

A Markov Chain Model of Coalescence With Recombination

Katy L. Simonsen* and Gary A. Churchill*,†

*Center for Applied Math, †Biometrics Unit, Cornell University, Ithaca NY 14853

October 14, 1996

Present Address of Corresponding Author:

Katy Simonsen
Department of Statistics
North Carolina State University
Raleigh NC 27695-8203

E-mail: simonsen@stat.ncsu.edu
Office: (919) 515-5769
Home: (919) 859-3644
Fax: (919) 515-7315

Abstract

Trees that describe the ancestry of DNA sequences sampled from a population may differ between loci because of genetic recombination. We seek to understand the relationship between such trees for loci that are linked with non-zero recombination rate. We consider a coalescent process model with recombination, as described by Hudson (1983; 1990). For two loci and a sample size of two sequences, a detailed analysis of this process yields the joint distribution of the two trees (one at each locus). A number of interesting results follow from this analysis, including the distribution of the number of recombination events in the history of the sample. For the general case of m loci and samples of size n , we describe an algorithm for simulating the tree building process. Because analytic results are difficult to obtain in this case, we use simulation to study properties of trees at multiple linked loci such as total tree time and number of recombination events.

1. INTRODUCTION

Most work on models of DNA sequence evolution has assumed either that there is no recombination, or that there is free recombination. In the first case, linked sites have exactly the same history, while in the latter case sites are essentially independent. The case where recombination rates are intermediate between these two extremes is more difficult to analyze. One example of the importance of recombination rate is its demonstrated correlation with levels of polymorphism in *Drosophila* (Begun and Aquadro, 1992). Models attempting to explain this correlation must be able explicitly to incorporate a variable recombination rate. We describe a model of sequence evolution that includes the recombination rate as a free parameter, and apply this model to study the effects of recombination on the sampling properties of DNA sequences.

Nucleotide substitution rates are generally believed to vary both locally, between sites in a gene sequence for example, and globally, across large regions of a genome. (Wakeley, 1993; Gaut and Weir, 1994; Wakeley, 1994; Gu, Fu and Li, 1995). There are many possible explanations for rate heterogeneity; including differences in selective constraint and unequal mutation rates. Another possible explanation that applies even under the assumptions of neutral evolution with homogeneous mutation rates is that the evolutionary time separating sites in a sample of sequences may vary due to recombination, since different loci may have different ancestries. We seek to understand the extent to which recombination can contribute to observed rate heterogeneity.

When highly variable “hot spots” or highly conserved regions are observed, it would be useful to know whether the effect could be explained simply by recombination, or whether a selective explanation is warranted. Although the problem of substitution rate heterogeneity is not addressed in this paper, it provides an important motivation for studying the effects of recombination on evolutionary time variation.

Suppose we have aligned DNA sequences from n randomly sampled individuals. Consider two nucleotide sites. If recombination has never separated these sites throughout the history of the sample, then they will have exactly the same genealogy. If, however, recombination has occurred between the two sites at some time in the history of the sample, then the genealogies at the two sites may differ. Imagine stepping along the sequence from site to adjacent site, examining the genealogies. A group of completely linked sites will all have the same genealogy. Then, passing a place where recombination has occurred, the genealogy will suddenly change. The process of changing genealogies as one moves from site to site is not Markovian (Bickeböllner and Thompson, 1996). However, the historical process that generates the linked genealogies is Markovian (in time). We exploit the Markov property to analyze and develop a general simulation technique for this process. Our goal is to understand the relationship between trees at different loci. To do this we will need to know the joint probability distribution of the genealogies at sets of adjacent loci. This distribution will be complex because it is defined on a space of trees with both topology and branch length components.

The structure of this paper is as follows. After giving some background we describe in Section 2 the simplest case of two loci and two sequences. This case is analyzed as a discrete-time Markov chain in Section 3, and as a continuous-time process in Section 4. The general case of n sequences and m loci is described in Sections 5 and 6. Section 7 gives the results of computer simulations of the general model.

1.1 Background

Griffiths (1981) considered the two-locus model and derived an expression for the covariance of the number of segregating sites at two loci in a sample of size 2, from which can be found the covariance of tree times at the two loci.

Hudson (1983; 1990) showed how to construct a set of correlated trees at adjacent loci by extending the coalescent process to incorporate recombination. This idea and the assumptions on which it is based form the basis for all the results of this paper. This technique has been very useful as a computational tool for simulating trees with recombination. However, the data structures and computer programs involved in constructing many correlated trees in this manner are extremely complex, and require large amounts of computer time and memory. Because it is of interest to explore through simulation deviations from the neutral model (Simonsen, Churchill and Aquadro, 1995), we would prefer a simpler computational tool for analyzing the properties of such trees. Such a tool is provided here.

Kaplan and Hudson (1985) pointed out that the coalescent process with recombination forms a Markov chain. They defined the states and transition probabilities of this Markov chain, which we review below, and from them derived a set of linear equations whose solution gives the covariance of the tree times at two loci, for a general sample size n . The number of equations to be solved grows as n^3 , but if the solutions for $n - 1$ are used and incorporated into the solution for n , the actual number of *new* equations to be solved each time grows only as n^2 . Numerical solutions were computed for particular values of the recombination rate (Kaplan and Hudson, 1985), but exact solutions (rational functions of the recombination rate) are practical at least up to $n = 10$. Covariances derived from this model have been used in estimates of sequence diversity (Pluzhnikov and Donnelly, 1996).

The problem of the number of recombination events in the history of a sample of size n was considered by Hudson and Kaplan (1985). They derived the probability of no recombination events, and estimated the expected number of events as a linear function of recombination rate. Hey and Wakeley (1996) used a related coalescent model to develop an estimator of the population recombination rate. We use the Markov chain model to derive the exact probability distribution of the number of recombination events for a sample of size two in the two-locus case.

2. MODEL: 2 LOCI AND 2 SEQUENCES

The problem of constructing trees with recombination is complicated by the large number of labeled tree topologies for n taxa. Let us consider first the case where the sample size is $n = 2$, so that there is only one

possible topology for the tree. In this case the only thing that can change from site to site is the height of the tree, that is, the amount of time from the present back to the common ancestor. Let us also restrict attention to $m = 2$ adjacent loci.

To analyze the joint distribution of tree heights at adjacent loci it will be necessary to analyze the tree-building process. The standard assumptions of the coalescent model apply, namely, large constant population size of N diploids, random mating, and non-overlapping generations. Time is measured in units of $2N$ generations. For a good review of the coalescent process, see (Hudson, 1990). The coalescent process proceeds backward in time from the present until the most recent common ancestor (MRCA) of the sample. A tree is constructed by joining the ancestral lineages of the sample when any two have a common ancestor.

Figure 1 shows a coalescent tree for two sequences and two loci, with no recombination. The process of tree-building begins at the bottom, where there are two separate sampled sequences at the present time. At some time back in the past, the two have a common ancestor, and a tree is constructed by coalescing the lineages. The time back to that common ancestor is exponentially distributed with mean 1 ($2N$ generations). With no recombination, the tree is the same at the two loci.

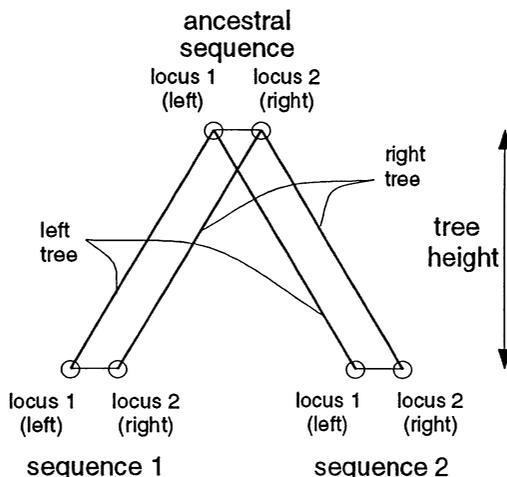


Figure 1: Two-locus tree with no recombination. The tree at the left locus is identical to the tree at the right locus.

Figure 2 shows an example of a coalescent tree for two sequences and two loci with one recombination event. The sampled sequence on the left is descended from a recombinant of two sequences. These two, along with an ancestor of the sampled sequence on the right, were all present in the population at the same time, and reach a common ancestor after two coalescent events. The process of building this tree is memoryless: the probability of two sequences having a common ancestor depends only on the population size at that time, and not on the history of their descendants. Therefore the process forms a Markov chain, as has been characterized by Kaplan and Hudson (1985). Since the notation used here differs from that in Kaplan and Hudson (1985), the Markov chain is described in detail below. The states of this Markov chain, as shown by Kaplan and Hudson (1985), are determined by the *number* and *type* of sequences present at any particular time t . The possible *types* of sequence are described below. Since ancestry may differ between loci, the word “lineage” will be used to refer to the ancestral history at a particular locus.

