Theoretical Population Biology **55**, 248–259 (1999)
Article ID tpbi.1998.1403, available online at http://www.idealibrary.com on IDEAL®

Recombination as a Point Process along Sequences

Carsten Wiuf and Jotun Hein

Institute of Biological Sciences, University of Aarhus, DK-8000 Aarhus, Denmark

Received January 15, 1998

Histories of sequences in the coalescent model with recombination can be simulated using an algorithm that takes as input a sample of extant sequences. The algorithm traces the history of the sequences going back in time, encountering recombinations and coalescence (duplications) until the ancestral material is located on one sequence for homologous positions in the present sequences. Here an alternative algorithm is formulated not as going back in time and operating on sequences, but by moving spatially along the sequences, updating the history of the sequences as recombination points are encountered. This algorithm focuses on spatial aspects of the coalescent with recombination rather than on temporal aspects as is the case of familiar algorithms. Mathematical results related to spatial aspects of the coalescent with recombination are derived. © 1999 Academic Press

INTRODUCTION

Understanding the genealogical relationship between sequences in a population has been central to recent analysis of the dynamics of sequence evolution at the population level. The stochastic process generating the genealogical relationship between k sampled sequences from a population with constant size N and no recombination was first described by Watterson (1975) and further developed into the theory of the coalescent by Kingman (1982). The process of evolution of sequences subject to both coalescence and recombination in a population was first described by Hudson (1983). In Hudson's setup a combined coalescent and recombination process is followed back in time until any nucleotide position in the extant sequences has only one ancestral nucleotide. The ancestral nucleotides can be located on different sequences. In this approach, operations are performed on sequences, operations being either coalescence or recombinations.

In this article an alternative algorithm which generates sample genealogies is given. The algorithm moves spatially along the sequences and updates the history of the sequences as recombination points are encountered. The algorithm is formulated as a point process indexed by sequence length measured in expected number of recombination events. It takes values in a set of graphs that increase with sequence length.

By definition, the coalescent with recombination is a Markov process since generation $t+1$ is determined from the previous generation t going backward in time. In contrast, the information necessary at any position p to determine the genealogical history at position $q>p$ includes information on the non-ancestral material that links regions of ancestral material. Sequences non-ancestral to p must be added. This implies that the coalescent tree describing the history of a single position p is not sufficient to build an algorithm which moves spatially along the sequences. The algorithm presented here is thus not Markovian in the above sense.

Some mathematical results related to the new algorithm and to spatial aspects of the coalescent with recombination are derived and discussed in the paper.

0040-5809/99 $30.00

EVOLUTION OF SEQUENCES SUBJECT TO RECOMBINATION

The model of a population of sequences subject to recombination is the following: Each sequence is L nucleotides long and recombination is assumed to occur to the left of a nucleotide. The population is constant and of size N and diploid; i.e., there are $2N$ sequences. A new generation is obtained from the present by (1) sampling $2N$ sequences in the old population with replacement and (2) forming random pairs of sequences and letting the pairs recombine at a random position between any two nucleotides with probability r. Time will start at the present and increase going backward in time.

This process is transformed to a continuous time and continuous sequence process by letting $N \to \infty$ and measuring time in $2N$ generations, and letting $L \to \infty$ and $r \to 0$, such that $4rLN \to \rho$, where $2rLN$ is the expected number of recombinations per $2N$ generations. Sequence length will be measured in expected number of recombinations per $2N$ generations. Hudson (1983) showed that the waiting time until a sequence is created by a recombination event from two sequences is exponentially distributed with intensity parameter $\rho_0/2$. For the extant sequences, $\rho_0/2$ is simply the length of the sequences; i.e., $\rho_0 = \rho$. For ancestral sequences, $\rho_0/2$ is the length of the interval spanned by regions that have ancestral material. This interval can include regions with non-ancestral material. The recombination point will be uniformly distributed within this material. The waiting time going backward in time until k sequences have only $k-1$ ancestors in the population is exponentially distributed with intensity parameter $k(k-1)/2$ and the two sequences that coalesce are chosen uniformly from all different, unordered pairs of sequences.

The history of k sequences can be simulated by going back in time, waiting for what occurs first, recombination or coalescence, and then performing the appropriate operation on the set of ancestral sequences. Recombination will increase the number of sequences carrying ancestral material by one, but will not increase the total amount of ancestral material. A coalescent event will decrease the number of sequences with ancestral material by one. It may increase the amount of material, where recombination can occur, because a coalescence can trap some non-ancestral material, called trapped material. When any position on the extant sequences has found one ancestor, all segments with ancestral material spliced together will constitute one sequence. Above this point coalescence cannot reduce the amount of ancestral material and all that will occur are redistributions of ancestral material on different sequences by recombination and coalescent events.

The tree that describes the history of a given column in the sequence alignment is called a *local tree*. Let p be a position in $(0, \rho/2)$ along the sequences, where $\rho/2$ is the sequence length. The tree that describes the phylogeny of the sequences at this point p is $T(p)$, the local tree at this point (or nucleotide). A local tree $T(p)$ can be found by starting at the present sequences and going back in time. When a recombination node is encountered the branch that describes the segment containing p is followed. That is, if the recombination point is to the right of p, then the arrow describing the fate of the left part of the sequence is to be followed and vice versa. The local tree $T(p)$ at p will be distributed like the coalescent process since one point cannot be subject to recombination.

Griffith and Marjoram (1997) embedded the coalescent with recombination in a birth and death process with birth rate $\mu_k = k(k-1)/2$ and death rate $\lambda_k = k\rho/2$. The parameter ρ is here defined by $\rho = \lim_{N \to \infty} 4Nr$ and is not dependent on sequence length. This process simplifies mathematics on the account that the notion of an ancestor will have a less restrictive meaning than usual: An "ancestral" sequence in the birth and death process need not have any genetic material in common with a sequence descended from it. Hence the genealogy of a sample of extant sequences described by the birth and death process will in general include more "ancestral" sequences than those carrying material ancestral to the sample. However, the process is useful as a general framework for studying genealogies, and as such will be used in this paper. The graph corresponding to this process is called the ancestral recombination graph (Griffiths and Marjoram, 1997).

THE SPATIAL ALGORITHM

We will now describe an algorithm which simulates the history of k sequences in the coalescent process with recombination. It starts at position 0 in the sequences and describes more of the history of the sequences by moving right, along the sequences; hence it is called a *spatial algorithm*. The history of the nucleotides at position 0 is automatically given by the traditional coalescent without recombination. As the algorithm moves right, along the sequences, recombinations that must be incorporated into the genealogical history of the sequences will be encountered.

