AN ANCESTRAL RECOMBINATION GRAPH

ROBERT C. GRIFFITHS, MONASH UNIVERSITY
PAUL MARJORAM, MONASH UNIVERSITY *

Abstract. This paper describes a model of a gene as a continuous length of DNA represented by the interval $[0,1]$. The ancestry of a sample of genes is complicated by possible recombination events, where a gene can have two parent genes.

An analogue of Kingman's coalescent process, in which the ancestry of a sample of genes at a single locus is described by a stochastic binary tree, is a stochastic ancestral recombination graph, with vertices where coalescent or recombination events occur. All the information about ancestry is contained in this graph.

The sample DNA lengths have marginal ancestral trees at each point in $[0,1]$ which are imbedded in the graph. An upper bound is found for the expected number of distinct most recent common ancestors of these trees, and the expected maximum waiting time to these ancestors.

Key words. Coalescent process, Genealogical process, Population genetics, Recombination graph.

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

1. Introduction.

An important ancestral process in population genetics is the coalescent process described in [10]. This represents the ancestry of a sample of n genes as a stochastic binary tree. A realization for 5 genes is illustrated in Figure 1.1. Measuring time backwards the number of ancestors $\{\xi_n(t), t \geq 0\}$ is a death process with $\xi_n(0) = n$ and rates $\mu_k = k(k-1)/2, k = n, \ldots, 2$. Vertices occur where two lines have a common ancestor. The rates are sufficiently fast to properly define a process $\{\xi_\infty(t), t \geq 0\}$ with an entrance boundary at infinity. The process has an absorbing state at unity, when the most recent common ancestor (MRCA) of the sample is found. The process arises as a limit from an ancestral process in a classical Wright-Fisher model with a fixed population size $2N$, and discrete generations when time is measured in units of $2N$ generations, and $N \to \infty$. The contents of a generation are formed by the $2N$ children choosing their parents at random from the previous generation. At a finer level a gene in this model might be thought of as a piece of DNA which does not break up along its ancestral lines (that is, there is no recombination).

This paper describes an analogue of the coalescent process when recombination is possible, the *ancestral recombination graph.* Such a graph for a two-locus model is described in [4].

It is convenient to represent a gene, thought of as a length of DNA, by the unit interval $[0,1]$. In a discrete Wright-Fisher model children in a generation choose one parent, with probability $1-r$, or two parents, with

* Mathematics Department, Monash University, Clayton 3168, AUSTRALIA, Supported in part by Australian Research Grant A19131517.

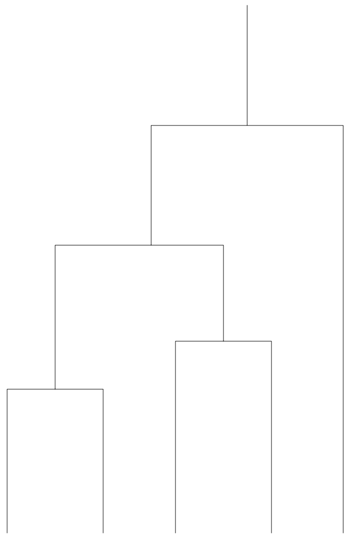

FIG. 1.1. *Coalescent tree.*

probability r, when a recombination event, looking back in time, takes place. If recombination occurs a position for the break point, Z, is chosen (independently from other break points) according to a given distribution, and the child gene is formed from the lengths $[0, Z]$ and $[Z, 1]$ from the first and second parents. Both of the parents are regarded as ancestors of any gene in a (forward) line of the child. Again time is measured in units of $2N$ generations and $N \to \infty$. The recombination rate per gene per generation r is scaled by holding $\rho = 2Nr$ fixed.

Particular cases for the break distribution are: Z is constant at 0.5, giving rise to a two-locus model; Z is discrete, taking values $\frac{1}{m}, \ldots, \frac{m-1}{m}$, giving rise to a m-locus model; and Z has a continuous distribution on $[0, 1]$, where breaks are possible at any point in $[0, 1]$.

A two-locus model is studied in [3], and finite-locus and continuous locus models in [5], [9].

