ELSEVIER

Unrooted Genealogical Tree Probabilities in the Infinitely-Many-Sites Model

R. C. GRIFFITHS
Department of Mathematics, Monash University, Clayton 3168, Australia

AND

SIMON TAVARÉ
Department of Mathematics and Biological Sciences, University of Southern California, Los Angeles, California 90089-1113

Received 2 February 1994; revised 11 July 1994

ABSTRACT

The infinitely-many-sites process is often used to model the sequence variability observed in samples of DNA sequences. Despite its popularity, the sampling theory of the process is rather poorly understood. We describe the tree structure underlying the model and show how this may be used to compute the probability of a sample of sequences. We show how to produce the unrooted genealogy from a set of sites in which the ancestral labeling is unknown and from this the corresponding rooted genealogies. We derive recursions for the probability of the configuration of sequences (equivalently, of trees) in both the rooted and unrooted cases. We give a computational method based on Monte Carlo recursion that provides approximants to sampling probabilities for samples of any size. Among several applications, this algorithm may be used to find maximum likelihood estimators of the substitution rate, both when the ancestral labeling of sites is known and when it is unknown.

1. INTRODUCTION

The infinitely-many-sites model is extremely popular in the population genetics literature as a description of the DNA sequence variability observed in samples of genes. Hudson [8] provides an excellent overview. For the most part, inference for this process has been based on the distribution of summary statistics of the sequences. One such is the number of segregating sites, which is the number of DNA sequence locations in the aligned sequences in which more than one nucleotide is represented. While this statistic is simple, it clearly does not make full use of the data. Strobeck [10] showed that certain genealogical trees play an important role in defining the probability distribution of the sample configuration. He showed how such probabilities could be calculated for samples that contained at most three distinct haplotypes.

MATHEMATICAL BIOSCIENCES 127:77–98 (1995)

655 Avenue of the Americas, New York, NY 10010
0025-5564/95/$9.50
SSDI 0025-5564(94)00044-Z

In this paper, we develop the theory of the genealogical trees that underpin the model and show how this theory can be used to calculate probabilities for samples of any size. One novel application of the theory is a way to compute maximum likelihood estimators of the substitution rate.

From a practical point of view, there are two important new aspects to our results. The first shows how sample configuration probabilities can be found when the ancestral labeling of the sites is *unknown*. The second is the development of a computationally feasible numerical algorithm for calculating such probabilities. This algorithm is based on the principles described by Griffiths and Tavaré [6], in which the probability of interest is represented as the mean of a functional of a particular Markov chain that can readily be simulated.

1.1. *THE COALESCENT*

The genealogy of a sample of n genes drawn at random from a large Wright–Fisher population of approximately constant size N individuals is often described by the coalescent [9]. The time T_j during which the sample has j distinct ancestors has an exponential distribution with mean $\mathbb{E}T_j = 2/(j(j-1))$, $j = 2,3,\ldots,n$, times for different j being independent. Time is conventionally measured in units of $2N$ generations. For reviews of the basic structure of the coalescent, see [8, 11], for example.

The coalescent can be thought of as the ancestral tree of the sample, showing how individuals are related back to their common ancestor. In the time scaling used here this tree is binary. Mutations are superimposed on the ancestral tree as follows. Suppose there is a probability of mutation of u per gene per generation, and set $\theta = 4Nu$. Conditional on the tree, put down mutations according to Poisson processes of rate $\theta/2$, independently for each branch in the tree. A typical sample path of a coalescent relating individuals in a sample of size $n = 7$, together with the mutations occurring in the tree, is given in Figure 1.

The coalescent tree with mutations can be condensed into a genealogical tree with no time scale by labeling each gene by a sequence of mutations up to the common ancestor. For the example in Figure 1, the sequences may be represented as follows:

gene 1	(9,7,3,1,0)
gene 2	(3,1,0)
gene 3	(11,6,4,1,0)
gene 4	(8,6,4,1,0)
gene 5	(8,6,4,1,0)
gene 6	(8,6,4,1,0)
gene 7	(10,5,2,0).

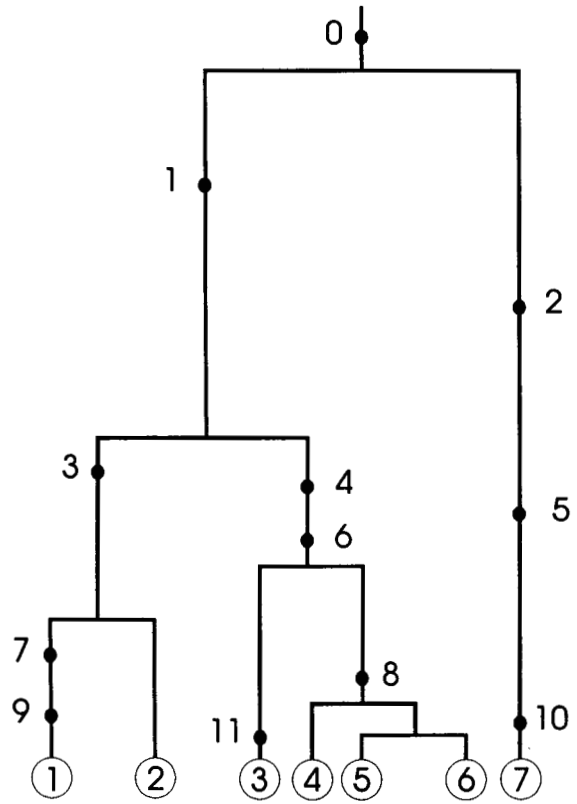

FIG. 1. Coalescent tree with mutations. ● denotes mutation, ○ denotes individuals.

The 0's in each sequence are used to indicate that the the sequences can be traced back to a common ancestor. The condensed genealogical tree is shown in Figure 2. The leaves in condensed tree correspond to the genes in the sample and the branches in the tree are the internal links between different mutations.

1.2. *TREES IN THE INFINITELY-MANY-SITES MODELS*

There are many different models in the literature that describe the effects of mutation on the type of each gene. In this paper, we focus on the infinitely-many-sites model of Watterson [12]. Under this model, a gene can be though of as an infinite sequence of completely linked sites, each labeled 0 or 1. A 0 denotes the ancestral (original) type, and a 1 a mutant type. The mutation mechanism is such that a mutant offspring gets a mutation at a single new site that has never before seen a