The mathematical formulation of the algorithm is cumbersome compared to the algorithm by Hudson

(1983), and the complexity of the algorithm measured in expected number of events has only partly been derived theoretically. Based on simulation studies, the complexity is believed to be of the same order of magnitude as that of Hudson's algorithm. The complexity of both algorithms is compared to the complexity of the birth and death process described above. The main objective of the algorithm is to discuss the spatial aspect of the coalescent with recombination in contrast to temporal aspects.

Some preliminary considerations are required before the algorithm is stated.

1. Sequence Length to Recombination Point

Essential to what follows is the distribution of the sequence length until a recombination is encountered conditional on the total branch length b of the genealogy. For a fixed position p in the sequences consider L nucleotides next to this position. Assume that the total branch length of the genealogy measured in generations is $2Nb$. The number of recombination points, n, in these segments of length L is binomially distributed,

$$p(n) = \frac{(2NbL)!}{n!(2NbL-n)!} r^n (1-r)^{2NbL-n} \sim \mathrm{Bi}(n; 2NbL, r),$$

which tends to a Poisson distribution for large population sizes N and long segments of nucleotides L,

$$\frac{(b\rho/2)^n}{n!} \exp(-b\rho/2) \sim \mathrm{Po}(n; b\rho/2).$$

From this it easily follows that the sequence length X until a recombination point is encountered conditional on b is given by

$$P(\text{no recombination} \mid b) = \exp(-b\rho/2),$$

and

$$P(X > x \mid b) = \exp(-bx)$$

for $x < \rho/2$. As $\rho \to \infty$, X conditional on b becomes exponentially distributed with parameter b, and otherwise X follows an exponential distribution truncated at $\rho/2$. Moreover, the recombination event will happen with equal chance in all generations and ancestors of the genealogy, i.e., $T \sim U(0, b)$, where T is the location of the event. This is an analogue to the statement that given total sequence length the waiting time until the first recombination event is exponential with an intensity half the total sequence length.

2. Waiting Time for a Sequence to Coalesce to Known Genealogy

The rate of coalescence between a pair of sequences is 1, so the conditional coalescence rate of an additional sequence to a group of k sequences known not to coalesce before the additional sequence does is k.

3. Notation

The notation used is adopted and modified from Griffiths and Marjoram (1997). A sequence S of length x is represented as an interval $[0, x[$ with a set $A \subseteq [0, x[$

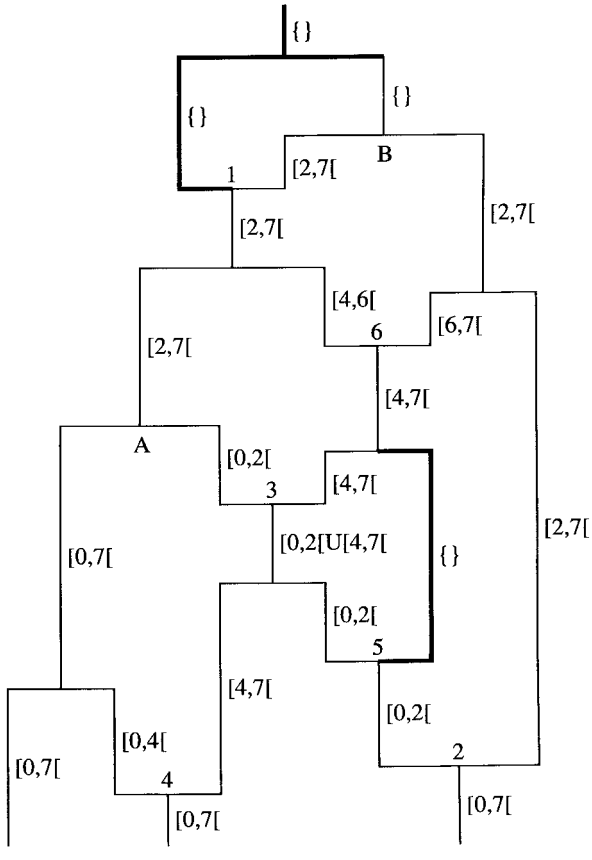

FIG. 1. The graph HUD. This example will be continued on subsequent figures. Sequences are assumed to have length 7. The history of a sample of size 3 is followed back in time until there is only one ancestral sequence. The whole graph shows the ancestral recombination graph, ARG, and the thin lines constitute HUD. Recombination points are written above the recombination event. The set $[x, y[$ written to the right of an edge is the set $A(e)$ of ancestral material that has not yet found a common ancestor. Positions $[0, 2[$ find a common ancestor at event A, whereas positions $[2, 7[$ find a common ancestor at event B. If event A happens at time t_A and event B at time t_B, then $\mathrm{CA}(t_A) = [0, 2[$ and $\mathrm{CA}(t_B) = [2, 7[$. For $t \neq t_A, t_B$ we have $\mathrm{CA}(t) = \varnothing$. Recombination points 1 and 5 are both outside ancestral and trapped material, and hence the two recombination events have no influence on the history. Recombination point 3 is in trapped material.

being associated. A is the set of ancestral material on the sequence S that has not yet found a common ancestor in the sample. For extant sequences $A = [0, x[$, and a sequence can have $A = \varnothing$ if it is created by a recombination event outside ancestral and trapped material, or if all ancestral positions in the sequence have found a common ancestor.

The genealogy G of a sample of size k is represented as a graph. Edges represent ancestors to one or more sequences in the sample. Vertices in the graph denote extant sequences or events: we denote the coalescence of edges e_1 and e_2 to e_3 by $e_3 = e_1 C e_2$, and the recombination of e_3 into e_1 (left part) and e_2 (right part) by $e_3 = e_1 R e_2$. Note that the C relation is symmetric, whereas the R relation is not. Positions at which breaks occur and time of events (measured from the present) are labeled on the graph. When necessary, the notation C_t, R_t, or R_p is used to indicate the time t of an event or the break at point p. Let $\mathrm{CA}(t)$ denote the set of positions in the sample with a most recent common ancestor at time t. An edge e has ancestral material in common with the sample if a sequence $S(e)$ representing e has either (1) $A(e) \neq \varnothing$ or (2) $A(e) = \varnothing$ and $\mathrm{CA}(t) \neq \varnothing$ assuming e is created by an event at time t. If $e_3 = e_1 R_p e_2$ then $A(e_1) = A(e_3) \cap [0, p[$ and $A(e_2) = A(e_3) \cap [p, x[$, and if $e_3 = e_1 C_t e_2$ then $A(e_3) = A(e_1) \cup A(e_2) \backslash \mathrm{CA}(t)$. Note that $\mathrm{CA}(t)$ can only be non-empty, if the event at time t is a coalescent event. The notation is illustrated in Fig. 1.