Figure 1.2 illustrates a recombination graph for a sample of n genes. Looking back in time, coalescences occur when two edges join to a vertex, and recombination occurs when one edge joins to two. Positions $Z_1, Z_2, \ldots$ where breaks occur are labeled on the graph. The number of ancestors of the sample back in time $\{\xi(t), t \geq 0\}$ is a birth and death process with rates $\mu_k = k(k-1)/2$ and $\lambda_k = k\rho/2$. Because of the quadratic death rate compared to the linear birth rate, with probability 1 there is a MRCA in the graph. As with the coalescent process, the the process can have an entrance boundary at infinity. It is implicit that the process is defined backward in time to negative infinity. Usually the graph is only of interest

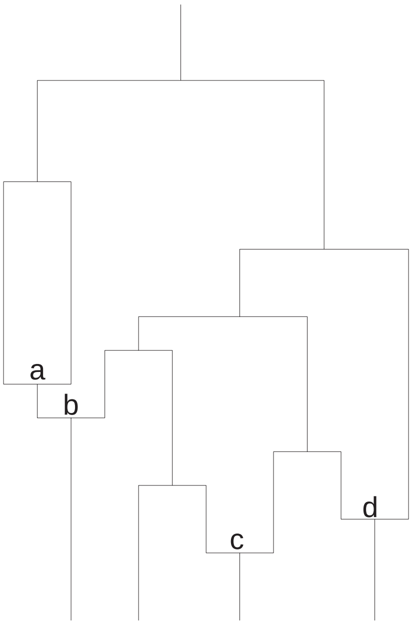

FIG. 1.2. *Ancestral recombination graph.*

to the MRCA, since the whole ancestry of the sample is determined by then. However if the ancestry of a single individual is followed back in time a graph is generated back to the next MRCA. Hitting a single edge (the MRCA) is a recurrent event, and the graphs between such hits are independent and identically distributed. Within a graph for n individuals a there is a subgraph for each individual which is either a single line to the grand MRCA, or is a graph which can be furthur decomposed into identically distributed subgraphs between hits of single edges. Subgraphs of the ancestry of n_0 of the n genes are consistent in the sense of being distributed as a recombination graph of a sample of n_0 genes.

It is shown in [4] that the expected waiting time to the grand MRCA from a sample of $n > 1$ genes is

$$2\rho^{-1} \int_0^1 \frac{1 - x^{n-1}}{1 - x} \big(e^{\rho(1-x)} - 1\big) dx. \tag{1.1}$$

The formula (1.1) also holds for the entrance boundary $n = \infty$. The expected time to a recombination event on a single line is $2/\rho$, therefore the expected time to generate a genealogy from a single gene to the next single edge is $2\big(e^{\rho} - 1\big)$.

The number of recombination vertices in the graph from a sample of n genes is distributed as the number of steps right in a random walk $\{\zeta_t, t = 0, 1, \ldots\}$ which starts at n, moves from k to $k + 1$ with probability $\rho/(\rho +$

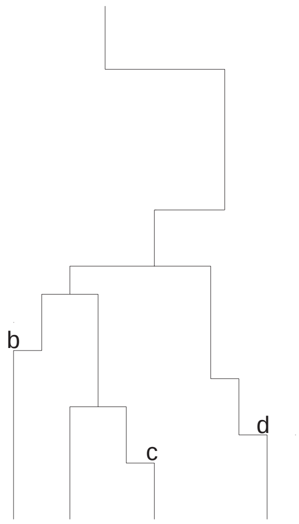

FIG. 1.3. *Marginal tree $\mathcal{T}(x)$, when $x > b$ and $x < c, d$.*

$k-1)$ or to $k-1$ with probability $(k-1)/(\rho+k-1)$, and has an absorbing state at 1. This is clearly true from the rates μ_k, λ_k. It is shown in [1] that the number of recombination vertices has a probability generating function $P_n(s) = Q_n(s)/Q_1(s)$, where

$$Q_n(s) = \int_0^1 x^{\rho(1-s)-1}(1-x)^{n-1}e^{-\rho s(1-x)}dx. \tag{1.2}$$

Each point $x \in [0,1]$ has a coalescent tree $\mathcal{T}(x)$ associated with its ancestry. These trees are imbedded in the recombination graph. To obtain $\mathcal{T}(x)$ trace from the leaves of the graph upward toward the MRCA in the graph. If there is a recombination vertex with label z, take the left path if $x \leq z$, or right path if $x > z$. The MRCA in $\mathcal{T}(x)$ may occur in the graph before the grand MRCA. Figure 1.3 shows an example of $\mathcal{T}(x)$ when $x > b$ and $x < c, d$.

