* (2004), which considers a stepwise-mutation model in a subdivided population in detail.

Although the IS proposal distributions are derived for a finite number of types, it may be possible to extend them to cases with an infinite number of types, depending on the model. The infinitely-many-alleles model and the stepwise-mutation model are two such models, considered in the companion papers.

References

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

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

Beaumont, M. (2001). Conservation genetics. In *Handbook of Statistical Genetics*, eds D. J. Balding, M. Bishop and C. Cannings, John Wiley, Chichester, pp. 779–809.

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.

Carbone, I. and Kohn, M. (2001). A microbial population–species interface: nested cladistic and coalescent inference with multilocus data. *Molecular Ecology* **10,** 947–964.

De Iorio, M. and Griffiths, R. C. (2004). Importance sampling on coalescent histories. II: Subdivided population models. *Adv. Appl. Prob.* **36,** 434–454.

De Iorio, M., Griffiths, R. C., Leblois, R. and Rousset, F. (2004). Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. Tech. Rep., Oxford University.

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

Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes. Characterization and Convergence*. John Wiley, New York.

Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetics data. *Genetics* **159,** 1299–1318.

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. In *Statistics in Molecular Biology and Genetics* (IMS Lecture Notes Monogr. Ser. **33**), Institute of Mathematical Statistics, Hayward, CA, pp. 163–185.

Griffiths, R. C. (1989). Genealogical-tree probabilities in the infinitely-many-sites model. *J. Math. Biol.* **27,** 667–680.

Griffiths, R. C. (2001). Ancestral inference from gene trees. In *Genes, Fossils, and Behaviour: An Integrated Approach to Human Evolution* (NATO Sci. Ser. A Life Sci. **310**), eds P. Donnelly and R. Foley, IOS Press, Amsterdam, pp. 137–172.

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). Ancestral inference in population genetics. *Statist. Sci.* **9,** 307–319.

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

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

Griffiths, R. C. and Tavaré, S. (1996). Markov chain inference methods in population genetics. *Math. Comput. Modelling* **23,** 141–158.

Griffiths, R. C. and Tavaré, S. (1997). Computational methods for the coalescent. In *Progress in Population Genetics and Human Evolution* (IMA Vols Math. Appl. **87**), eds P. Donnelly and S. Tavaré, Springer, Berlin, pp. 165–182.

Griffiths, R. C. and Tavaré, S. (1999). The ages of mutations in gene trees. *Ann. Appl. Prob.* **9,** 567–590.

Harding, R. M. *et al.* (1997). Archaic African *and* Asian lineages in the genetic ancestry of modern humans. *Amer. J. Human Genet.* **60,** 772–789.

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

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.

Kuhner, M. K., Yamato, J. and Felsenstein, J. (1997). Appliecations of Metropolis–Hastings genealogy sampling. In *Progress in Population Genetics and Human Evolution* (IMA Vols Math. Appl. **87**), eds P. Donnelly and S. Tavaré, Springer, Berlin, pp. 257–270.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

Markovtsova, L., Marjoram, P. and Tavaré, S. (2000a). The age of a unique event polymorphism. *Genetics* **156,** 401–409.

Markovtsova, L., Marjoram, P. and Tavaré, S. (2000b). The effects of rate variation on ancestral inference in the coalescent. *Genetics* **156,** 1427–1436.

Nath, M. and Griffiths, R. C. (1996). Estimation in an island model using simulation. *Theoret. Pop. Biol.* **3,** 227–253.

Nielsen, R. (1997). A likelihood approach to population samples of microsatellite alleles. *Genetics* **146,** 711–716.

Slade, P. (2000a). Simulation of selected genealogies. *Theoret. Pop. Biol.* **57,** 35–49.

Slade, P. (2000b). Most recent common ancestor probability distribution in gene genealogies under selection. *Theoret. Pop. Biol.* **58,** 291–305.

Stephens, M. (2001). Inference under the coalescent. In *Handbook of Statistical Genetics*, eds D. J. Balding, M. Bishop and C. Cannings, John Wiley, Chichester, pp. 213–238.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *J. R. Statist. Soc. B* **62,** 605–655.

Wakeley, J., Nielsen, R., Liu-Cordero, S. and Ardlie, K. (2001). The discovery of single nucleotide polymorphisms, and inferences about human demographic history. *Amer. J. Human Genet.* **69,** 1332–1347.

Wilson, I. J. and Balding, D. J. (1998). Genealogical inference from microsatellite data. *Genetics* **150,** 499–510.

Wilson, I. J., Weale, M. E. and Balding, D. J. (2003). Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. R. Statist. Soc. A* **166,** 155–201.