The ancestral recombination graph ARG possibly includes sequences with no material ancestral to the sample, and describes the history of a sample until there is only one sequence present in the ancestral sample. This sequence is called the grand most recent common ancestor (grand MRCA, Griffiths and Marjoram, 1997). The genealogy HUD obtained by Hudson's algorithm (Hudson, 1983) is the smallest genealogy containing all ancestral sequences carrying material ancestral to the sample, i.e.,

$$e \in \mathrm{HUD} \quad \text{iff} \quad A(e) \neq \varnothing \quad \text{or}$$
$$A(e) = \varnothing, \quad e = e_1 C_t e_2 \quad \text{and} \quad \mathrm{CA}(t) \neq \varnothing,$$

for some e_i, $i = 1, 2$ in ARG and some $t > 0$.

When subgraphs of ARG are considered, we use the convention that a vertex with only two outgoing edges will be ignored, so that the two edges reduce to one.

4. The Spatial Algorithm Graph

The spatial algorithm graph will be defined recursively as the limit of a series of subgraphs G_i of ARG, and will be explained in detail below and in Figs. 2–5. Define G_0 by

$$e \in G_0 \quad \text{iff} \quad 0 \in A(e) \quad \text{or} \quad e = e_1 C_t e_2, \quad 0 \in \mathrm{CA}(t),$$

for some e_i, $i = 1, 2$ inch ARG and some $t > 0$. Only one time point satisfies $0 \in \mathrm{CA}(t)$. G_0 is the graph that consists of all edges in ARG that describe the coalescent history of position 0 and corresponds to the local tree $T(0)$.

Moreover, G_0 describes the history of all positions between 0 and the recombination point p_1 nearest 0. The graph G_1 will he the graph that describes the genealogy of all positions before the second recombination point p_2, and so forth: G_i will describe the history of positions $[0, p_{i+1}[$.

Denote by f_{10} the unique edge in G_0 fulfilling $f_{10} = e_1 C_t e_2$ and $0 \in \mathrm{CA}(t)$ for some e_i, $i = 1, 2$ and some $t > 0$. The edge f_{10} is the most recent common ancestor

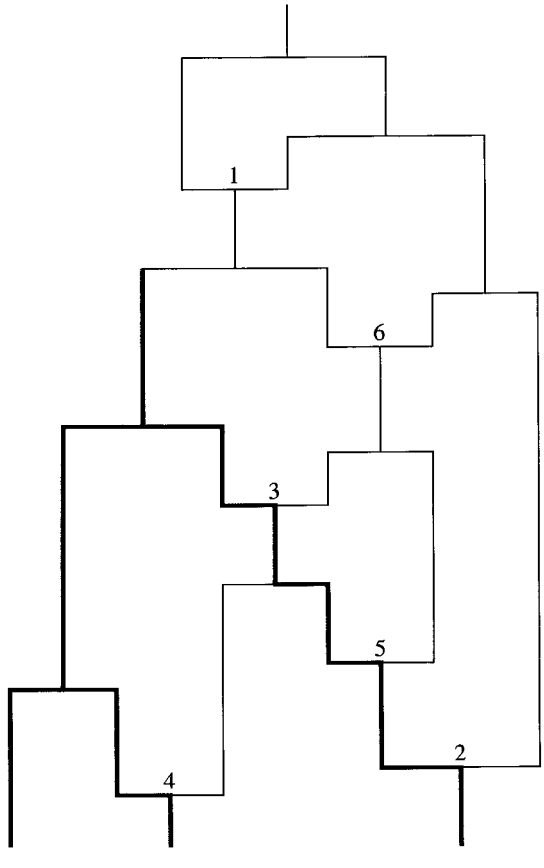

FIG. 2. Figures 2–5 illustrate the recursively defined G_i's. The ARG is as in Fig. 1. Thick lines show the coalescent tree G_0 for position 0 in the sample of size 3. At each recombination event encountered the edge describing position 0 is chosen, i.e., the left branch.

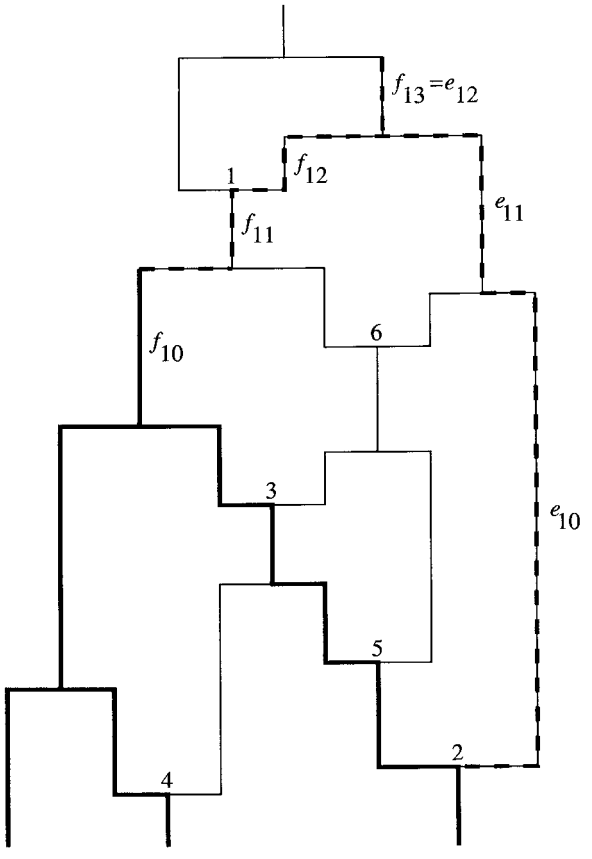

FIG. 3. First recombination point encountered. There are four recombination vertices in G_0 with points greater than $p_0=0$: 2, 3, 4, and 5. The smallest of these is 2; i.e., $p_1=2$. The history of position 2, $e_{10}, e_{11}, ...$, is followed until an edge e_{1j} either coalesces with an edge in G_0 or coalesces with an edge in the series of edges, $f_{10}, f_{11}, ...$, describing the history of position 2 in the edge ancestral to all edges in G_0, f_{10}. Here the latter is the case. G_1 consists of G_0 and the series $e_{10}, ..., e_{12}$ and $f_{10}, ..., f_{13}$, and f_{20} is defined by $f_{20}=f_{13}$.