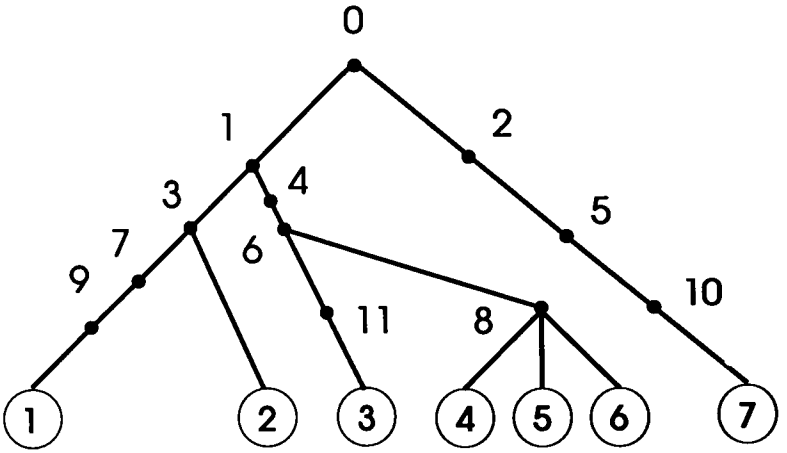

FIG. 2. Genealogical tree corresponding to Figure 1.

mutation. This changes the 0 to a 1 at that site and introduces another segregating site into the sample. By way of example, the segregating sites in a sample of 7 genes might have the following structure:

gene 1	1	0	1	0	0	0	1	0	1	0	0
gene 2	1	0	1	0	0	0	0	0	0	0	0
gene 3	1	0	0	1	0	1	0	0	0	0	1
gene 4	1	0	0	1	0	1	0	1	0	0	0
gene 5	1	0	0	1	0	1	0	1	0	0	0
gene 6	1	0	0	1	0	1	0	1	0	0	0
gene 7	0	1	0	0	1	0	0	0	0	1	0.

Each set of sequences from an infinitely-many-sites model corresponds to a *rooted* genealogical tree. The number of segregating sites is precisely the number of mutations in the tree. For example, the sequences listed above are equivalent to those in Figures 1 and 2. There are several algorithms for producing the rooted tree from a set of binary sequences. We use Griffiths [3]; see also Gusfield [7]. These trees are essentially the same as those arising from compatible binary characters in the systematics literature. Felsenstein [2] provides a review of this aspect.

If a set of sequences is from the infinitely-many-sites model, but the ancestral base at each site is unknown, then there is an *unrooted* genealogical tree that corresponds to the sequences. In these unrooted

trees, the vertices represent sequences and the number of mutations between sequences are represented by numbers along the edges.

Given a single rooted tree, the unrooted genealogy and therefore all other rooted trees, corresponding to other possible rootings, can be found. The constructive way to do this from a given rooted tree is to put in vertices corresponding to sequences, and also inferred sequences at each vertex of degree greater than two. A mutation at a vertex is counted in the unrooted tree edge leading toward the root in the given rooted tree. If there is no root type in the sample and the root is of degree two, then it is removed from the unrooted genealogy (as in our example). The unrooted tree constructed from any of these rooted trees is unique. It is convenient to label the vertices as to the genes they represent. The unrooted tree for the example sequences is shown in Figure 3.

Conversely, the class of rooted trees produced from an unrooted genealogy may be constructed by placing the root at one of the sequences or between mutations along an edge. Two examples are given in Figure 4. In the first the root corresponds to the third sequence, labeled 3 in Figures 1 and 3, and in the second it is between the two mutations between the two inferred sequences.

Trees are called *labeled* if the sequences are labeled. Two labeled trees are identical if there is a renumbering of the sites so that they are identical. An *ordered labeled* tree is one where the sequences are

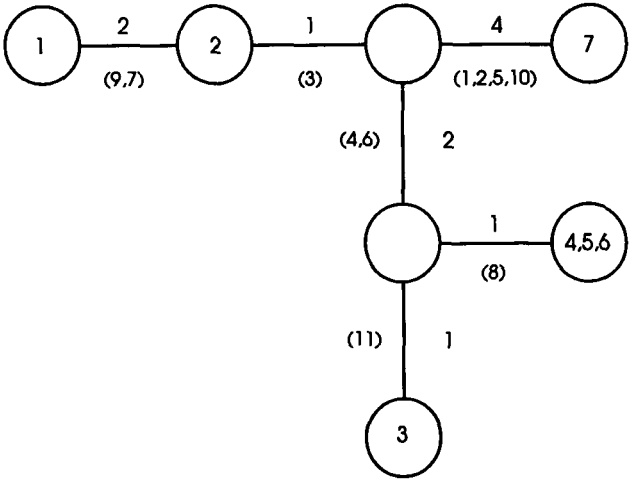

FIG. 3. Unrooted genealogical tree corresponding to Figure 1. The unlabeled circles are inferred individuals.

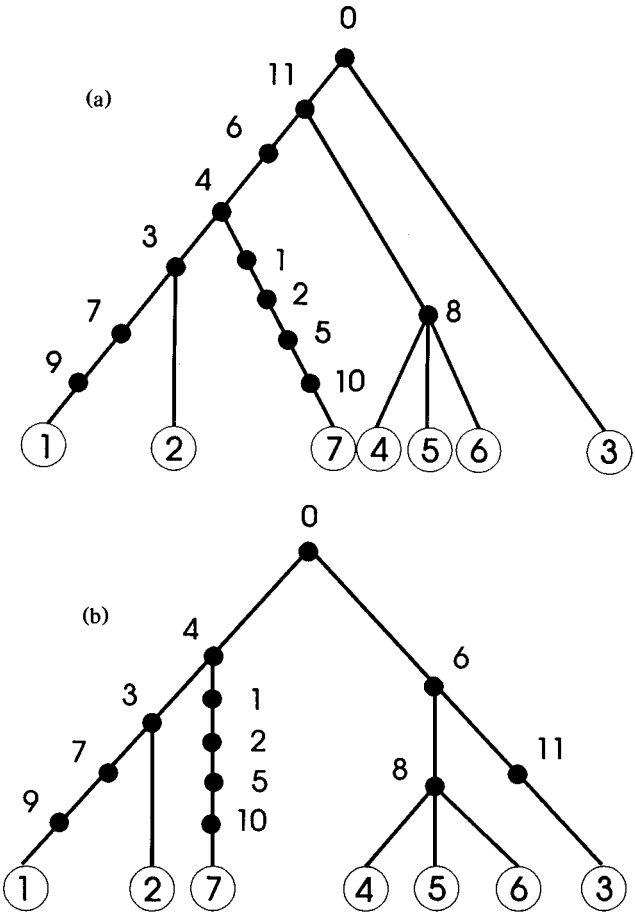

FIG. 4. Moving the root. (a) Tree with root the third sequence; (b) Tree with root between mutations.

labeled and considered to be in a particular order; visually this corresponds to a tree diagram with ordered leaves. An *unlabeled* (and so unordered) tree is a tree where the sequences are not labeled. Visually two unlabeled trees are identical if they can be drawn identically by rearranging the leaves and corresponding paths in one of the trees. (The sites of an unlabeled tree are, however, still labeled.)

Usually trees are unlabeled, with sequences and sites then labeled for convenience. However, it is easiest to deal with ordered labeled trees in a combinatorial and probabilistic sense and then deduce results

about unlabeled trees from labeled trees. If there are α sequences (including the inferred sequences) with $m_1, m_2, \ldots$ mutations along the edges and s segregating sites, then there are

$$\alpha + \sum_j (m_j - 1) = s + 1 \tag{1}$$

rooted trees when the sequences are labeled. There may be fewer unlabeled rooted trees, as some can be identical after unlabeling the sequences. In the example there are 11 segregating sites, and so 12 labeled rooted trees, which correspond to distinct unlabeled rooted trees as well.