There are three different types of sequence possible in the tree. Those that are ancestral to the sample only at the left locus have one lineage, represented by a single line on the tree, and will be called *single left* or *type 1* sequences. There are two such sequences shown in Figure 2. The corresponding type for the

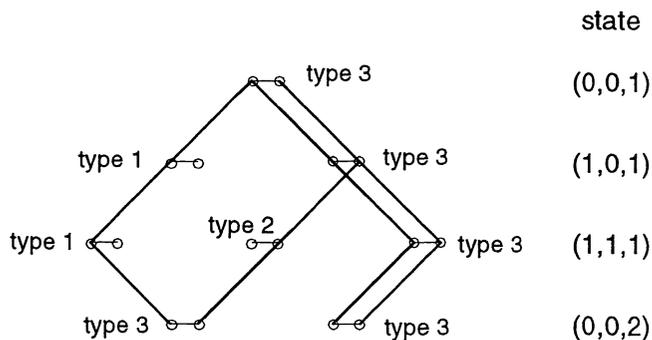


Figure 2: Two-locus tree with one recombination event. The type of each sequence is shown. Recombination is the most recent event, where one type 3 sample sequence is descended from a type 1 and a type 2 sequence. Two further coalescent events result in a common ancestor of type 3. The tree at the left locus is taller than the tree at the right locus. The series of states of the Markov chain corresponding to this tree are also shown.

right-hand locus is called *single right* or *type 2*, and there is one such sequence shown in Figure 2. *Type 3* or *double* sequences are ancestral to the sample at both loci, and have two lines passing through them on the tree. The sampled sequences are considered to be ancestral to themselves, and thus are always of type 3, as shown in Figure 2.

The state of the Markov chain at time t is defined by an ordered triple (i, j, k) , where i is the number of “single left” types, j is the number of “single right” types, and k is the number of “double” types present at time t . The total number of sequences present on the tree at time t is $i + j + k$. This ordered triple is subject to certain constraints, namely:

$$\begin{aligned}
 1 &\leq i + k \leq 2 \\
 1 &\leq j + k \leq 2 \\
 0 &\leq i, j, k
 \end{aligned}
 \tag{1}$$

The first two constraints result from the fact that at any time there must be at least one ancestor of the left (or right) locus, but at most $n = 2$ ancestors of the n left (or right) loci in the sample. The third constraint is simply that the number of each type of sequence is non-negative. These constraints result in nine possible states, which are given below.

The process moves from state to state through a series of recombination and coalescent events. Only certain transitions are possible at each state. For example, recombination (event **Re**) can only occur when a double type is present, that is, when $k > 0$. Several types of coalescent event are possible. Coalescence can occur between two double types (the event designated C_4), a left and a right type (C_2), a double type and a left (C_2^l) or right (C_2^r) single type, or between two left (C_1^l) or two right (C_1^r) single types. The nine states and the possible events at each state are listed in Table I.

At each state, the relative probabilities of the different types of coalescent events are simple functions of (i, j, k) . For example, a C_3^r event involves coalescence between one of the j right single types and one of the k double types, and therefore its probability is proportional to jk . There are a total of $\binom{i+j+k}{2}$ different ways to coalesce at each state. Let r be the per-generation probability of recombination between any two sequences, and let $R = 2Nr$, where the population is assumed to be of constant size N diploids. The seven types of events and their relative probabilities are as shown in Table II, where $x--$ means $x = x - 1$ and $x++$ means $x = x + 1$. The per-generation absolute probabilities of each event are the numerators of the values shown in Table II, divided by $2N$. We assume that N is sufficiently large that the probability of any two events occurring simultaneously is negligible.

Table I: States and Transition Probabilities of the Tree-Building Markov Chain

State	(i, j, k)	Possible Events	Probabilities	Next State
1	(0,0,2)	C_4	$1/(1+2R)$	9
		Re	$2R/(1+2R)$	2
2	(1,1,1)	C_2	$1/(3+R)$	1
		C_3^r	$1/(3+R)$	4
		C_3^l	$1/(3+R)$	5
		Re	$R/(3+R)$	3
		C_2	$2/3$	2
3	(2,2,0)	C_1^r	$1/6$	6
		C_1^l	$1/6$	7
		C_3^l	$1/(1+R)$	9
4	(1,0,1)	Re	$R/(1+R)$	6
		C_3^r	$1/(1+R)$	9
5	(0,1,1)	Re	$R/(1+R)$	7
		C_1^l	$1/3$	8
6	(2,1,0)	C_2	$2/3$	4
		C_1^r	$1/3$	8
7	(1,2,0)	C_2	$2/3$	5
		C_2	1	9
8	(1,1,0)	C_2	1	9
9	(0,0,1)			

Table II: Event Probabilities for the Markov Chain

Event	Relative Probability	State Transition
Re	$\frac{kR}{\binom{i+j+k}{2} + kR}$	$i++$
		$j++$
		$k--$
C_1^l	$\frac{\binom{i}{2}}{\binom{i+j+k}{2} + kR}$	$i--$
C_1^r	$\frac{\binom{j}{2}}{\binom{i+j+k}{2} + kR}$	$j--$
C_2	$\frac{ij}{\binom{i+j+k}{2} + kR}$	$i--$
		$j--$
		$k++$
C_3^l	$\frac{ik}{\binom{i+j+k}{2} + kR}$	$i--$
C_3^r	$\frac{jk}{\binom{i+j+k}{2} + kR}$	$j--$
C_4	$\frac{\binom{k}{2}}{\binom{i+j+k}{2} + kR}$	$k--$

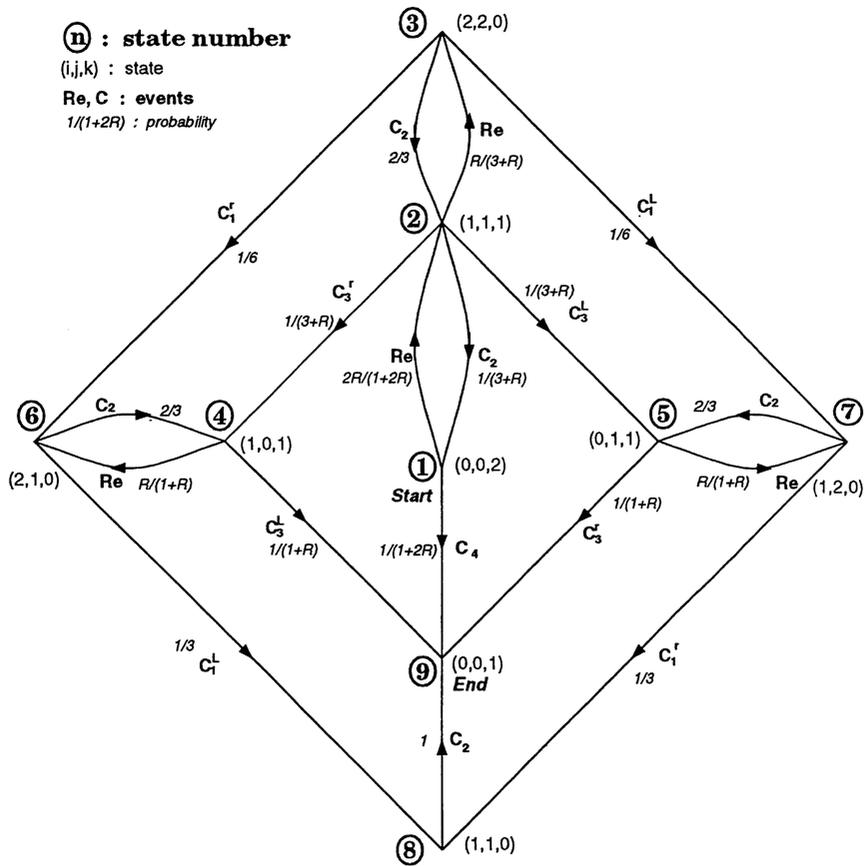


Figure 3: State transition diagram for the Markov chain with $n = 2$ and $m = 2$. The process begins at state 1 and terminates when state 9 is reached. Possible transitions are shown by arrows, with transition probabilities indicated. See text for more details.

The state transition diagram shown in Figure 3 summarizes the possible states and transitions of the Markov chain, that is, the information in Tables I and II. The states are numbered 1 through 9 (shown circled), and the corresponding (i, j, k) is shown for each state. The lines and arrows indicate the directions of possible transition events. For example, it is possible to move from state 2 to state 4 in one step (a coalescent event C_3^r), but not possible to move from state 6 to state 3. The process always begins in state 1 $(0, 0, 2)$ with the two sampled sequences. A coalescent event C_4 would bring the process to state 9, while a recombination event from state 1 would move the process to state 2. The process ends when state 9, the common ancestor state, is reached. The probabilities of each transition event are shown in italics next to the arrows.

An interesting property of Figure 3 is the following. While the process remains in the center section of the diagram (states 1, 2, and 3), the trees at both loci each accumulate the same amount of time. Therefore, if the process reaches state 9 directly from the center section, the two trees (left and right loci) will have the same height. While the process is on the left side of the diagram (states 4 and 6), the right locus has coalesced, but the left locus is accumulating time in its tree. Therefore, if the tree ever visits the left side of the diagram, the tree for the left locus will be taller than that of the right. Similarly, if the process ever visits the right side of the diagram (states 5 and 7), the tree for the right locus will be taller. Due to the one-way nature of the transition arrows, the process can not visit both the left and right sides of the diagram. The probabilities of the three possible outcomes are given by Theorem 2.