Since there are a finite number of recombination events in the graph, there are only a finite number of trees in $\{\mathcal{T}(x); x \in [0,1]\}$. There are potentially 2^R, if R recombination events occur, but some trees may be identical, or may not exist, depending on the ordering of the recombination break points. Recombination does not affect the marginal history of individual points, so for each $x \in [0,1]$, $\mathcal{T}(x)$ is distributed as a coalescent tree. Of course different trees share edges in the graph, and are not independently distributed.

Figure 1.4 contains all possible trees corresponding to the recombination graph in Figure 1.2. Trees 1 and 9 are identical; the other trees are all distinct. If $b > a$ then all trees exist as marginal trees in the graph,

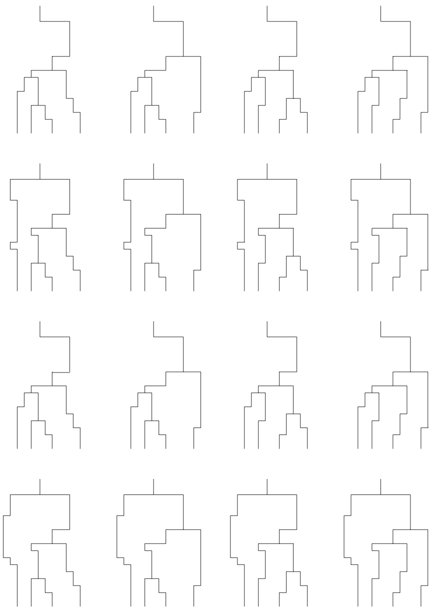

FIG. 1.4. *All possible marginal trees for the graph in Figure 1.2.*

otherwise if $b < a$ trees in Figure 1.4 with the right edge at vertex a do not exist as marginal trees.

Ancestor genes may only have part of their gametic material in common with the sample genes. It is even possible that some ancestor genes in the graph contain no material in common. A point x on an ancestor represented by an edge e in the graph has ancestral material in common with the sample if and only if e is included in $\mathcal{T}(x)$. Thus the subset of $[0,1]$ over which that ancestor has ancestral material in common with the sample is $\mathcal{P}_e = \{x; \mathcal{T}(x) \ni e, x \in [0,1]\}$. $\mathcal{P}_e$ is a union of a finite number of intervals, whose endpoints are a subset of the positions where recombination breaks have occured. If e and f are two edges, and $e \vee f$ denotes a coalesced edge from e and f, then $\mathcal{P}_{e\vee f} = \mathcal{P}_e \bigcup \mathcal{P}_f$. If a recombination break occurs at z, to an edge e, then the left and right hand edges from e in the graph are $\mathcal{P}_e \bigcap [0,z]$ and $\mathcal{P}_e \bigcap [z,1]$.

The embedded process describing the ancestral material in common with the sample is interesting. This Markov process has state $\{\mathcal{P}_{e_1}, \dots, \mathcal{P}_{e_k}\}$ while there are k ancestors, and makes transitions according to how the random walk ζ_t moves, by either choosing two edges e, f at random to coalesce, or an edge chosen at random to split at Z, a recombination break point.

Edges $e_1, \dots, e_k$ partition the interval $[0,1]$ of each of the n sample genes by their ancestral material in common. That is, $\mathcal{Q}_1, \dots, \mathcal{Q}_n$ are partitions of $[0,1]$, $\mathcal{Q}_i = \{\mathcal{Q}_{ij}, j = 1,\dots,k\}, i = 1,\dots,n$ such that $\mathcal{Q}_{ij}$ is the material on edge e_j in common with sample gene i. Then also $\mathcal{P}_{e_j} = \bigcup_{i=1}^n \mathcal{Q}_{ij}$. This way of thinking about ancestors is analogous to Kingman's scheme of labeling edges in a coalescent by an ancestral partition of the sample genes $1,\dots,n$.