The class of rooted trees corresponds to those constructed from toggling the ancestor labels 0 and 1 at sites. The number of the 2^s possible relabelings that are consistent with the sequences having come from a tree is

$$\alpha + \sum_j \sum_{k=1}^{m_j - 1} \binom{m_j}{k} = \alpha + \sum_j (2^{m_j} - 2). \tag{2}$$

This follows from the observation that if there is a collection of m segregating sites which correspond to mutations between sequences, then the corresponding data columns of the 0-1 sequences (with 0 the ancestral state) are identical or complementary. Any of the $\binom{m}{k}$ configurations of k identical and $m-k$ complementary columns correspond to the same labeled tree with a root placed after the kth mutation. The correspondence between different rooted labeled trees and the matrix of segregating sites can be described as follows: in order to move the root from one position to another, just toggle those sites that occur on the branches between the two roots.

Ethier and Griffiths [1] provide a simulation algorithm to generate the rooted genealogical tree of a sample of size n from the infinitely-many-sites model, starting from a common ancestor. The simulation proceeds as follows:

(1) Start with a tree with two leaves.
(2) At successive time instants where there are r leaves, either
  (i) grow a leaf (with probability $(r-1)/(r+\theta-1)$); or
  (ii) grow a branch (with probability $\theta/(r+\theta-1)$). (3)
(3) Stop when there are $n+1$ leaves in the tree, and delete the last leaf.

It is possible to simulate an unrooted tree using the same algorithm, by finding the unrooted tree that corresponds to the simulated rooted tree.

1.3. *INFERENCE ABOUT* θ

Watterson [12] was interested in estimating the parameter θ using the number K of segregating sites in the sample. Since the mean number of segregating sites in a sample of size n is $\theta\sum_{j=1}^{n-1} j^{-1}$, he proposed the unbiased estimator

$$\tilde{\theta} = K\left[\sum_{j=1}^{n-1} \frac{1}{j}\right]^{-1}. \tag{4}$$

This estimator does not make full use of the data since it ignores the genealogical relationships among the genes in the sample. It might be expected to be less efficient than an estimator that does. A start on the theory of likelihood methods for unlabeled genealogical trees was made by Strobeck [10]. In this paper, we extend the exact recursive method of Griffiths [4] to unrooted trees, and we develop a Markov chain Monte Carlo method for computing probabilities of either rooted or unrooted trees. This technique complements the recursive method and allows such probabilities to be computed for arbitrary data sets, which may have both large numbers of individuals and many segregating sites.

2. GENEALOGICAL TREE PROBABILITIES

The type of a gene i in the sample is described by a sequence $\mathbf{y}_i = (y_{i0}, y_{i1}, \ldots)$ of positive integers. Suppose that in a sample of n genes there are d distinct sequences, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d$, grouped according to their multiplicities $\mathbf{n} = (n_1, \ldots, n_d)$. We write the data in the form $(\mathbf{x}_1, \ldots, \mathbf{x}_d, \mathbf{n}) \equiv (T, \mathbf{n})$, where T represents a rooted tree and $\mathbf{n}$ the multiplicities of the leaves. In the example discussed earlier, there are $d = 5$ distinct sequences that can be represented as

$$\mathbf{x}_1 = (9,7,3,1,0)$$
$$\mathbf{x}_2 = (3,1,0)$$
$$\mathbf{x}_3 = (11,6,4,1,0)$$
$$\mathbf{x}_4 = (8,6,4,1,0)$$
$$\mathbf{x}_5 = (10,5,2,0).$$

Note that the label 0 is common to all sequences because we are assuming that the individuals in the sample can be traced back to a common ancestor.

Let $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ be collections of sequences, not necessarily distinct. Define equivalence relations $\sim$ and $\approx$ on trees by

$(\mathbf{x}_1,\ldots,\mathbf{x}_n)\sim(\mathbf{y}_1,\ldots,\mathbf{y}_n)$ if there exists a bijection $\eta\colon \mathbb{Z}_+\to\mathbb{Z}_+\equiv\{0,1,2,\ldots\}$ with $y_{ij}=\eta(x_{ij})$ for $i=1,\ldots,n$ and $j=0,1,\ldots$ and $(\mathbf{x}_1,\ldots,\mathbf{x}_n)\approx(\mathbf{y}_1,\ldots,\mathbf{y}_n)$ if there exists a bijection $\eta\colon \mathbb{Z}_+\to\mathbb{Z}_+$ and a permutation $\sigma\in S_n$, the set of permutations of $(1,\ldots,n)$, such that $y_{\sigma(i)j}=\eta(x_{ij})$ for $i=1,\ldots,n$ and $j=0,1,\ldots$. Equivalence classes under $\sim$ correspond to labeled trees, and those under $\approx$ to unlabeled trees.

Let $p(T,\mathbf{n})$ be the probability, under the stationary distribution of the process, of obtaining a particular ordered sample of distinct sequences $T=(\mathbf{x}_1,\ldots,\mathbf{x}_d)$ with multiplicities $\mathbf{n}=(n_1,\ldots,n_d)$ in a sample of size n under equivalence $\sim$. Ethier and Griffiths [1] prove that $p(T,\mathbf{n})$ satisfies the equation

$$n(n-1+\theta)p(T,\mathbf{n})=\sum_{k:n_k\geq 2} n_k(n_k-1)p(T,\mathbf{n}-\mathbf{e}_k)$$
$$+\theta\sum_{\substack{k:n_k=1,\,x_{k0}\text{ distinct}\\ \mathscr{S}\mathbf{x}_k\neq\mathbf{x}_j\ \forall j}} p(\mathscr{S}_kT,\mathbf{n})$$
$$+\theta\sum_{\substack{k:n_k=1\\ x_{k0}\text{ distinct}}}\ \sum_{j:\mathscr{S}\mathbf{x}_k=\mathbf{x}_j} p\big(\mathscr{R}_kT,\mathscr{R}_k(\mathbf{n}+\mathbf{e}_j)\big). \quad (5)$$

In (5), $\mathbf{e}_j$ is the jth unit vector, $\mathscr{S}$ is a shift operator which deletes the first coordinate of a sequence, $\mathscr{S}_kT$ deletes the first coordinate of the kth sequence of T, $\mathscr{R}_kT$ removes the kth sequence of T, and "x_{k0} distinct" means that $x_{k0}\neq x_{ij}$ for all $(\mathbf{x}_1,\ldots,\mathbf{x}_d)$ and $(i,j)\neq(k,0)$. In the inner summation in the last term of (5) there is at most one j such that $\mathscr{S}\mathbf{x}_k=\mathbf{x}_j$. The boundary condition is $p(T_1,\mathbf{e}_1)=1$. The system (5) is recursive in the quantity $\{n-1+\text{the number of vertices in } T\}$. Equation (5) can be validated by a simple coalescent argument by looking backwards in time for the first event in the ancestry of the sample. The first term on the right of (5) corresponds to a coalescence occurring first, while the last two terms correspond to mutations. There are two mutation terms, the first corresponding to the ancestor not being in the sample, the second to the ancestor being in the sample.