3. DISCRETE-TIME ANALYSIS

3.1 The Markov Chain

The tree-building process may be examined in two ways: first, as a discrete-step Markov chain in which only the order of the states is considered; secondly, as a continuous-time Markov chain embedded in the former, with exponentially distributed waiting times. The mean waiting time until the next event depends on the current state (i, j, k) . In this section the discrete-time process is analyzed.

The transition matrix corresponding to the Markov chain state transition diagram shown in Figure 3 is

$$P = \begin{bmatrix} 0 & \frac{2R}{1+2R} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{1+2R} \\ \frac{1}{3+R} & 0 & \frac{R}{3+R} & \frac{1}{3+R} & \frac{1}{3+R} & 0 & 0 & 0 & 0 \\ 0 & 2/3 & 0 & 0 & 0 & 1/6 & 1/6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{R}{1+R} & 0 & 0 & \frac{1}{1+R} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{R}{1+R} & 0 & \frac{1}{1+R} \\ 0 & 0 & 0 & \frac{2}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{2}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (2)$$

Properties: The initial state is always 1, which corresponds to $n = 2$ sequences sampled at the present time. The chain stops at state 9, which corresponds to one sequence ancestral to both sequences at both loci: the MRCA. There are five conjugacy classes of states: $\{1,2,3\}$, $\{4,6\}$, $\{5,7\}$, $\{8\}$, $\{9\}$. (A class consists of states which can be all be reached from one another.) All classes are transient and have period 2 except $\{9\}$ which is absorbing and recurrent. Classes $\{4,6\}$ and $\{5,7\}$ are mutually inaccessible, and class $\{1,2,3\}$ is inaccessible from both. The following theorem shows that the tree must reach a common ancestor.

Theorem 1 The Markov chain defined by the transition matrix P in Eq. (2) has the property that, beginning in state 1, the system will eventually reach state 9, with probability one.

Proof of Theorem 1 The theorem follows from standard results on Markov chains. See (Simonsen, 1996) for a complete proof. ■

Notation for paths: Each path or walk through the graph beginning in state 1 and ending in state 9 represents a particular pair of trees for the two sites. A path is represented by listing the sites visited in order, in bold type. The simplest such path is **19** where no recombination occurs and which has probability $\frac{1}{1+2R}$. Note that when $R = 0$ (no recombination) this path has probability 1. As a shorthand to represent paths where a particular loop is traversed a number of times, we introduce the following exponent notation. The path **12121219** will be represented **1(21)³9**. It will also be necessary to represent sets of paths with certain common properties. Square brackets are used to represent all permutations of their contents, in the sense that the elements inside the brackets may be traversed in any order. Thus **12[(32)(12)]19** represents the two paths **12321219** and **12123219**. Combining this with exponents leads to a fairly compact notation. For example, **12[(12)(32)²]19** represents the three paths **1212323219**, **1232123219**, **1232321219**.

3.2 Equal tree heights:

One question of interest that can be answered using only the discrete-time Markov chain is the following. What is the probability that two adjacent loci have the same tree height? The left tree will be shorter when the left lineage coalesces more recently than the right lineage, that is, when either two single left types coalesce (event C_1^l), or a single left type coalesces with a double type (event C_3^l), before the corresponding event at the right locus. These two events are MRCA events for the left locus. Similarly, if either of the MRCA events at the right locus, C_1^r or C_3^r , happen first, the right locus will have a shorter tree. The two trees will have exactly the same height if and only if these four events never happen. This is equivalent to never visiting states 4, 5, 6, 7, or 8.

Theorem 2 The probability that the tree heights T_l and T_r for two adjacent loci are identical when $n = 2$, which is the same as the probability of never visiting states 4, 5, 6, 7 or 8 before reaching state 9, is

$$\text{Prob}(T_l = T_r) = \frac{9 + R}{9 + 13R + 2R^2} \quad (3)$$

Proof of Theorem 2 The theorem follows from an analysis of the transition matrix Eq. (2). For a complete proof, see (Simonsen, 1996). ■

This agrees with the result derived by Griffiths (1981). By symmetry,

$$\begin{aligned} \text{Prob}(T_l < T_r) = \text{Prob}(T_r < T_l) &= \frac{1 - \text{Prob}(T_l = T_r)}{2} \\ &= \frac{R(6 + R)}{9 + 13R + 2R^2} \end{aligned} \quad (4)$$

Note that as the recombination rate approaches 0, the probability that the left and right trees have the same height goes to 1. Since R has a maximum of N , however, the expression (3) is never zero.

3.3 Joint Distribution of the Number of Visits

Consider the 8-vector $\mathbf{V} = (v_1, \dots, v_8)$, where v_h is the number of times state h is visited. We wish to determine the probability distribution of this vector-valued random variable.

Table III: Constraints on v , the Number of Visits to Each State; Cases with Non-zero Probability.

Case	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	Paths
1	≥ 1	$v_1 + v_3 - 1$	≥ 0	0	0	0	0	0	19
2	≥ 1	$v_1 + v_3$	≥ 0	≥ 1	0	v_4	0	1	24,689
3	≥ 1	$v_1 + v_3$	≥ 0	≥ 1	0	$v_4 - 1$	0	0	24,49
4	≥ 1	$v_1 + v_3 - 1$	≥ 1	≥ 0	0	$v_4 + 1$	0	1	36,689
5	≥ 1	$v_1 + v_3 - 1$	≥ 1	≥ 1	0	v_4	0	0	36,49
6	≥ 1	$v_1 + v_3$	≥ 0	0	≥ 1	0	v_5	1	25,789
7	≥ 1	$v_1 + v_3$	≥ 0	0	≥ 1	0	$v_5 - 1$	0	25,59
8	≥ 1	$v_1 + v_3 - 1$	≥ 1	0	≥ 0	0	$v_5 + 1$	1	37,789
9	≥ 1	$v_1 + v_3 - 1$	≥ 1	0	≥ 1	0	v_5	0	37,59

Theorem 3 The probability distribution of V , the number of visits to each state, is non-zero only in the 9 cases shown in Table III. Let $\text{Prob}[V = v] = Q(v)$ be denoted $Q_c(v)$ in case c , and let

$$H(x, y, z) = \left[\frac{2R}{(1+2R)(3+R)} \right]^x \left[\frac{2R}{3(3+R)} \right]^y \left[\frac{2R}{3(1+R)} \right]^z \binom{x+y-1}{y}. \quad (5)$$

These $Q_c(v)$ are given by

$$\begin{aligned} Q_1(v) &= \frac{1}{1+2R} H(v_1 - 1, v_3, 0) \\ Q_2(v) &= \frac{1}{2} H(v_1, v_3, v_4) & Q_6(v) &= \frac{1}{2} H(v_1, v_3, v_4) \\ Q_3(v) &= \frac{1}{1+R} H(v_1, v_3, v_4 - 1) & Q_7(v) &= \frac{1}{1+R} H(v_1, v_3, v_5 - 1) \\ Q_4(v) &= \frac{R}{18} H(v_1, v_3 - 1, v_4) & Q_8(v) &= \frac{R}{18} H(v_1, v_3 - 1, v_5) \\ Q_5(v) &= \frac{1}{6} H(v_1, v_3 - 1, v_4) & Q_9(v) &= \frac{1}{6} H(v_1, v_3 - 1, v_5). \end{aligned} \quad (6)$$

The proof of Theorem 3 is given in the Appendix. The nine cases of Table III represent different patterns of walks through Figure 3, and are described in detail in the proof.

To verify that Eq. (6) does in fact describe a density, we show that the sum over the space \mathbf{Z}^8 is equal to 1. Each case is summed separately, using well-known results for geometric and negative binomial distributions. Let S_c be the total probability of case c :

$$S_c = \sum_{\mathbf{v}} Q_c(\mathbf{v}) \quad (7)$$

Then

$$S_1 = \sum_{v_1=1}^{\infty} \sum_{v_3=0}^{\infty} Q_1(v) = \frac{9+R}{9+13R+2R^2} \quad (8)$$

$$S_2 = \sum_{v_1=1}^{\infty} \sum_{v_3=0}^{\infty} \sum_{v_4=1}^{\infty} Q_2(v) = \frac{6R^2}{(3+R)(9+13R+2R^2)} = S_6 \quad (9)$$

$$S_3 = \sum_{v_1=1}^{\infty} \sum_{v_3=0}^{\infty} \sum_{v_4=1}^{\infty} Q_3(v) = \frac{18R}{(3+R)(9+13R+2R^2)} = S_7 \quad (10)$$

$$S_4 = \sum_{v_1=1}^{\infty} \sum_{v_3=1}^{\infty} \sum_{v_4=0}^{\infty} Q_4(\mathbf{v}) = \frac{R^2(1+R)}{(3+R)(9+13R+2R^2)} = S_8 \quad (11)$$

$$S_5 = \sum_{v_1=1}^{\infty} \sum_{v_3=1}^{\infty} \sum_{v_4=1}^{\infty} Q_5(\mathbf{v}) = \frac{2R^2}{(3+R)(9+13R+2R^2)} = S_9 \quad (12)$$

From the above expressions, it follows that $\sum_{c=1}^9 S_c = 1$. Note from Table III that only case 1 involves no visits to states 4, 5, 6, or 7; therefore S_1 is equal to the probability of equal tree heights derived in Theorem 2.