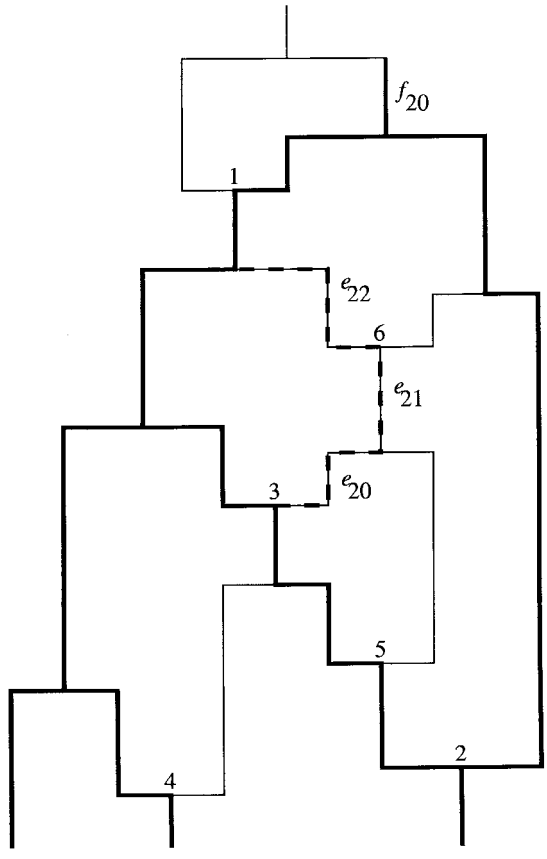

FIG. 4. Second recombination point encountered. Continuing similarly to Fig. 3, the smallest recombination point larger than $p_1=2$ is position 3. Hence $p_2=3$. Moving spatially along the sequences, the sequence e that experiences the recombination event has $A(e)=[0, 2[$ (Fig. 1, only ancestral material up to position 3 is taken into account), and the event would possibly not affect the sample's history. However, the third recombination point encountered makes sure that the previous event cannot be disregarded (Fig. 1). In this case the series $e_{20}, ..., e_{22}$ describing the history of position 2 in e_{20} coalesces with an edge in G_1.

(MRCA) to position 0 in the sample. Let $p_0=0$ and assume that G_{i-1} is defined for some $i>0$: Put

$$P_i=\{p<p_{i-1} \mid R_p \text{ vertex in } G_{i-1} \text{ with } e_2=e_1R_pe,\ e_1, e_2\in G_{i-1}\}.$$

P_i is the set of recombination points in G_{i-1} larger than p_{i-1}. Define $p_i=\min P_i$, and denote by e_{i0} the edge such that $e_2=e_1R_{p_i}e_{i0}$ for some e_i, $i=1, 2$. If $P_i=\varnothing$ let $G_i=G_{i-1}$, else let $f_{i0}, f_{i1}, ..., f_{im}$ be the edges describing the history of position p_i in f_{i0} until the grand MRCA, f_{im}. Let $e_{i0}, e_{i1}, ..., e_{in}, e_{i(n+1)}$ be the finite series of edges describing the history back in time of p_i in e_{i0} until either

(1) $e_{i(n+1)}\in G_{i-1}$ or

(2) $e_{i(n+1)}\in\{f_{ij}\}_{j=1, ..., m}$.

In the first case, the recombined segment coalesces with an edge in G_{i-1} and otherwise it coalesces with an ancestral edge to f_{i0}. There are only these two possibilities, since ARG is continued until there is one ancestral edge, the grand MRCA, f_{im}. Define $f_{(i+1)0}=f_{il}$, where $l=0$ if $e_{i(n+1)}\in G_{i-1}$, and else l is determined uniquely by (2). Define G_i by

$$G_i=G_{i-1}\cup\{e_{ij}\}_{j=0, ..., n}\cup\{f_{ij}\}_{j=1, ..., l},$$

and SAG by

$$\text{SAG}=\bigcup_{i=0}^{\infty} G_i.$$

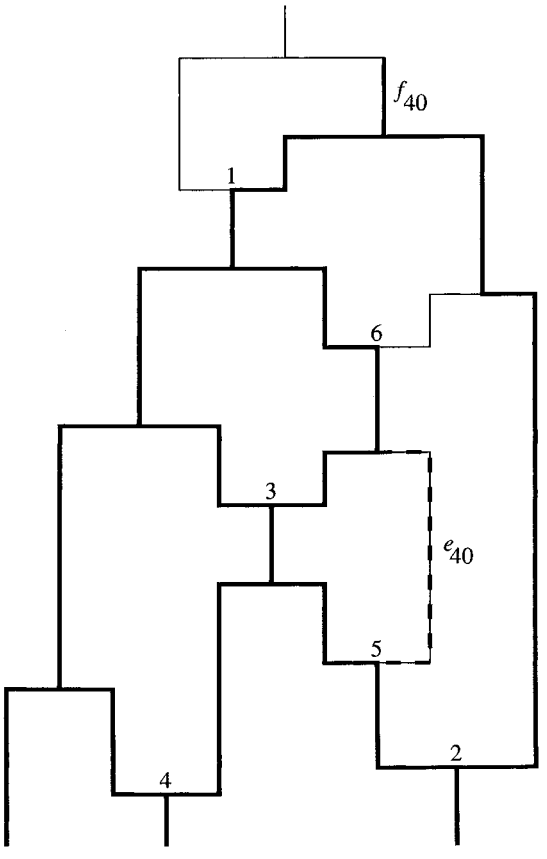

FIG. 5. Fourth recombination point encountered. Similar to the second one the fourth recombination point happens outside ancestral material. The event has no effect on the history of the sample, but if the sequences were prolonged, the event potentially could affect the genealogy (compare Fig. 4). Since the edge e_{40} does not carry any ancestral material to the sample it is not included in HUD (Fig. 1).

The edge f_{i0} belongs to G_{i-1} for all i, and is thus not included in the definition of G_i. Figures 2–5 illustrate the recursive definition of G_i's.

Due to the definition of G_0, a single edge is the ancestor to all edges in G_0, and by induction on i, it follows that G_i has the same property.

Since ARG is a finite graph, there can only be a finite number of different G_i's; therefore SAG is finite and there is a single edge being ancestral to all edges in SAG.