For the purposes of an algorithm for computing these probabilities, it is more convenient to consider the recursion satisfied by the quantities $p^0(T,\mathbf{n})$ defined by

$$p^0(T,\mathbf{n})=\frac{n!}{n_1!\cdots n_d!}p(T,\mathbf{n}). \quad (6)$$

$p^0(T,\mathbf{n})$ is the probability of the labeled tree T without regard to the order of the sequences in the sample. Using (5), this may be written in

the form

$$
\begin{aligned}
&n(n-1+\theta)p^0(T,\mathbf{n}) \\
&\quad = \sum_{k:n_k \geqslant 2} n(n_k-1)p^0(T,\mathbf{n}-\mathbf{e}_k) \\
&\qquad + \theta \sum_{\substack{k:n_k=1,x_{k0}\text{ distinct} \\ \mathscr{S}\mathbf{x}_k \neq \mathbf{x}_j\ \forall_j}} p^0(\mathscr{S}_k T,\mathbf{n}) \\
&\qquad + \theta \sum_{\substack{k:n_k=1 \\ x_{k0}\text{ distinct}}} \sum_{j:\mathscr{S}\mathbf{x}_k=\mathbf{x}_j} (n_j+1)p^0\big(\mathscr{R}_k T,\mathscr{R}_k(\mathbf{n}+\mathbf{e}_j)\big). \qquad (7)
\end{aligned}
$$

Let $p^*(T,\mathbf{n})$ be the probability of a corresponding unlabeled tree with multiplicity of the sequences given by $\mathbf{n}$ under equivalence $\approx$. p^* is related to p^0 by a combinatorial factor as follows. Define $T_\sigma = (\mathbf{x}_{\sigma(1)},\ldots,\mathbf{x}_{\sigma(d)})$, and $\mathbf{n}_\sigma = (n_{\sigma(1)},\ldots,n_{\sigma(d)})$ for σ in S_d. Letting

$$
a(T,\mathbf{n}) = \big|\{\sigma \in S_d : T_\sigma = T, \mathbf{n}_\sigma = \mathbf{n}\}\big|, \qquad (8)
$$

we have

$$
p^*(T,\mathbf{n}) = \frac{1}{a(T,\mathbf{n})} p^0(T,\mathbf{n}). \qquad (9)
$$

The number of distinct ordered labeled trees corresponding to the unlabeled tree is

$$
\frac{n!}{n_1!\cdots n_d! a(T,\mathbf{n})}.
$$

In the tree shown in Figure 2, $a(T,\mathbf{n}) = 1$. A subsample of three genes (9,7,3,1,0), (11,6,4,1,0), (10,5,2,0), forming a tree T' with frequencies $\mathbf{n}' = (1,1,1)$, has $a(T',\mathbf{n}') = 2$ because the first two sequences are equivalent in an unlabeled tree.

2.1. *RECURSIONS FOR UNROOTED TREES*

Ethier and Griffiths [1] consider the case of rooted trees, which we extend to cover the unrooted case. A labeled unrooted genealogical tree of a sample of sequences has a vertex set V which corresponds to the labels of the sample sequences and any inferred sequences in the tree. Let $\mathbf{Q}$ be the edges of the tree, described by $(m_{ij}, i,j \in V)$, where m_{ij} is the number of mutations between vertices i and j. Let $\mathbf{n}$ denote the multiplicities of the sequences. It is convenient to include the inferred

sequences $l \in V$ with $n_l = 0$. Then the unrooted genealogy is described by $(\mathbf{Q}, \mathbf{n})$.

Define equivalence relations $\sim_u$ and $\approx_u$ on trees, expressing equivalence of having the same number of segregating sites between pairs of sequences. For a tree $T = (\mathbf{x}_1, \ldots, \mathbf{x}_d)$ define a $d \times d$ matrix $C(T)$ by

$$c_{ab}(T) = \left|\{x_{ai} \in \mathbf{x}_a, x_{ai} \notin \mathbf{x}_b\} \cup \{x_{ai} \notin \mathbf{x}_a, x_{ai} \in \mathbf{x}_b\}\right|, \qquad a, b \in \{1, \ldots, d\}.$$

$c_{ab}(T)$ is the number of segregating sites between $\mathbf{x}_a$ and $\mathbf{x}_b$. For two trees T_1 and T_2 (with possibly nondistinct sequences), $T_1 \sim_u T_2$ if $C(T_1) = C(T_2)$, and $T_1 \approx_u T_2$ if there exists $\sigma \in S_n$ such that $C(T_{1\sigma}) = C(T_2)$. Unrooted trees are thus considered equivalent if they have the same matrix of numbers of differences between the sequences at their vertices.

Define $p(\mathbf{Q}, \mathbf{n})$, $p^0(\mathbf{Q}, \mathbf{n})$, $p^*(\mathbf{Q}, \mathbf{n})$ analogously to the probabilities for $(T, \mathbf{n})$ where in the first and third probabilities $(\mathbf{Q}, \mathbf{n})$ is a representative of equivalence classes under $\sim_u$ and $\approx_u$. Suppose that $\mathscr{C}$ is an equivalence class of trees under $\sim_u$ with pairwise segregating site matrix C_0. Then $\mathscr{C}$ is the union of the equivalence classes of trees T under $\sim$ having $C(T) = C_0$. The probability of a labeled unrooted genealogical tree $\mathbf{Q}$ is thus

$$p(\mathbf{Q}, \mathbf{n}) = \sum_{T: C(T) = C_0} p(T, \mathbf{n}), \tag{10}$$

where the trees T in the summation are distinct under $\sim$. If $\mathbf{Q}$ has a total of s mutations, there are $s+1$ terms in the sum in (10). The same relationship holds in (10) if p is replaced by p^0, and q by q^0. The combinatorial factor relating $p^*(\mathbf{Q}, \mathbf{n})$ and $p^0(\mathbf{Q}, \mathbf{n})$ is

$$a(\mathbf{Q}, \mathbf{n}) = \left|\{\sigma \in S_{|V|} : \mathbf{Q}_\sigma = \mathbf{Q}, \mathbf{n}_\sigma = \mathbf{n}\}\right|. \tag{11}$$