3.4 Expected Values and Variances

In this section, the expected values and variances of the components of \mathbf{V} will be derived from the distribution Q . From these values, other quantities of interest, such as total tree time, will be derived in later sections. Let M_{ch} be the expected value of v_h in case c :

$$M_{ch} = \sum_{\mathbf{v}} v_h Q_c(\mathbf{v}). \quad (13)$$

Exact expressions for the M_{ch} are lengthy, and are not reproduced here. The authors would be happy to provide them on request, as well as *Maple* (Char et al., 1991) code for computing them from the distribution $Q(\mathbf{v})$. Complete expressions are given in (Simonsen, 1996).

The expected values of \mathbf{V} are obtained by summing over the cases. Since v_2 depends on v_1 and v_3 , and v_6 on v_4 , their expected values are computed indirectly. For example, in case 4, $v_2 = v_1 + v_3 - 1$, and so $M_{42} = M_{41} + M_{43} - S_4$.

$$\mathbf{E}(v_1) = M_{11} + 2(M_{21} + M_{31} + M_{41} + M_{51}) = \frac{(1+2R)(9+R)}{9+13R+2R^2} \quad (14)$$

$$\begin{aligned} \mathbf{E}(v_2) &= (M_{11} + M_{13} - S_1) + 2[(M_{21} + M_{23}) \\ &\quad + (M_{31} + M_{33}) + (M_{41} + M_{43} - S_4) + (M_{51} + M_{53} - S_5)] \\ &= \frac{6R(3+R)}{9+13R+2R^2} \end{aligned} \quad (15)$$

$$\begin{aligned} \mathbf{E}(v_3) &= M_{13} + 2(M_{23} + M_{33} + M_{43} + M_{53}) \\ &= \frac{6R^2}{9+13R+2R^2} \end{aligned} \quad (16)$$

$$\begin{aligned} \mathbf{E}(v_4) &= \mathbf{E}(v_5) = M_{24} + M_{34} + M_{44} + M_{54} \\ &= \frac{2R(1+R)(9+R)}{(3+R)(9+13R+2R^2)} \end{aligned} \quad (17)$$

$$\begin{aligned} \mathbf{E}(v_6) &= \mathbf{E}(v_7) = M_{24} + (M_{34} - S_3) + (M_{44} + S_4) + M_{54} \\ &= \frac{3R^2(7+R)}{(3+R)(9+13R+2R^2)} \end{aligned} \quad (18)$$

$$\mathbf{E}(v_8) = 2(S_2 + S_4) = \frac{2R^2(7+R)}{(3+R)(9+13R+2R^2)} \quad (19)$$

To calculate variance and covariance terms, it will be necessary to compute second moments and cross-terms for each case. Let

$$C_{cgh} = \sum_{\mathbf{v}} v_g v_h Q_c(\mathbf{v}) \quad (20)$$

for $c = 1, \dots, 9$, $g = 1, \dots, 8$, and $h = 1, \dots, 8$. Then

$$\mathbf{E}(v_g v_h) = \sum_{c=1}^9 C_{cgh}, \quad \mathbf{E}(v_h^2) = \sum_{c=1}^9 C_{chh} \quad (21)$$

$$\text{Var}(v_h) = \sum_{c=1}^9 C_{chh} - (\mathbf{E}(v_h))^2 \quad (22)$$

$$\text{Cov}(v_g, v_h) = \sum_{c=1}^9 C_{cgh} - \mathbf{E}(v_g)\mathbf{E}(v_h). \quad (23)$$

Exact expressions for these are lengthy and thus are not reproduced here. They are available from the authors, with *Maple* (Char et al., 1991) code for computing them from the distribution $Q(\mathbf{v})$.

3.5 The Number of Recombination Events

The number U of recombination events in the history of the sample was studied by Hudson and Kaplan (1985). They considered the general n -sequence m -loci problem and derived an exact expression for $\text{Prob}(U = 0)$ and asymptotic expressions for $\text{Prob}(U = 1)$ and $\mathbf{E}(U)$. The probability distribution of U for $n = 2$ and $m = 2$ can be derived from the distribution of the number of visits $Q(\mathbf{v})$. Without recombination, there must be $n - 1 = 1$ coalescent event before reaching the MRCA. Each recombination event requires an additional coalescent event to compensate for the increased number of sequences in the genealogy. The total number of events is given simply by the number of states visited. Therefore U satisfies

$$2U + 1 = \sum_{h=1}^8 v_h \quad (24)$$

The relationships among the v_h given in Table III, combined with Eq. (24) result in the following possibilities.

$$U = \begin{cases} v_1 + v_3 - 1 & \text{case 1} \\ v_1 + v_3 + v_4 & \text{cases 2 and 4} \\ v_1 + v_3 + v_4 - 1 & \text{cases 3 and 5} \\ v_1 + v_3 + v_5 & \text{cases 6 and 8} \\ v_1 + v_3 + v_5 - 1 & \text{cases 7 and 9} \end{cases} \quad (25)$$

The distributions of these quantities are computed using the appropriate convolutions applied to $Q_c(\mathbf{v})$ for each case. For example, the joint probability of U and case 2 is

$$\text{Prob}(v_1 + v_3 + v_4 = u, \text{case 2}) = \sum_{z=1}^{u-1} \sum_{y=0}^{u-z-1} \frac{1}{2} H(u - y - z, y, z), \quad u \geq 2 \quad (26)$$

The overall probability distribution of U is obtained by summing over all the cases. Some examples for small values of u are given here, and the probabilities for $u = 0, \dots, 5, 10$ as functions of R are shown in Figure 4.

$$\text{Prob}(U = 0) = \frac{1}{1 + 2R} \quad \text{c.f. (Hudson and Kaplan, 1985, Eq.(2))} \quad (27)$$

$$\text{Prob}(U = 1) = \frac{2R(3 + 5R)}{(1 + 2R)^2(3 + R)(1 + R)} \quad (28)$$

$$\text{Prob}(U = 2) = \frac{2R^2(135 + 531R + 685R^2 + 317R^3 + 56R^4 + 4R^5)}{(1 + 2R)^3(3 + R)^2(1 + R)^2} \quad (29)$$

Observe that $\text{Prob}(U = 0) > \text{Prob}(U = 1)$ for all R . An intuitive explanation is as follows. For low recombination rates ($R < 0.5$), the probability of no recombination is high ($\text{Prob}(U = 0) > 0.5$). For higher R , a first recombination event (i.e. visit to state 2) is quite likely. Once that first event has occurred and there are three ancestral sequences, many other possible future events are possible; most of these involve a second recombination event. Thus, the probability of only one recombination event is low.

Histograms of the distribution of U for four different values of R are shown in Figure 5; note that the distribution is bimodal for larger R .

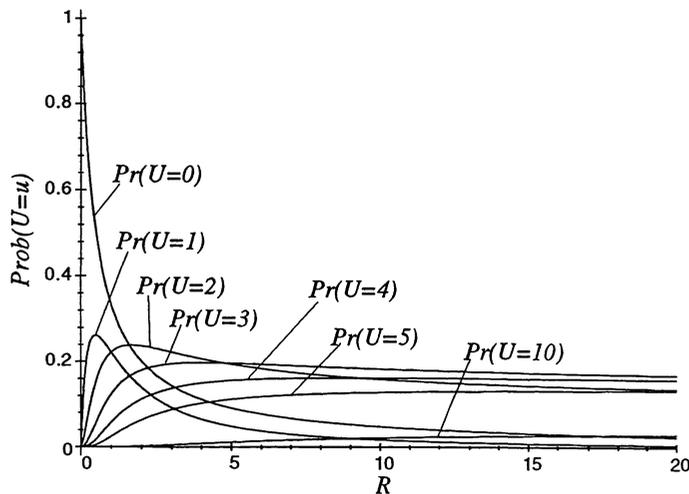


Figure 4: Probabilities of 0, 1, 2, 3, 4, 5, and 10 recombination events in the history of two sequences and two loci, as functions of the recombination rate. Values are based on exact formulas (e.g. Eqs. 27 – 29), not on simulation.