For example the imbedded ancestral partitions corresponding to the coalescent tree in Figure 1.1, with individuals labeled 1 to 5 from the left, are $\{\{1\},[2],\{3\},\{4\},\{5\}\}, \{\{1,2\},\{3\},\{4\},\{5\}\}, \{\{1,2\},\{3,4\},\{5\}\}, \{\{1,2,3,4\},\{5\}\}, \{1,2,3,4,5\}$.

In the recombination graph each ancestor can be labeled by which sample genes, and subsets of material it is ancestral to. The sample is represented as $\bigotimes_{i=1}^n (i,[0,1])$ and the ancestral partition is of this set. That is, while k ancestors, the jth ancestor would be labeled by $\bigotimes_{i=1}^n (i,\mathcal{Q}_{ij})$. The material on the n sample genes is partitioned by this labeling, as

$$\bigotimes_{i=1}^n (i,\mathcal{Q}_{ij}) \bigcap \bigotimes_{i=1}^n (i,\mathcal{Q}_{i\ell}) = \phi, j \neq \ell, \text{ and}$$
$$\bigcup_{j=1}^k \bigotimes_{i=1}^n (i,\mathcal{Q}_{ij}) = \bigotimes_{i=1}^n (i,[0,1]).$$

2. Recombination events and MRCAs. In the following it is assumed that the recombination break distribution of Z is uniform in $[0,1]$, unless otherwise mentioned.

THEOREM 2.1. *Let $R_{n,x,\delta}$ be the number of recombination events in $[x-\delta/2, x+\delta/2]$ before the MRCA at x. Define $h_n(x)$, the recombination density at $x \in [0,1]$, by*

$$h_n(x) = \lim_{\delta \to 0} \delta^{-1} P(R_{n,x,\delta} = 1) \ . \tag{2.1}$$

Then

$$h_n(x) = \sum_{k=1}^{n-1} \frac{\rho}{k} \ , \tag{2.2}$$

and

$$\lim_{\delta \to 0} \delta^{-1} P(R_{n,x,\delta} > 1) = 0. \tag{2.3}$$

Proof. When there are k ancestors of the sample fragments $[x-\delta/2, x+\delta/2]$ then coalescence occurs at a rate $k(k-1)/2$, and recombination at rate $k\rho\delta/2$. The probability of no recombination events in $[x-\delta/2, x+\delta/2]$ in the graph, while k ancestors, is the probability that coalescence occurs before recombination,

$$\frac{k(k-1)/2}{k(k-1)/2 + k\rho\delta/2} = \frac{k-1}{k-1+\rho\delta}.$$

The probability of no recombination events in $[x-\delta/2, x+\delta/2]$ before the MRCA at x is therefore

$$\prod_{k=2}^{n} \frac{k-1}{k-1+\rho\delta} \ ,$$

and

$$\lim_{\delta\to 0} \delta^{-1} P(R_{n,x,\delta} \geq 1) = \lim_{\delta\to 0} \delta^{-1}\Big(1 - \prod_{k=2}^{n} \frac{k-1}{k-1+\rho\delta}\Big) = \sum_{k=1}^{n-1} \frac{\rho}{k} \ .$$

Whatever the number of ancestors the recombination rate is proportional to δ, so $P(R_{n,x,\delta} > 1) = o(\delta^2)$ as $\delta \to 0$, and (2.1), (2.3) follow. □

Corollary 2.2. *The expected number of recombination events before the marginal MRCAs along the genes is* $\int_0^1 h_n(x)dx = \sum_{k=1}^{n-1} \frac{\rho}{k}$.

The result in this corollary is derived in [7].

Let $A_n(x,t)$ be the number of distinct ancestors of the sample of n, in $\mathcal{T}(x)$, at time t back. $A_n(\cdot,t)$ is a random step function. Eventually $A_n(x,\tau) = 1, x \in [0,1]$ at the time τ when the MRCA of the graph is reached. If a cross-section of the graph at time t back has edges $e_1,\dots,e_m$, then $A_n(x,t) = |\{e_i; x \in \mathcal{P}_{e_i}, i = 1,\dots,m\}|$. From one locus coalescent theory, $E(A_n(x,t)) = E(\xi_n(t))$, not depending on x.