The quantities $p(\mathbf{Q}, \mathbf{n})$ and $p^0(\mathbf{Q}, \mathbf{n})$ satisfy recursions similar to (5) and (7). The recursion for $p(\mathbf{Q}, \mathbf{n})$ is

$$\begin{aligned} n(n-1+\theta)p(\mathbf{Q}, \mathbf{n}) = & \sum_{k: n_k \geq 2} n_k(n_k - 1)p(\mathbf{Q}, \mathbf{n} - \mathbf{e}_k) \\ & + \theta \sum_{\substack{k: n_k = 1, |k| = 1 \\ k \to j, m_{kj} > 1}} p(\mathbf{Q} - \mathbf{e}_{kj}, \mathbf{n}) \\ & + \theta \sum_{\substack{k: n_k = 1, |k| = 1 \\ k \to j, m_{kj} = 1}} p(\mathbf{Q} - \mathbf{e}_{kj}, \mathbf{n} + \mathbf{e}_j - \mathbf{e}_k), \end{aligned} \tag{12}$$

where $|k|=1$ means that the degree of the vertex k is 1 (that is, k is a leaf), and $k \to j$ means that vertex k is joined to vertex j. In the last term on the right of (12), vertex k is removed from $\mathbf{Q}$. The boundary conditions in (12) for $n=2$ are

$$p((0),2\mathbf{e}_1)=\frac{1}{1+\theta},$$

and

$$p((m),\mathbf{e}_1+\mathbf{e}_2)=\left(\frac{\theta}{1+\theta}\right)^m\frac{1}{1+\theta}, m=1,2,\ldots.$$

To see why (12) is true, consider the possible rooted trees in (10) constructed by placing a root in the unrooted tree and sum their probabilities using (5). The first term on the right of (12) is clearly correct. For the second term note that a tree $\mathscr{S}_k T$ is a valid tree contributing to $\mathbf{Q}-\mathbf{e}_k$ for all trees T except when the root is at the tip of the sequence $\mathbf{x}_k$, and that the union of the class of trees $\mathscr{S}_k T$ over different positions for the root produces $\mathbf{Q}-\mathbf{e}_k$. A similar argument gives the third term.

Strobeck [10] derived equations for the probabilities of unrooted genealogies for two and three distinct haplotypes. His equations for these particular cases correspond to (12), with the factor (11) included in the recursions. There appear to be corrections needed to his Equations (3) and (4). The last two terms in his Equation (3) have multipliers $\delta(n_2+1,n_3)+1$ and $\delta(n_1,n_2+1)+1$ missing, and a divisor of n is missing from the last three terms in his Equation (4).

2.2. *RECURSIONS FOR SEQUENCES*

The recursions for the probabilities of rooted or unrooted trees can be translated directly into recursions for the probabilities of samples of sequences. Let $\mathbf{S}$ be the distinct binary sequences of the segregating sites in a sample of sequences, the rows of $\mathbf{S}$ representing distinct alleles, and let $\mathbf{n}$ be their multiplicities. If 0 represents an ancestral site, then

$$\begin{aligned} n(n-1+\theta)p(\mathbf{S},\mathbf{n}) = & \sum_{k:n_k\geqslant 2} n_k(n_k-1)p(\mathbf{S},\mathbf{n}-\mathbf{e}_k) \\ & +\theta \sum_{\substack{k:n_k=1,\mathbf{s}_{\cdot l}=\mathbf{e}_k \\ \mathbf{s}^l_{k\cdot}\neq\mathbf{s}^l_{j\cdot},k\neq j}} p(\mathbf{S}^l,\mathbf{n}) \\ & +\theta \sum_{\substack{k:n_k=1,\mathbf{s}_{\cdot l}=\mathbf{e}_k \\ \mathbf{s}^l_{k\cdot}=\mathbf{s}^l_{j\cdot}}} p\left(\mathbf{S}^{kl},\mathbf{n}^k+\mathbf{e}_j\right). \end{aligned} \tag{13}$$

In (13), $\mathbf{s}_{i\cdot}$ and $\mathbf{s}_{\cdot l}$ denote the ith row and the transpose of the lth column of $\mathbf{S}$. A superscript denotes removal of that column, and a double superscript notation $\mathbf{S}^{kl}$ indicates removal of row k and column l. For each k the second summation on the right includes just one term where $\mathbf{s}_{\cdot l} = \mathbf{e}_k$ though there may be multiple columns $\mathbf{e}_k$ in $\mathbf{S}$. The boundary condition in (13) is $p(\Phi,(1)) = 1$, where Φ is an empty matrix which may occur on the right side of (13) when $n = 2$.

The combinatorial factor relating $p^*(\mathbf{S},\mathbf{n})$ and $p^0(\mathbf{S},\mathbf{n})$ is

$$a(\mathbf{S},\mathbf{n}) = \left|\left\{\sigma \in S_d : \mathbf{n}_\sigma = \mathbf{n}, \exists \quad \tau \in S_s \text{ such that } \left(\mathbf{S}_{\sigma(i),\tau(j)}\right) = \mathbf{S}\right\}\right|. \quad (14)$$

The factors in (8) and (14) are identical.

In the case where the ancestral state is unknown, $\mathbf{S}$ is unique only up to arbitrary labels 0 and 1 for the two types in the columns, and

$$\begin{aligned} n(n-1+\theta)p(\mathbf{S},\mathbf{n}) = & \sum_{k:n_k \geq 2} n_k(n_k-1)p(\mathbf{S},\mathbf{n}-\mathbf{e}_k) \\ & + \theta \sum_{\substack{k:n_k=1, s_{\cdot l}=\mathbf{e}_k \\ \text{or } \mathbf{e}_k^c, \mathbf{s}^l_{k\cdot} \neq \mathbf{s}^l_{j\cdot}, k \neq j}} p(\mathbf{S}^l,\mathbf{n}) \\ & + \theta \sum_{\substack{k:n_k=1, \mathbf{s}_{\cdot l}=\mathbf{e}_k \\ \text{or } \mathbf{e}_k^c, \mathbf{s}^l_{k\cdot} = \mathbf{s}^l_{j\cdot}}} p\left(\mathbf{S}^{kl}, \mathbf{n}^k + \mathbf{e}_j\right), \end{aligned} \quad (15)$$

where $\mathbf{e}_k^c$ is the complement of the kth unit vector. Again, for each k the second summation on the right includes just one term where $\mathbf{s}_{\cdot l} = \mathbf{e}_k$ or $\mathbf{e}_k^c$ though there may be multiple columns $\mathbf{e}_k$ and $\mathbf{e}_k^c$ in $\mathbf{S}$. The boundary conditions in (15) for $n = 2$ are

$$p(\Phi,(2)) = \frac{1}{1+\theta},$$

and

$$p(\mathbf{S}_m,(1,1)) = \left(\frac{\theta}{1+\theta}\right)^m \frac{1}{1+\theta}, m = 1,\ldots,$$

where $\mathbf{S}_m$ represents m segregating sites between two sequences. The combinatorial factor relating $p^*(\mathbf{S},\mathbf{n})$ and $p^0(\mathbf{S},\mathbf{n})$ is

$$\begin{aligned} a(\mathbf{S},\mathbf{n}) = \Big|\{\sigma \in S_d : \mathbf{n}_\sigma = \mathbf{n}, \exists \tau \in S_s, \kappa \in G_s, \\ \text{such that } \left(\kappa(\mathbf{S})_{\sigma(i),\tau(j)}\right) = \mathbf{S}\}\Big|, \end{aligned} \quad (16)$$

where κ is one of the 2^s functions in G_s which map $d \times s\, 0-1$ matrices to the same class of matrices by toggling 0 and 1 entries in columns. The factors in (16) and (11) are identical.