The expected value of U follows simply from Eq. (24) and from the expected value of V in Eqs. (14)-(19).

$$\mathbf{E}(2U + 1) = \sum_{h=1}^8 \mathbf{E}(v_h) = \frac{3 + 13R}{3 + R} \quad (30)$$

$$\mathbf{E}(U) = \frac{6R}{3 + R} \quad (31)$$

Similarly, the variance of U follows from the variances and covariances of the v_h .

$$\begin{aligned} \text{Var}(U) &= \frac{1}{4} \text{Var} \left(\sum_{h=1}^8 v_h \right) \\ &= \frac{1}{4} \left[\sum_{h=1}^8 \text{Var}(v_h) + 2 \sum_{g < h} \text{Cov}(v_g, v_h) \right] \\ &= \frac{6R(27 + 84R + 29R^2 + 4R^3)}{(9 + 13R + 2R^2)(3 + R)^2} \end{aligned} \quad (32)$$

The mean and variance of U have limits of 6 and 12 respectively as R increases.

4. CONTINUOUS-TIME ANALYSIS

In this section the full continuous-time Markov chain is considered. The previous analysis of the discrete-time chain giving the distribution of the number of visits to each state will be used to calculate properties of the total tree times at each locus. The waiting time at each state (measured in units of $2N$ generations) is an exponentially-distributed random variable with parameter $\binom{i+j+k}{2} + kR$. Therefore the total time spent in state h is the sum of v_h exponentially distributed random variables with the same parameter, and hence follows a gamma distribution. Let λ_h be the parameter for state h as shown in Table IV.

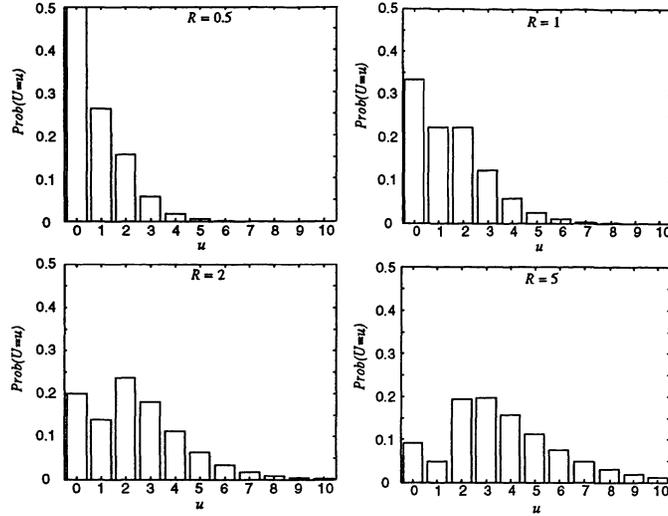


Figure 5: Histogram of the number of recombination events in the history of two sequences and two loci, with recombination rates $R = 0.5, 1, 2,$ and 5 .

Table IV: Parameters in the Continuous-time Markov Chain

State h	(i, j, k)	$\lambda_h = \binom{i+j+k}{2} + kR$
1	(0,0,2)	$1 + 2R$
2	(1,1,1)	$3 + R$
3	(2,2,0)	6
4	(1,0,1)	$1 + R$
5	(0,1,1)	$1 + R$
6	(2,1,0)	3
7	(1,2,0)	3
8	(1,1,0)	1

For a particular outcome of the number of visits v , if state h is ever visited, the total time T_h spent in state h is distributed $\Gamma(v_h, \lambda_h)$, which has density

$$g(t; v_h, \lambda_h) = \begin{cases} \frac{\lambda_h^{v_h} t^{v_h-1} e^{-\lambda_h t}}{(v_h-1)!}, & t \geq 0 \\ 0, & t < 0 \end{cases}. \quad (33)$$

For $h = 1 \dots 8$, let

$$f_h = \begin{cases} \Gamma(v_h, \lambda_h), & v_h \geq 1 \\ \delta, & v_h = 0 \end{cases} \quad (34)$$

where δ is the delta-distribution (all its density at 0). The dependence of the functions f_h on v_h will not be written explicitly, to simplify notation; $f_h(t)$ will represent $f_h(t; v_h, \lambda_h)$. Times of interest are the sum of T_h for different h , and so the distributions of those times are convolutions of gamma distributions. The height T_l of the tree at the left locus is $T_l = T_1 + T_2 + T_3 + T_4 + T_6$, while the height of the tree at the right locus is $T_r = T_1 + T_2 + T_3 + T_5 + T_7$. To simplify notation, let $T_c = T_1 + T_2 + T_3$ be the time common to both

trees. Then $T_c \sim f_c$, where

$$f_c(t) = \int_X f_1(t-x) \int_Y f_2(x-y) f_3(y) dy dx. \quad (35)$$

(The limits of integration are $(-\infty, \infty)$ in all cases; the capital letters shown serve only to clarify the variable of integration.) Therefore, the marginal distributions of these quantities, conditional on the number of visits $\mathbf{V}=\mathbf{v}$, are

$$p_{T_l|\mathbf{V}}(t|\mathbf{V}=\mathbf{v}) = \int_W f_c(t-w) \int_Z f_4(w-z) f_6(z) dz dw \quad (36)$$

$$p_{T_r|\mathbf{V}}(t|\mathbf{V}=\mathbf{v}) = \int_W f_c(t-w) \int_Z f_5(w-z) f_7(z) dz dw \quad (37)$$

Let Ψ denote the joint distribution of the two tree times, conditional on \mathbf{v} :

$$\begin{aligned} \Psi(t_1, t_2; \mathbf{v}) &= p_{T_l, T_r|\mathbf{V}}(t_1, t_2|\mathbf{V}=\mathbf{v}) = \\ &\int_W f_c(t_2-w) \left[\int_{Z_1} f_4(t_1-t_2-w-z_1) f_6(z_1) dz_1 \right] \left[\int_{Z_2} f_5(w-z_2) f_7(z_2) dz_2 \right] dw \end{aligned} \quad (38)$$

Note that since either f_4 and f_6 (cases 2–5), or f_5 and f_7 (cases 6–9), or all four (case 1) are equal to δ , Eq. (38) will be equivalent to one of the following.

$$\Psi(t_1, t_2; \mathbf{v}) = \begin{cases} \delta(t_1 - t_2) f_c(t_1) & \text{(case 1)} \\ f_c(t_2) \int_Z f_4(t_2 - t_1 - z) f_6(z) dz & \text{(cases 2 - 5)} \\ f_c(t_1) \int_Z f_5(t_2 - t_1 - z) f_7(z) dz & \text{(cases 6 - 9)} \end{cases} \quad (39)$$

The joint distribution of T_l and T_r now follows from Bayes' rule:

$$p_{T_l, T_r}(t_1, t_2) = \sum_{\mathbf{v}} \Psi(t_1, t_2; \mathbf{v}) Q(\mathbf{v}) \quad (40)$$

A scatter plot of this distribution for 1000 realizations of the pair (T_r, T_l) with $R = 0.723$ is shown in Figure 6. This bivariate density has the interesting property that a fixed fraction of its mass, namely $(9 + R)/(9 + 13R + 2R^2)$ (Eq. 3), is exactly along the diagonal. The value 0.723 of R used for the scatter plot places approximately half the points along the diagonal $T_l = T_r$.

4.1 Expected Values and Variances

The total tree times at each locus depend on the time spent in each state. The expected value of T_h , the total time spent in state h , is easily computed from the expected value of v_h , since

$$\mathbf{E}(T_h) = \mathbf{E}_{\mathbf{V}} \mathbf{E}_{T_h|\mathbf{V}}(T_h|\mathbf{V}=\mathbf{v}) = \frac{\mathbf{E}_{\mathbf{V}}(v_h)}{\lambda_h}. \quad (41)$$

Therefore,

$$\mathbf{E}(T_1) = \frac{9 + R}{9 + 13R + 2R^2} \quad (42)$$

$$\mathbf{E}(T_2) = \frac{6R}{9 + 13R + 2R^2} \quad (43)$$

$$\mathbf{E}(T_3) = \frac{R^2}{9 + 13R + 2R^2} \quad (44)$$

$$\mathbf{E}(T_4) = \mathbf{E}(T_5) = \frac{2R(9 + R)}{(3 + R)(9 + 13R + 2R^2)} \quad (45)$$

$$\mathbf{E}(T_6) = \mathbf{E}(T_7) = \frac{R^2(7 + R)}{(3 + R)(9 + 13R + 2R^2)} \quad (46)$$

$$\mathbf{E}(T_8) = \frac{2R^2(7 + R)}{(3 + R)(9 + 13R + 2R^2)}. \quad (47)$$

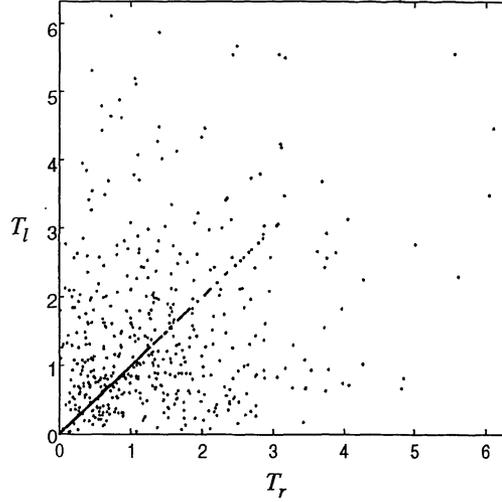


Figure 6: Scatter plot (1000 points) of the joint distribution of T_l, T_r (Eq. 40) with $R = 0.723$. For this value of R , $\text{Prob}(T_l = T_r) = 0.5$, so that half the points are on the diagonal (equal tree heights), one quarter are above the diagonal (left tree taller), and one quarter are below the diagonal (right tree taller).