Of particular interest is the number of distinct MRCAs of a sample in the marginal trees at points along the genes,

$$|\{v(x); v(x) \text{ is the root of } \mathcal{T}(x), x \in [0,1]\}| \ .$$

Clearly if there are R recombination events in the recombination graph then there can be at most $R+1$ MRCAs of a sample along the genes.

For example suppose $a < b < c < d$ for the graph in Figure 1.1. Then for the sample $(0,b] \cup (d,1]$ has the grand MRCA as the MRCA, and $(b,c]$, $(c,d]$ have other distinct MRCAs; a total of 3 distinct MRCAs.

The next theorem shows that multiple recombination events are rare in the ancestral lines of a sample of n genes as $n \to \infty$.

Theorem 2.3. *Let R_n be the number of recombination events affecting the ancestory of a sample of n genes, R_n^0 the number of recombination events which occur to ancestors not having a previous recombination event in their lineage from the sample, and R_n^1 the number of recombination events to the grand first common ancestor. Then*

$$R_n^0 \leq R_n \leq R_n^1, \tag{2.4}$$

$$\textstyle\sum_{j=2}^{n} \frac{\rho}{j+\rho-1} \leq E(R_n^0) \leq 1 + \sum_{j=2}^{n} \frac{\rho}{j+\rho-1}, \tag{2.5}$$

$$E(R_n^1) = \rho \int_0^1 \frac{1-(1-x)^{n-1}}{x} e^{\rho x} dx. \tag{2.6}$$

As $n \to \infty$, all three of $E(R_n^0)$, $E(R_n)$, $E(R_n^1)$ are asymptotic to $\rho\log(n)$, and $E(R_n^1 - R_n^0)$ is uniformly bounded above for $n = 2, 3, \ldots$.

Proof. For R_n^0 disregard the genealogy of any genes once they have been involved in recombination events, and consider the coalescence of ancestors where no recombination has taken place. Then ancestral lines in this modified coalescent are lost by coalescence or recombination at rates $k(k-1)/2$ and $k\rho/2$. The probability that the kth line is lost by recombination is $\rho/(k+\rho-1)$, $k = n, \ldots, 1$. The last recombination event, if it occurs when one line, may, or may not, be before the MRCAs of the genes. This accounts for the inequality (2.5). The formula (2.6) for $E(R_n^1)$, and the result $E(R_n^1) \sim \rho\log(n)$ are derived in [1].

$E(R_n^1 - R_n^0)$ is shown to be uniformly bounded above by the following.

$$\begin{aligned}
E(R_n^1 - R_n^0) &= \rho\int_0^1 \frac{1-(1-x)^{n-1}}{x} e^{\rho x}dx - E(R_n^0) \\
&= \rho\int_0^1 \frac{1-(1-x)^{n-1}}{x}(e^{\rho x}-1)dx \\
&\qquad + \rho\sum_{j=0}^{n-2}\int_0^1 (1-x)^j dx - E(R_n^0) \\
&\le \rho\int_0^1 \frac{1-(1-x)^{n-1}}{x}(e^{\rho x}-1)dx \\
&\qquad + \rho\sum_{j=2}^{n}\Big(\frac{1}{j-1} - \frac{1}{j+\rho-1}\Big) \\
&\le \rho\int_0^1 \frac{e^{\rho x}-1}{x}dx + \rho^2\sum_{j=1}^{\infty}\frac{1}{j(j+\rho)} \\
&< \infty.
\end{aligned}$$

□

This theorem provides an estimate for the expected logarithm of the number of distinct trees in the recombination graph, $\phi_n = E\log|\{\mathcal{T}_n(x); x \in [0,1]\}|$. There are two possible paths to the grand MRCA in the graph along lines where there has been exactly one recombination event. Therefore

$$(2.7) \qquad (\log 2)E(R_n^0) \le \phi_n \le (\log 2)E(R_n^1)$$

and $\phi_n \sim \rho(\log 2)(\log n)$ as $n \to \infty$.