Equation (13) and (15) have the benefit that they do not require a knowledge of the genealogical tree structure of a sample, and are possibly easier to code than the corresponding tree recursions. However whatever inference can be drawn about the genealogy of a sample is of great interest in practice.

2.3. *A NUMERICAL EXAMPLE*

In this example we suppose that the ancestral states are unknown and that the sequences, each with multiplicity unity, are:

$$\begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{matrix}$$

For convenience, label the segregating sites 1, 2, 3, and 4 starting from the left. When 0 is the ancestral state, a possible rooted tree for these sequences has paths to the root of (1,0), (2,3,0), and (4,0). It is then straightforward to construct the corresponding unrooted genealogy, which is shown in Figure 5. The central sequence is inferred.

There are five possible labeled rooted trees constructed from the unrooted genealogy, corresponding to the root being at one of the sequences, or between the two mutations on the edge between the inferred individual and gene 3. These five trees are shown in Figure 6, together with their probabilities $p(T,\mathbf{n})$, computed exactly from the recursion (5) when $\theta = 2.0$. $p(\mathbf{Q},\mathbf{n})$ is the sum of these probabilities, 0.00497256. The factor in (11) is 2, and the multinomial coefficient $3!/1!1!1! = 6$ so $p^*(\mathbf{Q},\mathbf{n}) = 3 \times 0.00497256 = 0.0149177$. Note that the trees (b) and (e) are identical unlabeled rooted trees, but are distinct labeled rooted trees, so are both counted in calculating $p^*(\mathbf{Q},\mathbf{n})$.

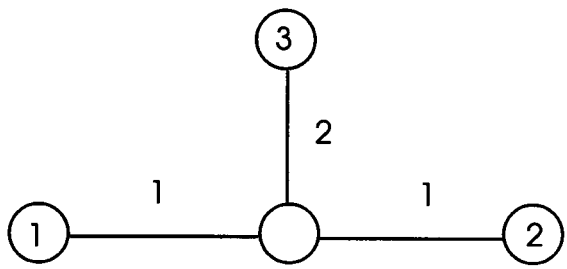

FIG. 5. Unrooted genealogy for numerical example.

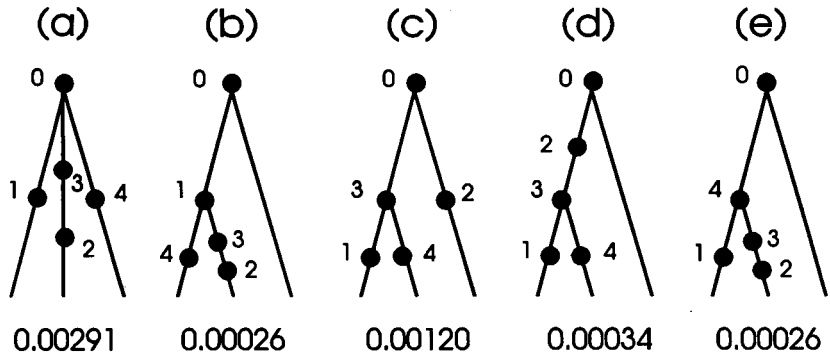

FIG. 6. Labeled rooted tree probabilities.

The genealogy in Figure 5 fits into Strobeck's [10] genealogy in his Figure 1(c), and his (revised) algorithm gives the same probability.

In this small genealogy the coalescent trees with four mutations can be enumerated to find the probability of the genealogy. The trees which produce the tree in Figure 5 are shown in Figure 7, with correspondence to the trees in Figure 6 highlighted.

Let T_3 be the time during which the sample has three ancestors and T_2 the time during which it has two. T_3 and T_2 are independent exponential random variables with rates 3 and 1, respectively. By considering the Poisson nature of the mutations along the edges of the coalescent tree and adding the probabilities of the trees in Figure 7 we find that

$$p^*(\mathbf{Q},\mathbf{n}) = \mathbb{E}\left(\left(\frac{\theta}{2}\right)^4 e^{-(3T_3+2T_2)\theta/2}\left(T_3^2(T_2+T_3)^2/2! + 2T_3^3(T_2+T_3)/2! + 2T_3^3T_2/2! + T_3^2T_2(T_2+T_3) + T_3^2T_2^2/2!\right)\right)$$

Evaluating this expectation gives the correct probability 0.0149176.

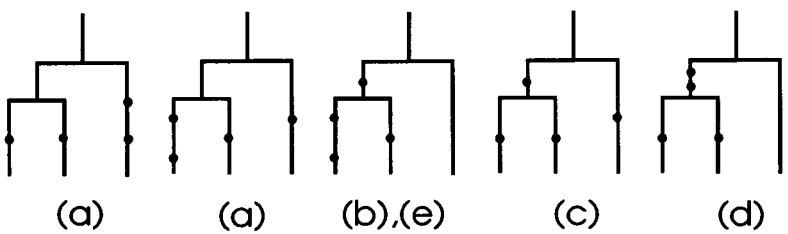

FIG. 7. Possible coalescent trees producing Figure 6 trees.

3. CALCULATING GENEALOGICAL TREE PROBABILITIES

For small sample sizes, probabilities like $p^0(\mathbf{Q},\mathbf{n})$ can be computed numerically by solving the recursion (7) for each rooted tree using Griffiths' program PTREE [4], and then combining the results using (10). In practice this direct combinatorial approach is impractical for large sample sizes. In this section, we present an alternative method for computing probabilities like $p^0(T,\mathbf{n})$. The obvious strategy of using the Ethier–Griffiths simulation method in (3) down from the common ancestor is not useful in practice because most of the time the *observed* $(T,\mathbf{n})$ will not be hit. Rather, we base our approach on simulating back up the tree from the sample to the ancestor. The basic ingredients of the method are discussed in Griffiths and Tavaré [6], where simulation algorithms for probabilities of sample configurations of DNA sequence data generated by the finite-sites models are provided.