From these, it is easy to verify that the marginal expectations and variances of T_l and T_r are equal to 1.

The covariance of the left and right tree heights is computed using the fact that $T_4T_5 = T_4T_7 = T_5T_6 = T_6T_7 = 0$.

$$\text{Cov}(T_l, T_r) = \mathbf{E}(T_l T_r) - \mathbf{E}(T_l)\mathbf{E}(T_r) \quad (48)$$

$$\mathbf{E}(T_l T_r) = \mathbf{E}\left(T_1^2 + T_2^2 + T_3^2 + 2(T_1 T_2 + T_1 T_3 + T_2 T_3) + \sum_{g=1}^3 \sum_{h=4}^7 T_g T_h\right) \quad (49)$$

$$\mathbf{E}(T_h^2) = \mathbf{E}_{\mathbf{V}} \mathbf{E}_{T|\mathbf{V}} T_h^2 = \mathbf{E}_{\mathbf{V}} \frac{v_h^2 + v_h}{\lambda_h^2} = \frac{\mathbf{E}(v_h^2)}{\lambda_h^2} + \frac{\mathbf{E}(v_h)}{\lambda_h^2} \quad (50)$$

$$\mathbf{E}(T_g T_h) = \mathbf{E}_{\mathbf{V}} \mathbf{E}_{T|\mathbf{V}}(T_g T_h) = \frac{\mathbf{E}_{\mathbf{V}}(v_g v_h)}{\lambda_g \lambda_h}, \quad (g \neq h) \quad (51)$$

since for fixed $\mathbf{V}=v$, T_g and T_h are independent if $g \neq h$. Substituting Eqs. (50), (51), and (21) into Eq. (48) gives

$$\text{Cov}(T_l, T_r) = \frac{9 + R}{9 + 13R + 2R^2}, \quad (52)$$

which agrees with the result of Griffiths (1981) and satisfies the recurrence equations of Kaplan and Hudson (1985).

This expression for the covariance has implications for the variance of the number of segregating sites observed in DNA sequences. Suppose an observed set of two sequences is comprised of two loci between which recombination can occur, and let S_l and S_r be the number of segregating sites observed at the left and right loci. The standard assumption is that S_i is Poisson-distributed with parameter $\theta_i T_i$ where $\theta = 4N\mu$ is the mutation rate, (note that here T is the tree *height*, not total tree time) and this should be true for each locus. Without recombination, T_l must equal T_r , and they have common expectation and variance 1. This leads to the standard result (for $n = 2$) that $\mathbf{E}(S_i) = \theta_i$ and $\text{Var}(S_i) = \theta_i + \theta_i^2$. The total number of segregating sites observed is $S_l + S_r$; with no recombination its expectation and variance are just as given

above except with θ_i replaced by $\theta_l + \theta_r$. However, if recombination is allowed, the variance is reduced below this value, as shown.

$$\begin{aligned}
\text{Var}(S_l + S_r) &= \text{Var}(S_l) + \text{Var}(S_r) + 2\text{Cov}(S_l, S_r) \\
&= (\theta_l + \theta_l^2) + (\theta_r + \theta_r^2) + 2\theta_l\theta_r\text{Cov}(T_l, T_r) \\
&= (\theta_l + \theta_r) + (\theta_l^2 + \theta_r^2) + 2\theta_l\theta_r\frac{9 + R}{9 + 13R + 2R^2}
\end{aligned} \tag{53}$$

If $R = 0$, of course, this last expression becomes simply $(\theta_l + \theta_r) + (\theta_l + \theta_r)^2$.

5. MORE LOCI AND SEQUENCES

The Markov chain can be extended to larger sample sizes n and number of loci m . Defining the states and transition probabilities is fairly straightforward, but an exact analysis is quite complicated in general. The case $n = 2$, $m = 2$ shown above was relatively simple to analyze because there is only one possible labeled topology for a tree with two sequences, and therefore the only possible change between loci is the height of the tree. For larger n , however, more information is required in order to keep track of topology information. Each sequence may be ancestral to the sample at any combination of loci, and a lineage at any locus may be ancestral to any of the sequences in the sample. As a first approach, let us for the moment ignore topology information and be concerned only with the total time in the tree at each locus. Then information about the sequences to which each lineage is ancestral is not important.

In the case $m = 2$ there were three types of sequence, which were referred to as “single left”, “single right”, and “double”. More generally, a sequence will be called *active* at a locus if it is ancestral to any sampled sequence at that locus. Thus the three types for $m = 2$ are called active at locus 1, at locus 2, and at both loci, respectively. In general, sequences may be active at any of m loci, so that there are $M = 2^m - 1$ possible types of sequence (they must be active at at least one locus). For convenience, we order these types according to counting in binary, so that type 1 is active only at locus 1, type 3 is active at loci 1 and 2, type 2^{m-1} is active only at locus m , type M is active at all loci, and so on. Then the state is given by the vector $\mathbf{Y} = (y_1, y_2, \dots, y_M)$, where y_i is the current number of sequences of type i . Let $X = \sum_{i=1}^M y_i$ be the current total number of sequences.

Of interest is the total number of states for any given n and m . When $m = 2$ there are $n(2n^2 + 9n + 1)/6$ states, and when $m = 3$ the number of states is $\frac{203}{60}n^5 - \frac{287}{8}n^4 + \frac{685}{3}n^3 - \frac{5729}{8}n^2 + \frac{69077}{60}n - 733$. A general formula for the number of states has not been derived, but it is equal to the volume of a region in M -dimensional space defined by the following $m + M$ constraint equations:

$$1 \leq \sum_{i \in A_j} y_i \leq n, \quad A_j = \{1 \leq i \leq M : i \bmod 2^j \geq 2^{j-1}\}, \quad j = 1, \dots, m \tag{54}$$

$$0 \leq y_i, \quad i = 1, \dots, M. \tag{55}$$

These constraints correspond to Eqs. (1) for the $n = 2, m = 2$ case, since A_j is the set of all types that are active at locus j . The first set of equations means that there must always be between one and n lineages for each locus.

5.1 Transitions

When X sequences are present in the population, there are $\binom{X}{2}$ equally likely coalescent events possible. The time back to the most recent coalescent event is exponentially distributed with parameter $\binom{X}{2}/2N$. Coalescence may occur between two sequences of the same type i , which has relative probability $\binom{y_i}{2}$ and results in an ancestor also of type i and a state change y_i-- (meaning $y_i = y_i - 1$). Coalescence may also occur between two sequences of different types i_1 and i_2 , with relative probability $y_{i_1}y_{i_2}$, resulting in an ancestor of type $i_3 = i_1 | i_2$ ($|$ means binary OR) and a state change $y_{i_1}--, y_{i_2}--, y_{i_3}++$ (meaning $y_{i_3} = y_{i_3} + 1$).

Recombination may occur between any two adjacent loci, and hence at any of $m - 1$ places, at $m - 1$ possible rates. Let $\mathbf{R} = 2N(r_1, \dots, r_{m-1})^T$, where r_j is the per-generation probability of recombination between loci j and $j + 1$. Each type $i = 1 \dots M$ has an associated recombination probability (that is, of being the recombinant descendant of two sequences in the previous generation) that depends on which loci are active for that type. Suppose a certain sequence is active for exactly two loci: j_1 and j_2 . This means that only the DNA at those two loci is ancestral to our sample, and the history at the other loci is irrelevant. For a recombination event affecting this sequence to be relevant to the history of the sample, and hence count as an “event” in the tree-building process, it must affect the two active loci for this sequence. Therefore recombination must take place somewhere between the two loci. There are $|j_2 - j_1|$ places this can happen. Recombination events at all other places are irrelevant for this type.

For a general type i , recombination is relevant anywhere between the two most distant active loci for that type. For each type i and locus j define Λ_{ij} to be 1 if recombination between locus j and $j + 1$ is relevant for a sequence of type i , and 0 otherwise. As a simple procedure to determine Λ , let $C_j = 2^j - 1$ and $\overline{C}_j = M - C_j$. Then $\Lambda_{ij} = 1 \iff$ both $i \& C_j$ and $i \& \overline{C}_j$ are non-zero ($\&$ means binary AND). Λ is an M by $m - 1$ indicator matrix. The total rate of recombination when in state Y is equal to $Y\Lambda R$; the probability of recombination in a sequence of type i between loci j and $j + 1$ is $y_i \Lambda_{ij} R_j$. The latter is the transition probability associated with the state change $y_i --, y_{i_1} ++, y_{i_2} ++$, where i_1 and i_2 are the types obtained by splitting a sequence of type i between locus j and $j + 1$, that is, $i_1 = i \& C_j$, and $i_2 = i \& \overline{C}_j$.