If $e \in G_i$ then an induction argument on i shows that there exists a path from $e \in G_i$ to an extant edge in $G_0 \subseteq G_i$. This property and the definition of the series e_{ij} ensure that there are only two possibilities for e_{in},

$$z = yC\, e_{in},$$

where either

$$y = f_{i(l-1)}, \qquad z = f_{il} \qquad \text{or} \qquad y, z \in G_{i-1}.$$

Moreover for $j < n$ either

$$e_{i(j+1)} = e_{ij} Cx, \qquad e_{ij} = e_{i(j+1)}\, R_q x,\ q > p_i,$$

or

$$e_{ij} = xR_q\, e_{i(j+1)},\ p_i > q,$$

x denoting different edges not in G_{i-1}.

As defined G_0 is the history of position 0. Moving spatially along the sequences the first recombination event happens in position p_1, and hence G_0 describes the history of all the positions in $[0, p_1[$. The fate of position p_1 in the sequence e_{10} recombining at position p_1 is followed through the edges $e_{11}, ..., e_{1n}$ until the sequence containing p_1 finds a common ancestor, either with a sequence included in G_0 or with one ancestral to p_1 in f_{10} ($y, z \in G_0$ and $y = f_{1(l-1)}$, $z = f_{1l}$ respectively). This describes the history of all positions $[0, p_2[$ in the sample, since the next recombination point effecting the genealogy is p_2. The edges $f_{10}, ..., f_{1l}$ describe the history of the positions $[p_1, p_2[$. The history of the positions $p > p_2$ in f_{10} might be affected by the recombination break in p_2, whereas the history of the positions $[0, p_1[$ in f_{10} is not important, because the positions have found a common ancestor, f_0. This argument applies for all $i > 1$, thereby completing an increasing part of the samples history. Since ARG is finite, one will complete the history of the full sequences for some finite i. Recombination events that do not affect the history of the sample might be encountered as one moves along the sequences, thus making both the number of sequences and the number of events greater than the same numbers in HUD.

The total length of the series $e_{i0}, ..., e_{in}$ is the coalescent time until e_{i0} coalesces with one of the sequences in G_{i-1} or an ancestral sequence to position p_i (Figs. 2–5).

5. Spatial Algorithm Simulating the History of k Sequences

With the above considerations, **1–4**, and the description of SAG, the spatial algorithm for simulating histories of a sample of extant sequences can be stated:

Start leftmost at the sequences.

(1) Choose a coalescent tree (graph) $T(0)$ for the first position $p_0 = 0$ in the k sequences according to the distribution of the coalescent processes.

(2) Let b_0 be the total branch length of $T(0)$, and put $B_0 = b_0$, $P_0 = p_0$, and $G_0 = T(0)$.

For $i = 1, 2, \ldots$ repeat the following procedure as long as $P_{i-1} \leqslant \rho/2$.

(3) Choose recombination point $p_i \sim E(B_{i-1})$ and location $t_i \sim U(0, B_{i-1})$ as described in (**1**).

(4) Coalesce the recombined edge (sequence) e_i to the spatial algorithm graph G_{i-1} (**4**) according to the distribution of the coalescent process (**2**).

(5) Set $G_i = G_{i-1} \cup e_i$, $B_i = B_{i-1} + b_i$, and $P_i = P_{i-1} + p_i$.

Since the ancestral recombination graph is finite with probability one (Griffiths and Marjoram, 1997), and the spatial algorithm graph created by the algorithm is embedded in the ancestral recombination graph, this algorithm will stop eventually.

The spatial algorithm is formulated as a Markov process with state space being the set of spatial algorithm graphs with no sequence information added. The Markov property is however artificial in the sense that coalescence information will discretely be added as one encounters more recombination points along the sequences. No information will be subtracted. Figure 6 illustrates that it is not possible to formulate an algorithm as a Markov process on the set of local trees: The local tree $T(p)$ at position p is not sufficient to determine the local trees $T(q)$ for $q > p$. If only the local tree were considered then the branch representing sequences with the ancestral material $[0, p[$ would be ignored and the probability that $T(p)$ has the same tree height as $T(0)$ would be zero.

Break points are only needed to determine the local trees at each position, and are as such no essential part of the algorithm. The graphs are strictly growing as one moves along sequences in number of edges and vertices, as well as in total branch length. In Hudson's algorithm the amount of ancestral material is a decreasing function with time, but not strictly decreasing.

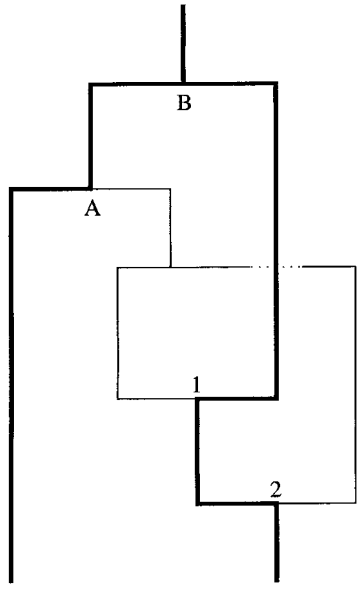

FIG. 6. Counterexample against "only local tree" algorithm. The genealogy of two sequences of length 3 is shown. Recombinations occur in positions 1 and 2 in ancestral sequences. The positions $[0, 1[\cup [2, 3[$ find a common ancestor at event A, and the positions $[1, 2[$ find a common ancestor at event B. If only the local tree of a position $p \in [1, 2[$ (marked with thick lines) is known, we will have no chance to determine whether the positions $[2, 3[$ find a common ancestor at the same time as the positions $[0, 1[$. Information on the local trees for the positions $q < p$ is necessary to account for the possibility of common ancestry of the positions $[0, 1[$ and $[2, 3[$.

6. Complexity

For all types of graphs discussed here, the following two relations are valid,

$$\text{coal} = \text{rec} + k - \text{anc}, \qquad \text{and hence}$$

$$\text{ev} = \text{rec} + \text{coal} = 2\text{rec} + k - \text{anc},$$

where k denotes sample size, ev the number of events in the graph, coal the number of coalescent events, rec the number of recombination events, and anc the number of "ancestral" sequences. In ARG and SAG anc is one, and in HUD it is stochastic.