The idea is to use (7) to construct a Markov chain with a set of absorbing states in such a way that the probability $p^0(T,\mathbf{n})$ can be written as the expected value of a functional of the Markov chain up to the time the absorbing states are hit. The appropriate Markov chain $\{X(l),\ l=0,1,\ldots\}$ has a tree state space, and makes transitions as follows:

$$(T,\mathbf{n}) \to (T,\mathbf{n}-\mathbf{e}_k) \text{ with probability } \frac{(n_k-1)}{f(T,\mathbf{n})(n+\theta-1)} \tag{17}$$

$$\to (\mathscr{S}_k T,\mathbf{n}) \text{ with probability } \frac{\theta}{f(T,\mathbf{n})n(n+\theta-1)} \tag{18}$$

$$\to \left(\mathscr{R}_k T,\mathscr{R}_k(\mathbf{n}+\mathbf{e}_j)\right) \text{ with probability } \frac{\theta(n_j+1)}{f(T,\mathbf{n})n(n+\theta-1)}. \tag{19}$$

In (17), (18), and (19), $k=1,2,\ldots,d$. The first type of transition is only possible if $n_k>1$, and the second or third if $n_k=1$. In the last two transitions a distinct singleton first coordinate in a sequence is removed. The resulting sequence is still distinct from the others in (18), but in (19) the shifted kth sequence is equal to the jth sequence. The scaling factor is

$$f(T,\mathbf{n}) \equiv f_\theta(T,\mathbf{n}) = \sum_{k=1}^{d} \frac{(n_k-1)}{(n+\theta-1)} + \frac{\theta m}{n(n+\theta-1)},$$

where m is given by

$$m = \left|\{k : n_k = 1, x_{k,0} \text{ distinct}, \mathscr{S}\mathbf{x}_k \neq \mathbf{x}_j \forall j\}\right| + \sum_{\substack{k: n_k = 1 \\ x_{k,0} \text{ distinct}}} \sum_{j: \mathscr{S}\mathbf{x}_k = \mathbf{x}_j} (n_j + 1).$$

The idea is to simulate the X process starting from an initial tree $(T, \mathbf{n})$ until the time τ at which there are two sequences $(x_{10}, \ldots, x_{1i})$ and $(x_{20}, \ldots, x_{2j})$ with $x_{1i} = x_{2j}$ (corresponding to the root of the tree) representing a tree T_2. The probability of such a tree is

$$p^0(T_2) = (2 - \delta_{i+j,0}) \binom{i+j}{j} \left[\frac{\theta}{2(1+\theta)}\right]^{i+j} \frac{1}{1+\theta}.$$

The representation of $p^0(T, \mathbf{n})$ is now

$$p^0(T, \mathbf{n}) = \mathbb{E}_{(T,\mathbf{n})}\left[\prod_{l=0}^{\tau-1} f(T(l), \mathbf{n}(l))\right] p^0(T_2), \tag{20}$$

where $X(l) \equiv (T(l), \mathbf{n}(l))$ is the tree at time l. Equation (20) may be used to produce an estimate of $p^0(T, \mathbf{n})$ by simulating independent copies
of the tree process $\{X(l),\ l = 0, 1, \ldots\}$, and computing $[\Pi_{l=0}^{\tau-1} f(T(l), \mathbf{n}(l))] p^0(T_2)$ for each run. The average over all runs is then an unbiased estimator of $p^0(T, \mathbf{n})$. An estimate of $p^*(T, \mathbf{n})$ can then be found by dividing by $a(T, \mathbf{n})$.

It is possible to construct a simulation method to find the likelihood of an unrooted genealogy based on (12). However, it seems best to proceed by finding all the possible rooted labeled trees corresponding to an unrooted genealogy and their individual likelihoods.

3.1. *COMPUTING A LIKELIHOOD SURFACE*

The method of Griffiths and Tavaré [6] can be modified to compute $p^0(T, \mathbf{n})$ for fixed $(T, \mathbf{n})$ as a function of θ from a single realization of the process $\{X(l),\ l = 0, 1, \ldots\}$. In the context of estimating θ, this provides a Monte Carlo approximant to a likelihood surface. This method is as follows: Simulate the chain $\{X(l),\ l = 0, 1, \ldots\}$ with a particular value θ_0 as parameter and obtain the likelihood surface for other values of θ using the representation

$$p_\theta^0(T, \mathbf{n}) = \mathbb{E}_{(T,\mathbf{n})}^{\theta_0}\left[\prod_{l=0}^{\tau-1} h((T(l), \mathbf{n}(l)), (T(l+1), \mathbf{n}(l+1)))\right] p_\theta^0(T_2), \tag{21}$$

where (as before) $(T(l),\mathbf{n}(l))$ is the tree at time l, and h is determined by

$$h((T,\mathbf{n}),(T,\mathbf{n}-\mathbf{e}_k)) = f_{\theta_0}(T,\mathbf{n})\frac{n+\theta_0-1}{n+\theta-1},$$

and

$$h((T,\mathbf{n}),(T',\mathbf{n}')) = f_{\theta_0}(T',\mathbf{n}')\frac{\theta(n+\theta_0-1)}{\theta_0(n+\theta-1)},$$

this last holding both for transitions of the form (18), when $(T',\mathbf{n}') = (\mathscr{S}_k T,\mathbf{n})$, and of the form (19), when $(T',\mathbf{n}') = (\mathscr{R}_k T,\mathscr{R}_k(\mathbf{n}+\mathbf{e}_j))$.

3.2. *CHECKING THE ALGORITHM*

In order to check the Monte Carlo algorithm, we use an example for which exact results can be computed using Griffiths' PTREE algorithm [4]. The sample of size $n = 30$ is described in Table 1. The sample has the rooted labeled genealogy given in Figure 8. For illustration, we assume that the labeling of ancestral and mutant sites is known.

In Table 2 exact values of the tree probability $p^*(T,\mathbf{n})$ are given, together with simulation estimates based on 30,000 runs, for a variety of θ values. Notice that the approximate confidence intervals, which are based on the simulation variance, cover the true value in each case.

To check the surface simulation method for approximating probabilities, we ran the simulation method determined by (21) for $\theta = 0.6(0.2)3.0$, $\theta_0 = 1.0, 1.8, 2.6$, using 30,000 replicates each. The approximating curves of $\log p^*(T,\mathbf{n})$ for different θ_0 values are combined by weighting inversely proportional to the estimated variance. This composite curve

TABLE 1

Alleles and Their Frequencies

Sequences								Frequency
1	0	0	1	0	0	0	1	3
2	0	0	0	0	0	0	1	4
3	0	0	0	0	0	0	0	4
4	1	0	0	1	0	0	0	11
5	1	0	0	0	0	0	0	1
6	0	1	0	0	0	0	0	2
7	0	0	0	0	1	0	1	2
8	0	0	0	0	1	1	1	3

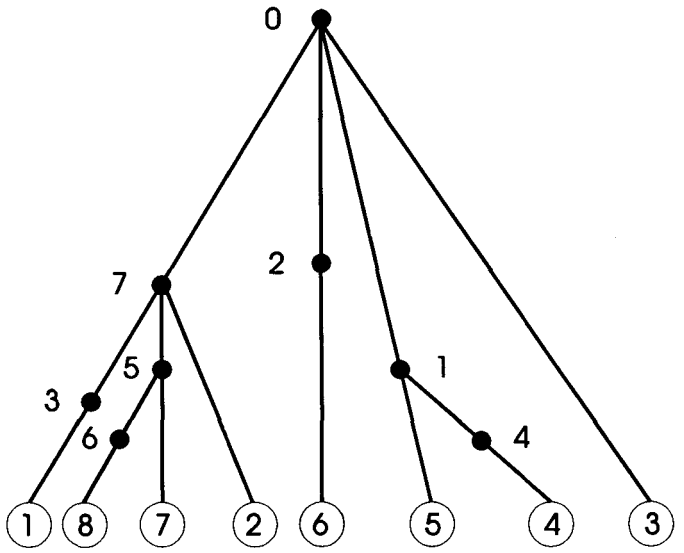

FIG. 8. Genealogical tree corresponding to Table 1.

and the true curve computed by Griffiths' algorithm are shown in Figure 9. The agreement is extremely good. If the ancestral labeling is unknown, then the probabilities of the different rooted trees can be computed individually and summed according to (10) to approximate the probability of the unrooted tree.