5.2 Example: $m = 3, n = 4$

To make the above ideas concrete, suppose $m = 3$ and $n = 4$. Then $M = 7$ and there are 1308 possible states Y . There are two possible sites for recombination, with recombination rates $\mathbf{R} = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}$. The 7 possible types and their binary representations are the following: 1=[001], 2=[010], 3=[011], 4=[100], 5=[101], 6=[110], 7=[111]. Types 1, 2, and 4 are ineligible for recombination, since they are active for only one locus. For type 3, recombination is relevant only between loci 1 and 2, while for type 6 it is relevant only between locus 2 and 3. Recombination at either place is relevant for types 5 and 7. Since $C_1 = [001]$, $\overline{C}_1 = [110]$, $C_2 = [011]$, and $\overline{C}_2 = [100]$, the indicator matrix is

$$\Lambda = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}^T. \quad (56)$$

The 1308 possible states are all those $\mathbf{Y} = (y_1, \dots, y_7)$ satisfying the following constraint equations:

$$1 \leq y_1 + y_3 + y_5 + y_7 \leq 3 \quad (57)$$

$$1 \leq y_2 + y_3 + y_6 + y_7 \leq 3 \quad (58)$$

$$1 \leq y_4 + y_5 + y_6 + y_7 \leq 3 \quad (59)$$

$$0 \leq y_i, \quad i = 1, \dots, 7. \quad (60)$$

The initial state is $Y(0) = (0, 0, 0, 0, 0, 4)$, and in this state the total rate of recombination is $Y(0)\Lambda R = 4R_1 + 4R_2$, and the total rate of coalescence is $\binom{4}{2} = 6$. Thus the time spent in the initial state is exponentially distributed with parameter $4R_1 + 4R_2 + 6$, and there are 14 different ways to move to a new state, via 8 recombination events (four with probability $R_1/(4R_1 + 4R_2 + 6)$, four with probability $R_2/(4R_1 + 4R_2 + 6)$) and 6 coalescent events (each with probability $1/(4R_1 + 4R_2 + 6)$).

6. CHANGES IN TREE TOPOLOGY

Incorporating topology information into the model makes the Markov chain more complex. Each lineage may be ancestral to any of the n sequences in the sample, with $2^n - 1$ possibilities. In the previous section, when topology information was ignored, a “type” of sequence was defined by its active loci. Now, we must also specify, for each active locus, those members of the sample to which the lineage is ancestral. A “type” can now be described by an $n \times m$ matrix of 1’s and 0’s, where the (i, j) th element is 1 if and only if that type is active for locus j , and is ancestral to sample member i at locus j . A “vertical OR” of this matrix (i.e. column is 1 if there are any 1’s in that column) reduces this matrix to the m -digit binary types of the

previous section, destroying topology information. There are $2^{nm} - 1$ possible types. Transitions have similar effects, in that coalescence produces a new ancestor whose type is the binary OR of the two descendants, and recombination produces two new ancestors whose types are those formed by splitting the matrix in two and padding either half with zeros. Another, and perhaps simpler way to think of these new “types” is to consider each column of the matrix as a number in base 2^n , and a type as a number in that base (instead of in binary, as in the previous section).

Suppose $n = 4$ and $m = 3$. At each locus, a type may be ancestral to any of the 4 sample members, with $2^4 = 16$ possibilities; if it is not ancestral to any then the type is not active for that locus. Thus a type is defined by $m = 3$ hexadecimal digits, with $(2^4)^3 - 1 = 4095$ possibilities. Thus there is a natural ordering of types that corresponds to counting in hexadecimal; each type is represented by a hexadecimal integer with m digits, ranging from $1 = [001]$ (active only at locus 1, ancestral only to sample member 1) to $4095 = [fff]$ (active at all loci for all sequences; the type of the MRCA). The initial sample members are of types $[111] = 273$, $[222] = 546$, $[444] = 1092$, and $[888] = 2194$, so the initial state is a 4095-vector with a 1 in those four positions.

7. COMPUTER SIMULATION

Constructing many-locus trees with recombination, as outlined by Hudson (1983; 1990), is a complicated task. Computer programs to build such trees require complex data structures, and a large amount of memory and computer time to run. Furthermore, the necessary code is difficult to write, understand, and debug. In contrast, a computer program to iterate a Markov chain is relatively simple. Such a program consists of repeating three basic steps:

1. in the current state, determine the probabilities of moving to any other state
2. randomly choose the next state
3. update the current state

The simplicity and speed of this process makes it feasible to perform large scale computer simulation studies, such as those needed to analyze statistical tests. Results should be identical to those obtained by the simulation method of Hudson (1983). One obvious use of such simulations is to verify the formulas derived in earlier sections. We can also use simulations to estimate identical quantities for larger m and n . For

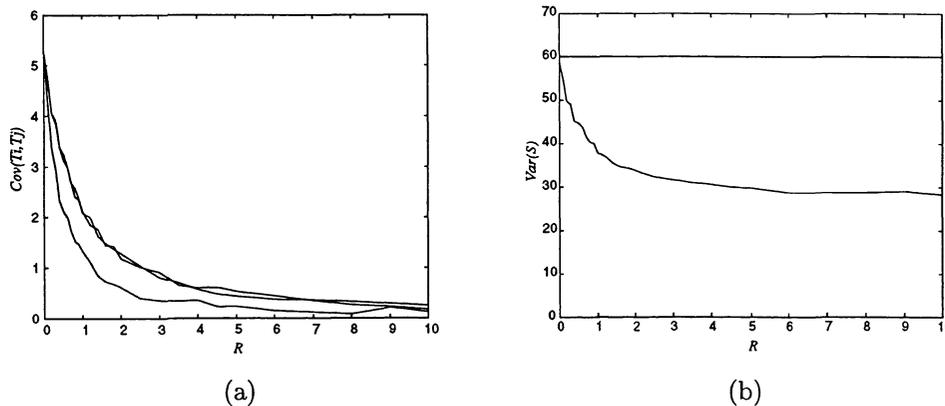


Figure 7: (a) Covariance of total tree times and (b) Variance of the number of segregating sites, for $n = 4$, $m = 3$, based on 10,000 iterations of the Markov chain. The upper two curves in (a) are $\text{Cov}(T_1, T_2)$ and $\text{Cov}(T_2, T_3)$; the lower curve is $\text{Cov}(T_1, T_3)$, where T_j is the total tree time for locus j . The recombination rate R is the same between locus 1 and locus 2 as between locus 2 and locus 3 ($R_1 = R_2 = R$), and the mutation rate for (b) is $\theta_1 = \theta_2 = \theta_3 = 2$.

example, the total number of segregating sites at m loci has variance

$$\text{Var}(S) = \sum_{j=1}^m \text{Var}(S_j) + 2 \sum_{i<j}^m \text{Cov}(S_i, S_j). \quad (61)$$

By simulating total tree times at m loci this expression can be estimated. Figure 7(a) shows the pairwise covariances of tree times for $n = 4, m = 3$, and Figure 7(b) shows the resulting variance of the number of segregating sites at all three loci. In a similar manner, the distribution of the number of recombination events in the history of a sample of any particular size could be estimated.

If topology information is included in the simulations, information can be obtained about the probability of changing topologies between loci, and its dependence on recombination rate. Figure 8 shows an estimate of the probability that trees at two adjacent loci have the same labeled topology, as computed by simulating the Markov chain 10,000 times for $n = 3$ and $m = 2$. When $n = 3$ there are three possible labelled topologies for the trees, and so the probability of identical labelled topology levels off at $1/3$ as the recombination rate grows large.

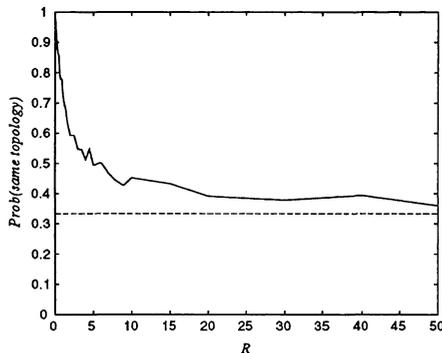


Figure 8: Probability that trees at adjacent loci have the same labeled topology, based on 10,000 iterations of the Markov chain with $n = 3$ and $m = 2$. There are three possible topologies.

8. DISCUSSION

We have analyzed in detail the special case $n = 2, m = 2$ of the Markov chain of Kaplan and Hudson (1985). This case was amenable to exact analytic results because it involves only nine states. A similar approach might be used to derive results for other small values of m and n , but the number of states, and hence the complexity of the analysis, grows very quickly. The results given here serve two purposes: to illustrate the technique by which such results are derived, and to provide exact solutions whenever sequences and loci are considered pairwise. It is also possible that under some circumstances the results for the special case may serve as good approximations to the results for larger n and m .

For the $n = 2, m = 2$ case the key result is the probability distribution of the number of visits to each state. From this distribution is derived the probability distribution of the number of recombination events in the history of the sample, and the joint distribution of tree times at two loci. Both of these have implications for statistical properties of DNA sequences.