The complexity of the algorithms will be measured in number of events, and only asymptotic results for $k \to \infty$ and $\rho \to \infty$ will be considered. Due to the above relations, attention can be restricted to the number of recombination events rec and the number of ancestral sequences anc. Ethier and Griffiths (1990) calculated the expected number of recombination events in the birth and death process and found that for $k \to \infty$ and ρ fixed

$$E(\text{rec}) \sim \rho \log(k),$$

and for $\rho \to \infty$ and k fixed

$$E(\text{rec}) \sim \exp(\rho).$$

For all ρ and k Ethier and Griffiths (1990) found

$$E(\text{rec}) \geqslant \exp(\rho) - 1.$$

The last inequality means that even for large sample sizes the dominant term will be $\exp(\rho)$, and not $\rho \log(k)$. The expectation of the number of recombination events is not known for Hudson's algorithm, but since this number is greater than the number of recombination events R_k within ancestral material (i.e., not counting recombination events within trapped material), R_k is a lower bound. Hudson and Kaplan (1985) found that

$$ER_k = \rho \sum_{i=1}^{k-1} \frac{1}{i} \sim \rho \log(k).$$

Moreover Griffiths and Marjoram (1997) found that

$$1 \leqslant E(\text{anc}) \leqslant 1 + \left(1 - \frac{2}{k^2 + k}\right) \rho \leqslant 1 + \rho.$$

Thus for fixed ρ and increasing sample size the spatial algorithm and Hudson's algorithm perform equally well in terms of complexity, since $\rho \log(k)$ is the asymptotic growth of the expected number of recombination events in the birth and death process. For increasing sample size the number of events ev will grow like k, and not like $\rho \log(k)$, because $E(\text{ev}) = 2E(\text{rec}) + k - E(\text{anc}) \approx 2\rho \log(k) + k \sim k$.

For increasing sequence length the bound on ER_k is too low to be used in comparison: For large sequences there will be large regions of trapped material enlarging the rate of recombination considerably and the linear bound on ER_k does not reflect this. Simulation results (not shown) indicate that the number of events grows subexponentially in ρ in SAG. Hence both algorithms are considerably faster than the birth and death process for large values of ρ.

MATHEMATICAL RESULTS

The coalescent process with recombination has been studied to some extent in the literature (Hudson, 1983; Hudson and Kaplan, 1985; Kaplan and Hudson, 1985; Griffiths and Marjoram, 1996, 1997; Wiuf and Hein, 1997; among others). Some new results will be derived which are related to the spatial algorithm, and the spatial aspects of the coalescent with recombination.

We consider sequences of infinite length, i.e., $\rho = \infty$.

1. Sequence Length until First Recombination Point

As proved in the previous section the distribution of sequence length X_k until the first recombination point conditional on the total branch length b of the genealogy of the first nucleotide is exponential with intensity parameter b. Moreover it is shown in the Appendix that the total branch length $B_k = \sum_{i=2}^{k} iW_i$, $W_i \sim E(i(i-1)/2)$, in the coalescent process is distributed like the maximum of $k-1$ exponential variables with intensity $\frac{1}{2}$; i.e., the density takes the form

$$f_k(b) = \tfrac{1}{2}(k-1)\{1 - \exp(-b/2)\}^{k-2} \exp(-b/2).$$

It follows that the unconditional sequence length until the first recombination point is

$$g_k(x) = \int_0^\infty b \exp(-bx)\, f_k(b)\, db,$$

which by induction is

$$g_k(x) = 2 \sum_{i=1}^{k-1} (-1)^{i-1} \frac{(k-1)!}{(i-1)!\,(k-i-1)!} \frac{1}{(i+2x)^2}.$$

For $k = 2$ and $k = 3$ this reduces to

$$g_2(x) = \frac{2}{(1+2x)^2}, \quad g_3(x) = \frac{4}{(1+2x)^2} - \frac{1}{(1+x)^2}.$$

The mean value of X_k is given by

$$\begin{aligned} EX_k &= \int_0^\infty x g_k(x)\, dx \\ &= \frac{1}{2} \sum_{i=1}^{k-2} (-1)^{i-1} \frac{(k-1)!}{i!(k-i-2)!} \log(i+1) \end{aligned}$$

for $k \geqslant 3$ and

$$EX_2 = \infty$$

for $k = 2$.

2. Tree Heights

If one compares the coalescent tree of the first position and the tree just to the right of the first recombination point, there are three possibilities: The trees are of equal height, the first tree is higher than the second tree, or the first tree is lower than the second. In general it is difficult to calculate the probabilities for these three possibilities, but for samples of size 2 this can be done. Let H_1 denote the tree height of the first tree, and H_2 the tree height of

the second. Formally $H_2 = H(X_2)$, where $H(x)$ is the height of the local tree in position x. Then (see Appendix)

$$P(H_1 = H_2) = P(H_1 > H_2) = \tfrac{1}{2} - \tfrac{1}{4}\log(3) \approx 0.225,$$

and

$$P(H_1 < H_2) = \tfrac{1}{2}\log(3) \approx 0.549.$$

Hence $P(H_1 < H_2) + P(H_1 > H_2) \approx 0.775$ and in about three out of four cases the first recombination will result in a change in most recent common ancestor.

An expression for the probability that $H_2 = H(X_2)$ is higher than H_1 can be given for samples of arbitrary size (see Appendix):

$$\begin{aligned} P(H_1 < H_2) &= \frac{1}{2}(k-1)\int_0^\infty \frac{1}{x}\{1+\exp(-x/2)\} \\ &\quad \times \{1-\exp(-x/2)\}^{k-1}\exp(-x/2)\,dx \\ &= \frac{1}{2}(k-1)\sum_{i=1}^{k}(-1)^i \\ &\quad \times \frac{(k-1)!}{(i-1)!(k-i)!}\{\log(i+1)+\log(i)\}. \end{aligned}$$

Even for larger samples this probability is high, tending slowly to zero; e.g., for $k = 2$, $P(H_1 < H_2) = 0.549$; $k = 3$; 0.405; $k = 5$; 0.292; $k = 10$; 0.206; $k = 100$; 0.102; and $k = 1000$; 0.069.

3. Second Recombination Point

The number of different genealogies makes it difficult to deduce general results on the length until the ith recombination point. To give an idea of this, the density h_2 of the length between the first and second recombination break in a sample of size 2 is stated here. The proof is sketched in the Appendix:

$$\begin{aligned} h_2(x) &= \frac{1}{(2x+1)^2}\left\{\frac{4x+3}{(2x+2)^2}\log(2x+3)\right. \\ &\quad \left. + \frac{4x+1}{(2x)^2}\log(2x+1) - \frac{1}{2x(x+1)}\right\}. \end{aligned}$$

For $x \to 0$, $h_2(x)$ tends to $\frac{3}{4}\log(3) + 2 \approx 2.824$. The chance that the length between the first and second recombination point is small is thus higher than the chance that the length until the first recombination point is small ($g_2(x) \to 2$ for $x \to 0$). Moreover in contrast to the mean value of g_2, the mean value of h_2 is finite.