4. DISCUSSION

One important application of the theory we have developed here is maximum likelihood estimation of the substitution rate θ in the typical

TABLE 2

Exact and Estimated Tree Probabilities

θ	scale	True probability	Estimated probability	95% Confidence interval
1.0	10^{-12}	4.77	4.25	(3.52,4.97)
2.0	10^{-11}	1.27	1.35	(1.13,1.58)
4.0	10^{-12}	3.33	3.22	(2.85,3.59)
6.0	10^{-13}	3.85	4.01	(3.42,4.60)
8.0	10^{-14}	4.12	3.57	(2.95,4.20)
10.0	10^{-15}	4.75	4.28	(2.81,5.75)

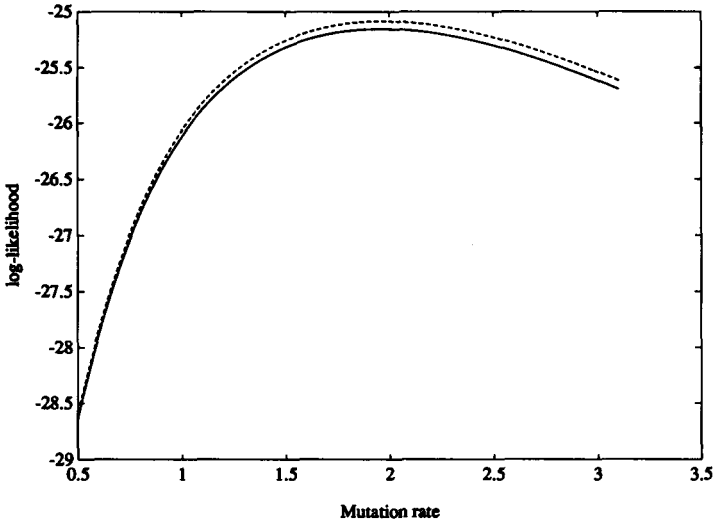

FIG. 9. Log-likelihood curves (dashed line: exact values; solid line: Monte Carlo approximant).

case where the ancestral labeling of sites is unknown. Further investigation of the statistical properties of the MLE $\hat{\theta}$ is certainly worthwhile. We are currently investigating how much better the MLE is than the simpler estimator (4) based on the number of segregating sites.

In addition to maximum likelihood estimation of the substitution rate θ, there are several other applications of the theory and simulation method we have described here. Among these is ancestral inference. For example, we may use the likelihoods of individual rooted trees to compare different ancestral labelings of the sites. The methods may also be extended to make inferences about the distribution of the time to the most recent common ancestor of a sample, conditional on the observed sequences.

A program PTREESIM implements the algorithm for simulating the likelihood of a sample of genes from the infinitely-many-sites model using the representation in (20). Maximum likelihood estimation of θ is achieved by simulating likelihood surfaces with respect to θ, as described in Section 3.1. The best way to treat the case when the ancestral base type at segregating sites is unknown is to produce a listing of all the labeled rooted genealogical trees with their likelihoods and then add up these likelihoods to find the likelihood of the unlabeled unrooted genealogical tree. PTREESIM automates this procedure for

single likelihoods and for surface simulation. Checking consistency of a collection of sequences with the infinitely-many-sites model, and the production of a (rooted or unrooted) tree from these sequences is an option in PTREESIM. The algorithm used is that in Griffiths [3]. PTREESIM is available from the authors on request in portable *C* source code and in executable form for a PC.

The infinitely-many-sites process described here is certainly an oversimplified picture of the variability observed in DNA sequence data. The assumption that once a mutation has occurred at a site there may be no further mutation there is clearly in conflict with the occurrence of back substitutions, for example. Griffiths and Tavaré [6] show how the Monte Carlo recursion technique can be used to approximate the likelihood of samples from coalescent models which allow for back substitutions. These techniques are considerably more expensive computationally than the present ones.

The sampling theory developed here applies to samples from populations that have maintained approximately constant size for many generations, an assumption that is clearly unreasonable in some cases. The effects of deterministically varying population size can be incorporated into the analysis; see Griffiths and Tavaré [5]. The combinatorial and topological structure of the tree $(T, \mathbf{n})$ is precisely the same as that described here, but the probabilistic structure has to be modified to account for the fact that the times T_j while the sample has j distinct ancestors are dependent random variables.

The authors were supported in part by NSF grant DMS 90-05833, and by the IMA at the University of Minnesota.

REFERENCES

1 S. N. Ethier and R. C. Griffiths, The infinitely-many-sites model as a measure valued diffusion, *Ann. Probab.* 15:414–545 (1987).
2 J. Felsenstein, Numerical methods for inferring evolutionary trees, *Quart. Rev. Biol.* 57:379–404 (1982).
3 R. C. Griffiths, An Algorithm for Constructing Genealogical Trees, Statistics Research Report #163, Department of Mathematics, Monash University, 1987.
4 R. C. Griffiths, Genealogical-tree probabilities in the infinitely-many-site model, *J. Math. Biol.* 27:667–680 (1989).
5 R. C. Griffiths and S. Tavaré, Sampling theory for neutral alleles in a varying environment, *Phil. Trans. Roy. Soc. London B*, 344:403–410 (1994).
6 R. C. Griffiths and S. Tavaré, Simulating probability distributions in the coalescent, *Theor. Pop. Biol.* 46:131–159 (1994).

7 D. Gusfield, Efficient algorithms for inferring evolutionary trees, *Networks* 21:19–28 (1991).
8 R. R. Hudson, Gene genealogies and the coalescent process, in *Oxford Surveys in Evolutionary Biology*, vol. 7, D. Futuyma and J. Antonovics, eds., Oxford University Press, 1991, pp. 1–44.
9 J. F. C. Kingman, On the genealogy of large populations, *J. Appl. Probab.* 19A:27–43 (1982).
10 C. Strobeck, Estimation of the neutral mutation rate in a finite population from DNA sequence data, *Theor. Pop. Biol.* 24:160–172 (1983).
11 S. Tavaré, Calibrating the clock: Using stochastic processes to measure the rate of evolution, in *Molecular Biology and Mathematics*, E. S. Lander, ed., National Academy Press, to appear.
12 G. A. Watterson, On the number of segregating sites in genetical models without recombination, *Theor. Pop. Biol.* 7:256–276 (1975).