The Markov chain model generalizes to any n and m . The states are m -digit binary numbers if topology information is ignored; m -digit numbers in base 2^n if topology information is included. The generalized model is a straightforward and useful computational tool. Simulations using this technique should give identical results to those using the method of Hudson (1983), but are easier to perform. Many interesting properties of DNA sequences, such as those depending on tree times or the number of recombination events, may be studied using simulations of the simpler Markov chain, with no topology information. If topology

information is of interest, the full Markov chain can be used to perform simulations more efficiently than previous methods allowed.

ACKNOWLEDGMENTS

We thank N. L. Kaplan, C. F. Aquadro, M. J. Ford, R. T. Durrett, and J. Wakeley for helpful discussions and comments. This work was completed as part of the first author's Ph.D. thesis at Cornell University.

REFERENCES

- Begun, D. J. and Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, 356:519–520.
- Bickeböllner, H. and Thompson, E. A. (1996). Distribution of genome shared ibd by half-sibs: Approximation by the poisson clumping heuristic. *Theoretical Population Biology*, 50(1):66–90.
- Char, B. W., Geddes, K. O., Gonnet, G. H., Leong, B. L., Monagan, M. B., and Watt, S. M. (1991). *Maple V Library Reference Manual*. Springer-Verlag, New York.
- Gaut, B. S. and Weir, B. S. (1994). Detecting substitution-rate heterogeneity among regions of a nucleotide sequence. *Molecular Biology and Evolution*, 11(4):620–629.
- Griffiths, R. C. (1981). Neutral two-locus multiple allele models with recombination. *Theoretical Population Biology*, 19:169–186.
- Gu, X., Fu, Y.-X., and Li, W.-H. (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Molecular Biology and Evolution*, 12(4):546–557.
- Hey, J. and Wakeley, J. (1996). A coalescent estimator of the population recombination rate. (Manuscript).
- Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23:183–201.
- Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7:1–44.
- Hudson, R. R. and Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample 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. *Theoretical Population Biology*, 28:382–396.
- Pluzhnikov, A. and Donnelly, P. (1996). Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics*. In press.
- Simonsen, K. L. (1996). *Models of DNA Sequence Evolution and Applications to Hypothesis Testing*. PhD thesis, Cornell University, Center for Applied Mathematics.
- Simonsen, K. L., Churchill, G. A., and Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, 141:413–429.
- Wakeley, J. (1993). Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *Journal of Molecular Evolution*, 37(6):613–623.
- Wakeley, J. (1994). Substitution-rate variation among sites and the estimation of transition bias. *Molecular Biology and Evolution*, 11(3):436–442.

APPENDIX

Proof of Theorem 3 The division of the space of possible v into 9 cases is a consequence of the constraints imposed on v by the fact that a state may only be entered from an adjacent state. Recall that the conjugacy classes of this Markov chain are $\{1,2,3\}$, $\{4,6\}$, $\{5,7\}$, $\{8\}$, and $\{9\}$. The structure dictates that classes $\{1,2,3\}$ and $\{9\}$ are always visited, and that either $\{4,6\}$ or $\{5,7\}$ or neither, but not both, may be visited. If $\{4,6\}$ is visited, it may be entered either by state 4 through path **24** or by state 6 through path **36**. The class may be left either by state 4 through path **49** or by state 6 through path **689**. This results in four possible cases. A similar choice applies to class $\{5,7\}$. Taking path **24** or **25** imposes the constraint $v_1 + v_3 = v_2$, since a visit to 1 or 3 is always followed by a visit to 2. On the other hand, taking path **36** or **37** imposes the constraint $v_2 = v_1 + v_3 - 1$, since after the final visit to state 3, state 2 is not visited again.

Case 1: Consider first the case that neither class $\{4,6\}$ nor $\{5,7\}$ is visited, which further implies that $\{8\}$ is not visited. In other words, suppose $v_4 = v_5 = v_6 = v_7 = v_8 = 0$. Then since state 2 lies between states 1 and 3, and state 9 must be reached through state 1, there is the constraint $v_1 + v_3 = v_2 + 1$. If $v_3 = 0$ then $v_2 = v_1 - 1$ and the set of possible paths can be written $1 (21)^{v_1-1} 9$, so that the probability is geometrically distributed:

$$\text{Prob}[\mathbf{V} = (v_1, v_1 - 1, 0, 0, 0, 0, 0, 0, 1)] = (P_{12}P_{21})^{v_1-1} P_{19}. \quad (62)$$

If $v_3 > 0$ then $v_2 = v_1 + v_3 - 1 > 0$ and the set of possible paths is written $12 \left[(12)^{v_1-2} (32)^{v_3} \right] 19$. Hence

$$\begin{aligned} Q_1(v) &= \text{Prob}[\mathbf{V} = (v_1, v_1 + v_3 - 1, v_3, 0, 0, 0, 0, 0, 1)] \\ &= (P_{12}P_{21})^{v_1-1} (P_{23}P_{32})^{v_3} \binom{v_1+v_3-2}{v_3} P_{19}. \end{aligned} \quad (63)$$

Note the similarity to a negative binomial distribution. This reduces to Eq. 62 when $v_3 = 0$. Since

$$H(x, y, z) = (P_{21}P_{12})^x (P_{23}P_{32})^y \binom{x+y-1}{y} (P_{64}P_{46})^z, \quad (64)$$

the result for $Q_1(v)$ follows.

Case 2: paths **24** and **689**. In this case class $\{4,6\}$ is entered at state 4 and left by state 6, so $v_4 = v_6$, and $v_8 = 1$. The set of all possible paths is $12 \left[(12)^{v_1-1} (32)^{v_3} \right] (46)^{v_4} 89$. This has probability

$$\begin{aligned} Q_2(v) &= \text{Prob}[\mathbf{V} = (v_1, v_1 + v_3, v_3, v_4, 0, v_4, 0, 1, 1)] \\ &= P_{12} (P_{21}P_{12})^{v_1-1} (P_{23}P_{32})^{v_3} \binom{v_1+v_3-1}{v_3} P_{24}P_{46} (P_{64}P_{46})^{v_4-1} P_{68}P_{89} \\ &= \frac{1}{2} H(v_1, v_3, v_4) \end{aligned} \quad (65)$$

Case 3: paths **24** and **49**. In this case class $\{4,6\}$ is both entered and left by state 4, so $v_6 = v_4 - 1$, and $v_8 = 0$. The set of possible paths is written

$12 \left[(12)^{v_1-1} (32)^{v_3} \right] 4 (64)^{v_4-1} 9$. This has probability

$$\begin{aligned} Q_3(v) &= \text{Prob}[\mathbf{V} = (v_1, v_1 + v_3, v_3, v_4, 0, v_4 - 1, 0, 0, 1)] \\ &= P_{12} (P_{21}P_{12})^{v_1-1} (P_{23}P_{32})^{v_3} \binom{v_1+v_3-1}{v_3} P_{24} (P_{46}P_{64})^{v_4-1} P_{49} \\ &= \frac{1}{1+R} H(v_1, v_3, v_4 - 1) \end{aligned} \quad (66)$$

Case 4: paths 36 and 689. In this case class $\{4,6\}$ is both entered and left by state 6, so $v_6 = v_4 + 1$, and $v_8 = 1$. The set of all possible paths is written

$12 \left[(12)^{v_1-1} (32)^{v_3-1} \right] 36 (46)^{v_4} 89$. This has probability

$$\begin{aligned}
Q_4(v) &= \text{Prob}[V = (v_1, v_1 + v_3 - 1, v_3, v_4, 0, v_4 + 1, 0, 1, 1)] \\
&= P_{12} (P_{21} P_{12})^{v_1-1} (P_{23} P_{32})^{v_3-1} \binom{v_1+v_3-2}{v_3-1} P_{23} P_{36} (P_{64} P_{46})^{v_4} P_{68} P_{89} \\
&= \frac{R}{18} H(v_1, v_3 - 1, v_4)
\end{aligned} \tag{67}$$

Case 5: paths 36 and 49. In this case class $\{4,6\}$ is entered by state 6 and left by state 4, so $v_4 = v_6$, and $v_8 = 0$. The set of all possible paths is written

$12 \left[(12)^{v_1-1} (32)^{v_3-1} \right] 364 (64)^{v_4-1} 9$. This has probability

$$\begin{aligned}
Q_5(v) &= \text{Prob}[V = (v_1, v_1 + v_3 - 1, v_3, v_4, 0, v_4, 0, 0, 1)] \\
&= P_{12} (P_{21} P_{12})^{v_1-1} (P_{23} P_{32})^{v_3-1} \binom{v_1+v_3-2}{v_3-1} P_{23} P_{36} P_{64} (P_{46} P_{64})^{v_4-1} P_{49} \\
&= \frac{1}{6} H(v_1, v_3 - 1, v_4)
\end{aligned} \tag{68}$$

This completes the proof for the cases where class $\{4,6\}$ is visited. The proof for the parallel cases $(6 - 9)$ where class $\{5,7\}$ is visited is exactly the same, with v_4 and v_6 replaced by v_5 and v_7 . ■