4. Expected Height of the Second Tree

For samples of size 2 the expected height of the tree just to the right of the first recombination point is (see Appendix)

$$EH_2 = 1 + \tfrac{3}{8}\log(3) \approx 1.412.$$

Griffiths and Marjoram (1997) calculated the expected time until the most recent common ancestor (TMRCA) at position p given a recombination event at that point, and found for samples of size 2 that

$$E(\text{TMRCA at } p \mid \text{recombination at } p) = 2.$$

These two expected values are not supposed to be equal: The former is the unconditional expected height of the tree in the first recombination point (which is a stochastic point, X_2), whereas the latter is the conditional expected value given a recombination event in a fixed position. Due to symmetry the expected tree height on both sides of the recombination point is 2. This symmetry is not present in the first case, since the expected height of the tree just to the left of the first recombination point is 1 ($H_1 \sim E(1)$ in the coalescent process).

Furthermore the expectation of H_2 given $H_1 = s$ is (see Appendix)

$$E(H_2 \mid H_1 = s) = \frac{1}{4} + \frac{3}{4}s + \frac{3}{8s}\{1-\exp(-2s)\}.$$

$E(H_2 \mid H_1 = s)$ is a growing function of s with $E(H_2 \mid H_1 = s) \approx \frac{3}{4}s$ for large values of s. For small values of s the distribution of H_2 given $H_1 = s$ is approximately exponential with intensity 1 (see Appendix), as expected, and in agreement with $E(H_2 \mid H_1 = s) \to 1$ as $s \to 0$.

DISCUSSION

In this paper a new algorithm, the spatial algorithm, for simulating genealogical histories of a sample of k sequences subject to recombination is developed and discussed. It points to the fact that the coalescent with recombination can be considered a point process along sequences. Some new results related to the algorithm are derived.

The complexity of the algorithm measured in terms of expected number of events in the genealogical history produced by the algorithm was shown to be of the same order of magnitude as Hudson's algorithm for constant sequence length and variable sample size. Simulation

studies indicated that for increasing sequence length ρ and constant sample size, both algorithms are sub-exponential in ρ.

Some of the derived results under Mathematical Results hold for the coalescent process with mutation in an infinite-site model: In **3** all results carry over by changing "sequence length until first recombination point" with "sequence length from the ith to the $(i+1)$th mutation point." For example, the mean length between mutation points in a sample of size 2 is infinite. The reason that this results holds for all i, and not just $i=0$, is loosely speaking that all positions share the same history.

The spatial algorithm can be extended to cover demographic scenarios other than a population of constant size. To see this, consider the distribution of the length until the next recombination point conditional on the total length of the genealogy, b. This on distribution depends on b only, and not on the population history, on sizes of subpopulations, or on which ancestral sequences belong to which subpopulations. If sequences are of different "types" (e.g., from different subpopulations), the above holds true as long as all sequences involved in a recombination event is of the same type. In the situation with several subpopulations, if two sequences recombine to create a new sequence, all sequences must necessarily belong to the same subpopulation, i.e., be of the same type. The distribution of the length until the next recombination point conditional on b will not change if the distributions of the waiting times between coalescent events are changed accordingly. Simulated sample histories from other demographic scenarios can then easily be obtained by modifying 1 and 4 in Section **5** under The Spatial Algorithm. Griffiths and Tavaré (1997) and references therein provide methods for sampling coalescent times from different scenarios.

APPENDIX

In this Appendix proofs of statements found under Mathematical Results are given. Bold numbers refer to the section numbers used under Mathematical Results.

1: PROPOSITION. *Assume that* $X_i \sim E(\lambda+\mu i)$, $i=0, 1, \ldots$. *Put* $Y_i = Y_{i-1} + X_i$, $i>0$, *and* $Y_0 = X_0$. *Then the density* f_i *of* Y_i *is given by*

$$f_i(y) = \exp(-\lambda y)\{1-\exp(-\mu y)\}^i \frac{1}{\mu^i\, i!} \prod_{j=0}^{i} (\lambda+\mu j).$$

Proof. The proof can easily be obtained by induction. ∎

COROLLARY. *Assume as in the above Proposition, and that* $\lambda=\mu$. *Then* Y_i *is distributed like the maximum of* $i+1$ *independent* $E(\lambda)$ *variables.*

A proof of a similar result can be found in Ross (1982, p. 144). The application to the branch length of a coalescent tree seems to be new; however, in Tavaré (1984, p. 153) one can find an indication of the result.

The formula for $g_k(x)$ can be found by conditioning on branch length

$$g_k(x) = \int_0^\infty b \exp(-xb)\, f_k(b)\, db,$$

and applying 3.432 in Gradshteyn and Ryzhik (1994). Similarly for the mean value of g_k,

$$\begin{aligned}\int_0^\infty x g_k(x)\, dx &= \int_0^\infty \int_0^\infty xb \exp(-xb)\, f_k(b)\, dx\, db \\ &= \int_0^\infty \frac{1}{b} f_k(b)\, db,\end{aligned}$$

and then apply formula 3.411 (19) in Gradshteyn and Ryzhik (1994).

3: The expression of the probability that the tree of the first position is lower than the tree of a position just to the right of the first recombination point can be derived in the following way:

If the second tree is higher than the first, the recombined sequence cannot coalesce with any ancestral sequence until all positions $[0, p[$, p denoting the recombination point, have found a common ancestor. Let w_i be the time while there are i ancestral sequences to the subsequences consisting of positions $[0, p[$ in the sample. Put $l=(l_1, l_2, \ldots, l_k)$ with $l_1=0$, $l_i=\sum_{j=2}^{i} jw_j$, and let k be sample size. l_i denotes the total branch length while there are at most i ancestral sequences to positions $[0, p[$. Hence

$$P(H_1<H_2) = \int_{R_+^{k-1}} P(H_1<H_2 \mid l)\, f(l)\, dl,$$

where $f(l)$ denotes the density function of l. If there are i lineages, there are i possible branches where a recombination event can occur, and conditional on the location x of the recombination point the probability takes the form

$$\begin{aligned}P(H_1<H_2) = \int_{R_+^{k-1}} \sum_{i=2}^{k} \int_0^{w_i} \frac{i}{l_k} \\ \times P(H_1<H_2 \mid l, i, x)\, f(l)\, dx\, dl,\end{aligned}$$

where $w_i = (1/i)(l_i - l_{i-1})$ and $f(x \mid l_k) = 1/l_k$ is the uniform density of the location of the recombination point. But

$$P(H_1 < H_2 \mid l, i, x) = e^{-(l_i - l_{i-1} - ix)} \prod_{j=2}^{i-1} e^{-(l_j - l_{j-1})} = e^{ix - l_i},$$

with the jth term being the probability that the recombined sequence does not coalesce with any of the ancestral sequences. Inserting the above expression in the integral and evaluating yield

$$\begin{aligned} P(H_1 < H_2) &= \frac{1}{2}(k-1) \int_0^\infty \frac{1}{l_k} (1 - e^{-l_k/2})^{k-1} e^{l_k/2} \, dl_k \\ &= \frac{1}{2}(k-1) \sum_{i=1}^{k} (-1)^i \\ &\quad \times \frac{(k-1)!}{(i-1)!\,(k-i)!} \{\log(i+1) + \log(i)\}. \end{aligned}$$