THEOREM 2.4. *Let $p_n(x)$ be the probability that the MRCAs of $\mathcal{T}(x-)$ and $\mathcal{T}(x+)$ are identical in a recombination graph of n genes, given that a recombination event has occurred at x before the MRCA of $\mathcal{T}(x)$. Then*

$$(2.8) \qquad \begin{aligned} p_n(x) &= 1 - \frac{n^2+n-2}{n(n+1)\sum_{k=1}^{n-1}\frac{1}{k}} \\ &\sim 1 - (\log(n))^{-1} \text{ as } n \to \infty. \end{aligned}$$

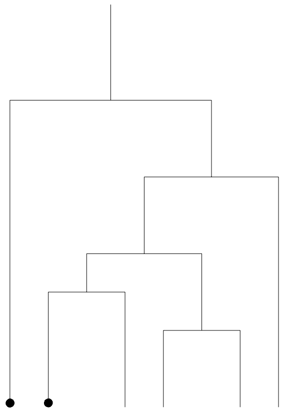

FIG. 2.1. *Ancestral tree of $\ell+1$ genes after recombination.*

Proof. It is sufficient to consider just the one recombination event, since others do not affect $p_n(x)$. Suppose that this event occurs while there are ℓ ancestors of the sample.

The two MRCAs of the sample to the left and right of x are distinct if and only if one of the last two lines in the coalescence of the $\ell+1$ genes after the recombination event is a single ancestor line of one of the genes involved in the recombination, and not an ancestor line of any of the other ℓ genes. Figure 2.1 illustrates this, with the two genes just after recombination shown by dots.

This occurs if and only if the subtree of the ℓ genes coalesces first, and so the probability of distinct ancestors is

$$2 \times \frac{\ell(\ell-1)}{(\ell+1)\ell} \cdot \frac{(\ell-1)(\ell-2)}{\ell(\ell-1)} \cdots \frac{2}{3} = \frac{4}{\ell(\ell+1)}.$$

The probability that there is one recombination event in $[x-\delta/2, x+\delta/2]$ while ℓ lines is

$$1 - \frac{\ell-1}{\ell-1+\rho\delta} + o(\delta^2) = \frac{\rho\delta}{\ell-1} + o(\delta^2),$$

so the conditional probability of a recombination event while ℓ lines is

$$\frac{1}{(\ell-1)\sum_{k=1}^{n-1}\frac{1}{k}}.$$

Finally, $p_n(x)$ is the sum of the probabilities, over ℓ, that the MRCAs of the sample of n genes immediately to the left and right of x are identical, and that the recombination event occurred while ℓ ancestors, given that a recombination event has occurred at x. That is

$$\begin{aligned} p_n(x) &= \sum_{\ell=2}^{n}\Big(1-\frac{4}{\ell(\ell+1)}\Big)\frac{1}{(\ell-1)\sum_{k=1}^{n-1}\frac{1}{k}} \\ &= 1-\frac{n^2+n-2}{n(n+1)\sum_{k=1}^{n-1}\frac{1}{k}}. \end{aligned}$$

□

THEOREM 2.5. *Let α_n be the expected number of distinct MRCAs of a sample of n genes. Then*

$$\begin{aligned} \alpha_n &\le 1+\Big(1-\frac{2}{n^2+n}\Big)\rho \qquad (2.9) \\ &\le 1+\rho, \quad \text{for } n=2,3,\dots. \end{aligned}$$

(2.10)

Proof.

$$\begin{aligned} \alpha_n &\le 1+\mathrm{E}(\text{Number of changes of ancestor along the genes}) \\ &= 1+\int_0^1 \mathrm{P}(\text{Change of ancestor at } x \mid \text{Recombination at } x)h_n(x)dx \\ &= 1+\int_0^1 (1-p_n(x))h_n(x)dx \\ &= 1+\Big(1-\frac{2}{n^2+n}\Big)\rho. \end{aligned}$$

□

It is interesting that α_n is uniformly bounded in n, even though the expected number of recombination events before the MRCAs along the genes is asymptotic to $\rho\log(n)$.

3. Waiting times to MRCAs. Let $\{W_n(x),\ 0\le x\le 1\}$ denote the collection of waiting times until the MRCAs at positions x for a sample of n genes. $W_n(x)$ is a random step function, depending on the number of recombination events in the ancestry.

Marginally

$$W_n(x)=T_n(x)+\cdots+T_2(x),$$

where $\{T_k(x),\ k=2,\dots n\}$ are the times while k ancestors, distributed as mutually independent exponential random variables with means

$\{\frac{2}{k(k-1)},\ k = 2, \dots n\}$, and $E(W_n(x)) = 2\big(1 - \frac{1}{n}\big)$.
Although the marginal distribution of $W_n(x)$ does not depend on x or the recombination history of the sample, the distribution of $\{W_n(x),\ 0 \le x \le 1\}$ does.

THEOREM 3.1. *Let* $W = \max_{0 \le x \le 1} W_n(x)$, *then*

$$\begin{aligned} E(W) &\le 2 + \rho \frac{n^2 + n - 2}{2n(n+1)} \\ &\le 2 + \frac{\rho}{2}, \quad \text{for } n = 2, 3, \dots . \end{aligned} \tag{3.1}$$

Proof. Let R_n be the number of recombination events before the MRCAs of the genes, occurring at positions $x_1, \dots, x_{R_n}$. Define $x_0 = 0,\ x_{R_n+1} = 1$ for convenience. Let W_i be the time to the MRCA for the interval $[x_i, x_{i+1}]$. Then

$$\begin{aligned} W &= \max_{i=0,\dots,R_n} W_i \\ &\le W_0 + \sum_{i=1}^{R_n} (W_i - W_{i-1}) \vee 0, \text{ and} \\ E(W) &\le 2\Big(1 - \frac{1}{n}\Big) + E(R_n) E\Big((W_1 - W_0) \vee 0\Big). \end{aligned}$$

$E(W_0) = 2\Big(1 - \frac{1}{n}\Big)$, since W_0 is the marginal waiting time to the MRCA at $x_0 = 0$. Now $P(W_1 > W_0) = \frac{1}{2}(1 - p_n(x_1))$, and $E(W_1 - W_0 | W_1 > W_0) = 1$, since $W_1 - W_0$ is the time taken for the last two lines of the genealogy at x_1 to coalesce, given $W_1 > W_0$. Thus

$$\begin{aligned} E(W) &\le 2\Big(1 - \frac{1}{n}\Big) + \frac{1}{2} E(R_n)(1 - p_n(x_1)) \\ &= 2\Big(1 - \frac{1}{n}\Big) + \rho \frac{n^2 + n - 2}{2n(n+1)} \\ &\le 2 + \frac{\rho}{2}. \end{aligned}$$

□

Bounds for α_n and $E(W)$, although derived for a uniform break distribution, carry through for any continuous distribution on $[0, 1]$.

4. Mutations on the graph. Let $\{V_n(x), x \in [0, 1]\}$ denote the collection of edge lengths of the marginal trees until the MRCAs at positions x for a sample of n genes. Using the notation in Section 3

$$V_n(x) = nT_n(x) + \cdots + 2T_2(x).$$

$V_n(x)$ is a random step function. A picture (essentially) of a simulated realization of $V_{10}(x)$ is shown in [6]. In a model with mutations occurring along the edges of the recombination graph according to a Poisson process of rate $\theta/2$, the total number of mutations occurring on lines with material in common to the sample is distributed as

$$N\Big(\frac{\theta}{2}\int_0^1 V_n(dx)\Big), \tag{4.1}$$

where $N(\cdot)$ is a Poisson process of unit rate.

The mean number of mutations is

$$\mu_m = \theta \sum_{j=1}^{n-1} \frac{1}{j}, \tag{4.2}$$

and variance

$$\begin{aligned} \sigma_m^2 &= E\Big((\theta/2)^2 \int_0^1 \int_0^1 V_n(dx)V_n(dy)\Big) + \mu_m - \mu_m^2 \\ &= (\theta^2/2)\int_0^1 (1-z)Q_n(z;\rho z)dz + \mu_m - \mu_m^2, \end{aligned} \tag{4.3}$$

where $Q_n(z;\rho z)$ is the product of edge lengths at two points distance z apart. This quantity is distributed as the product of edge lengths in a two locus model with recombination rate ρz between the loci.