The last equality is due to formula 3.411 (19) in Gradshteyn and Ryzhik (1994).

In a sample of size 2, $P(H_1 = H_2) = P(H_1 > H_2)$, and hence $2P(H_1 = H_2) = 1 - P(H_1 < H_2) \approx 1 - 0.549 = 0.451$.

4: The derivation of the expression of h_2 consists in evaluating a series of integrals, the first being the following conditional probability:

$$\begin{aligned} P(H_2 = s \mid H_1 = s) &= \frac{1}{2} \int_0^s \frac{1}{s} (1 - e^{-2(s-x)}) \, dx \\ &= \frac{1}{2} - \frac{1}{4s}(1 - e^{-2s}). \end{aligned}$$

The location x of the recombination point has conditional density $1/s$, and the probability given x that the two recombined sequences coalesce is $\frac{1}{2}(1 - e^{-2(s-x)})$. Similarly one obtains

$$P(H_2 \leqslant t < s \mid H_1 = s) = \frac{t}{2s} - \frac{1}{4s}(1 - e^{-2t}),$$

$$P(s < H_2 \leqslant t \mid H_1 = s) = \frac{1}{2s}(1 - e^{-2s})(1 - e^{s-t}).$$

The density function of H_2 can be derived using the above expressions. Obviously

$$P(H_2 \leqslant t) = P(H_2 \leqslant t, H_1 = H_2) + P(H_2 \leqslant t, H_1 \neq H_2).$$

Both probabilities on the right side can be calculated conditioning on H_1, and using the above conditional expressions,

$$\begin{aligned} F_1(t) = P(H_2 \leqslant t, H_1 = H_2) &= \int_0^t \frac{1}{2} e^{-s} \\ &\quad - \frac{1}{4s} e^{-s}(1 - e^{-2s}) \, ds, \end{aligned}$$

and

$$\begin{aligned} F_2(t) = P(H_2 \leqslant t, H_1 \neq H_2) &= \int_t^\infty \frac{t}{2u} e^{-u} \\ &\quad - \frac{1}{4u} e^{-u}(1 - e^{-2t}) \, du + \int_0^t \frac{1}{2} e^{-u} \\ &\quad + \frac{1}{4u} e^{-u}(1 - e^{-2u}) - \frac{1}{2u} e^{-t}(1 - e^{-2u}) \, du. \end{aligned}$$

Denote by f_1 and f_2 the derivatives of F_1 and F_2 with respect to t, respectively. Then h_2 can be written in the form

$$h_2(x) = \int_0^\infty 2t \, e^{-2tx} \{ f_1(t) + f_2(t) \} \, dt,$$

since sequence length until a recombination point is exponentially distributed with parameter $2t$, twice the height of the tree. Interchanging the order of integration in the double integral expression of h_2, and evaluating using formula 3.411 (19) in Gradshteyn and Ryzhik (1994), one obtains

$$\begin{aligned} h_2(x) = \frac{1}{(2x+1)^2} \Bigg\{ &\frac{4x+3}{(2x+2)^2} \log(2x+3) \\ &+ \frac{4x+1}{(2x)^2} \log(2x+1) - \frac{1}{2x(x+1)} \Bigg\}. \end{aligned}$$

5: The expected values of H_2, $H_2 \mid H_1 = s$ and the distribution of H_2 given $H_1 = s$ for small values of s are easily computed using results in **6** above.

ACKNOWLEDGMENTS

Anne-Mette Krabbe Pedersen is thanked for reading and commenting on the manuscript. J. H. was supported by the Danish Research Council under Grant SNF 94-0163-1.

REFERENCES

Ethier, S. N., and Griffiths, R. C. 1990. On the two-locus sampling distribution, *J. Math. Biol.* **29**, 131–159.

Gradshteyn, I. S., and Ryzhik, M. 1994. "Table of Integrals, Series, and Products" (A. Jeffrey, Ed.), 5th ed., Academic Press, San Diego.

Griffiths, R. C., and Marjoram, P. 1996. Ancestral inference from samples of DNA suquences with recombination, *J. Comp. Biol.* **3/4**, 479–502.

Griffiths, R. C., and Marjoram, P. 1997. An ancestral recombination graph, *in* "Progress in Population Genetics and Human Evolution" (P. Donnelly and S. Tavaré, Eds.), pp. 257–270, IMA Volumes in Mathematics and Its Applications, Vol. 87, Springer-Verlag, Berlin.

Griffiths, R. C., and Tavaré, S. 1997. Computational methods for the coalescent, *in* "Progress in Population Genetics and Human Evolution" (P. Donnelly and S. Tavaré, Eds.), pp. 165–182, IMA Volumes in Mathematics and Its Applications, Vol. 87, Springer-Verlag, Berlin.

Hudson, R. R. 1983. Properties of the neutral allele model with intergenic recombination, *Theor. Popul. Biol.* **23**, 183–201.

Hudson, R. R., and Kaplan, N. 1985. Statistical properties of the number of recombination events in the history of DNA sequences, *Genetics* **111**, 147–164.

Kaplan, N., and Hudson, R. R. 1985. The use of sample genealogies for studying a selectively neutral *m*-loci model with recombination, *Theor. Popul. Biol.* **28**, 382–396.

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

Ross, S. M. 1982. "Stochastic Processes," Wiley, New York.

Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models, *Theor. Popul. Biol.* **26**, 119–164.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination, *Theor. Popul. Biol.* **7**, 256–276.

Wiuf, C., and Hein, J. 1997. On the number of ancestors to a DNA sequence, *Genetics* **147**, 1459–1468.