Let M be the number of mutations and λ be the random variable $\frac{\theta}{2}\int_0^1 V_n(dx)$. Then (4.3) follows from

$$\begin{aligned} \sigma_m^2 &= E(M(M-1)) + \mu_m - \mu_m^2 \\ &= E(E(M(M-1))|\lambda) + \mu_m - \mu_m^2 \\ &= E(\lambda^2) + \mu_m - \mu_m^2. \end{aligned}$$

The formulae (4.2), (4.3) were derived by [5], (4.3) can be expressed as

$$\sigma_m^2 = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \frac{1}{2}\frac{\theta^2}{\rho^2}\int_0^\rho (\rho - z) f_n(z)dz, \tag{4.4}$$

where $f_n(z)$ is the covariance of the edge lengths in a 2-locus model with recombination rate ρ and sample size n. In [5], time is measured in twice our units, accounting for a factor of four different in front of the integral expression in (4.4).

A particular case for $n = 2$, the only explicit formula for $f_n(z)$, is

$$f_2(z) = \frac{4(\rho + 18)}{\rho^2 + 13\rho + 18}. \tag{4.5}$$

A model with mutation is obtained by specifying the distribution of the position on a gene where mutation takes place. Mutations which occur in material ancestral to the sample will be represented in the sample genes. Suppose mutation occurs according to a continuous distribution in $[0,1]$, and the label of a mutation is just the position where it occurs. An observed sample of n, then, is a collection of n sets of points where mutations have occurred. If there is a mutation at x_0, then some genes of the sample will contain the mutation, and others will not, being of the type of the MRCA at x_0.

There is an urn model representation of a two-locus sampling distribution in [1] which extends easily to the model in this paper where a gene is a length $[0,1]$ of DNA. The idea in this representation is to produce the relative order of coalescent, mutation, and recombination events in the imbedded process in the graph. Then the shape of graph is filled in later.

Let $\{M(t), t = 0,1,\dots\}$, $M(0) = n$ be a random walk on positive integers with absorbing state 1, and transition probabilities for $m \geq 2$,

$$m = \begin{cases} m-1, & \text{with probability } (m-1)/(m-1+\theta+\rho), \\ m, & \text{with probability } \theta/(m-1+\theta+\rho), \\ m+1, & \text{with probability } \rho/(m-1+\theta+\rho). \end{cases} \tag{4.6}$$

Suppose τ is the absorption time. Keep a record of $M(0),\dots,M(\tau)$. Begin with a sample of size 1 at τ, then construct samples at times $\tau-1,\dots,1$. Let $D(t) = M(t) - M(t-1)$, $t = \tau,\dots,1$. If $D(t) = -1$, then choose a sample member at random to duplicate. If $D(t) = 0$, then choose a sample member at random to mutate at a random position in $[0,1]$ according to a prescribed distribution. If $D(t) = 1$, then choose a pair at random to recombine at a position chosen according to a prescribed recombination break distribution. The sample at n at time 0 is distributed as a sample in the recombination graph.

REFERENCES

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

[2] Ethier, S.N. and Griffiths, R.C., *The neutral two-locus model as a measure-valued diffusion*, **22**, 773-786, Adv. Appl. Prob., (1991)

[3] Griffiths, R.C., *Neutral two-locus multiple allele models with recombination*, **19**, 169-186, Theoret. Popn. Biol., (1981).

[4] Griffiths, R.C., *The two-locus ancestral graph*, **18**, 100-117, Selected proceedings of the symposium on applied probability, Sheffield, 1989., Institute of Mathematical Statistics, *ed.* I.V. Basawa and R.L. Taylor, IMS Lecture Notes–Monograph Series, (1991).

[5] Hudson, R.R., *Properties of a neutral allele model with intragenic recombination*, **23**, 183-201,Theoret. Popn. Biol., (1983).

[6] Hudson, R.R., *Gene genealogies and the coalescent process*, **7**, 1-44, Oxford Surveys in Evolutionary Biology, *ed.* Futuyma, D. and Antonovics, J. (1991).

[7] Hudson, R.R. and Kaplan, N.L., *Statistical properties of the number of recombination events in the history of a sample of DNA sequences*, **111**, 147-164, Genetics, (1985).

[8] Hudson, R.R. and Kaplan, N.L., *The coalescent process in models with selection and recombination*, **120**, 831-840, Genetics, (1988).

[9] Kaplan, N.L. and Hudson, R.R., *The use of sample genealogies for studying a selectively neutral m-loci model with recombination*, **28**, 382-396, Theoret. Popn. Biol., (1985).

[10] Kingman, J.F.C., *The coalescent*, **13**, 235-248, Stoch. Proc. Applns., (1982).