Estimation in ancestral recombination graphs using Markov chain Monte Carlo

Paul Marjoram[1], **Lada Markovtsova**[2] **and Simon Tavaré**[1,2,3]

University of Southern California

Research Report. October 20, 2000 Updated May 27, 2001

ABSTRACT

We present a Reversible Jump Markov chain Monte Carlo [MCMC] approach to ancestral inference for sequence data that have undergone recombination. Our method uses the ancestral recombination graph to model the genealogy of chromosomes from a random sample of individuals. Our framework is Bayesian, although it can also produce maximum likelihood estimates.

---

[1] Biostatistics Division, Department of Preventive Medicine, University of Southern California, 1540 Alcazar Street, CHP 220, Los Angeles, CA 90033. [2]Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113. [3] Program in Molecular Biology, SHS 172, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-1340. The authors were supported in part by NIH grant GM 58897.

1 Introduction

In this research report we present details of an MCMC approach to inference in ancestral recombination graphs. Our application is an application of *Reversible Jump* MCMC methodology, first introduced by Green (1995). We describe the method in the context of a random mating population of constant size, but the approach can be generalized in a number of ways. Our long-term interest is in the analysis of phenotypic data from a disease, together with a collection of markers, such as Single Nucleotide Polymorphisms [SNPs], with a view to locating genes or functional mutations for the disease. We model the disease phenotype and allow for incomplete penetrance at the disease locus. For simplicity, we explain the methodology in the context of a two-locus, finite sites model. Our method allows for the estimation of penetrance probabilities, as well as the distance between the marker of interest and the functional mutation or disease gene. Related approaches in the context of marker maps include those of Griffiths and Tavaré (1989), Griffiths and Marjoram (1996), Kuhner et al. (2000) and Nielsen (2000). Wall (2000) provides a useful overview of the properties of many of these methods.

1.1 Ancestral Recombination Graphs

The evolution of a population that is not subject to recombination is usually described by the stochastic process known as the coalescent, introduced by Kingman (1982). This process models the ancestry of a random sample of chromosomal regions drawn from the present day population. As one looks back in time, lines of ancestry join, or coalesce, at times when the chromosomes in the sample share common ancestors. Eventually the sample can be traced back to its most recent common ancestor (MRCA).

The coalescent can be decomposed into two independent components: a tree topology that describes the coalescence of lines of ancestry, and a set of times recording the instances at which these coalescence events occur. Time is measured in units of N generations, where N is the current population size. Furthermore, time increases as we move back into time (up the tree), the time at which the sample is drawn being defined as $t = 0$. We refer to the point at which the sample was drawn as the 'bottom'

of the tree, and the MRCA as the top of the tree.

Let u be the mutation probability per sequence per generation, and define $\mu = 2Nu$. (We use the notation μ, rather than the more common θ, to avoid confusion with established notation for the recombination rate.) $\mu/2$ is the expected number of mutations per individual per unit time on the coalescent time scale. For a comprehensive discussion of the coalescent process the reader is referred to Hudson (1991), Donnelly and Tavaré (1995), and Nordborg (2000).

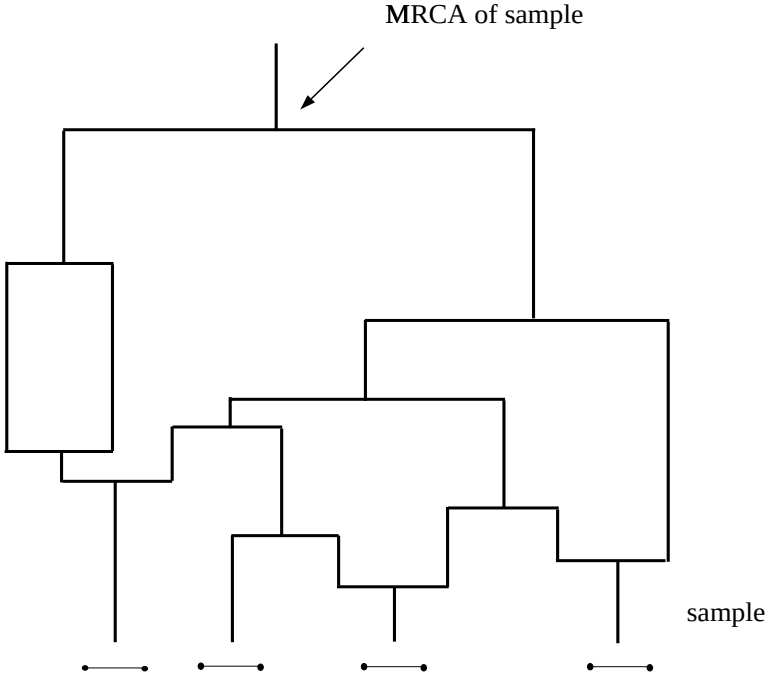
More recently, the coalescent has been used as the basis for estimation and inference via the use of the Importance Sampling approach of Griffiths and Tavaré (1994a, 1994b), and the Markov chain Monte Carlo [MCMC] approach initiated by Kuhner *et al.* (1995, 1998). Bayesian methods for inference in the coalescent are described by Tavaré *et al.* (1997), Wilson and Balding (1998), and Markovtsova *et al.* (2000a,b). The Bayesian approach has the advantage of being able to generate posterior distributions for parameters of interest that allow for prior information about these parameters. By dividing out the prior, one can still produce maximum likelihood estimates of relevant parameters if necessary.

It is relatively straightforward to allow for recombination in the context of the coalescent. This was first shown by Hudson (1983), and further developed by Griffiths (1992) and Griffiths and Marjoram (1997). Let r be the recombination rate per individual per generation, and define $\rho = 2Nr$; $\rho/2$ is the expected number of recombinations per individual per unit time on the coalescent time scale. Instead of coalescent trees we now have recombination graphs. Moving up the graph, as well as the usual coalescence events, recombination events are now possible. If $W(t)$ is the jump chain of events within the recombination graph, (so $W(\cdot)$ records the changes in the number of lines but not the times between the jumps), and if there are k lines currently in the graph, the jump probabilities are

$$\begin{aligned} \mathbb{P}(k \to k-1) &= \frac{k-1}{k-1+\rho}, \\ \mathbb{P}(k \to k+1) &= \frac{\rho}{k-1+\rho}. \end{aligned}$$

The first probability corresponds to a coalescence, the second to a recombination. If there are currently k lines, then the time to the next event is distributed as an

Figure 1: Example of a recombination graph

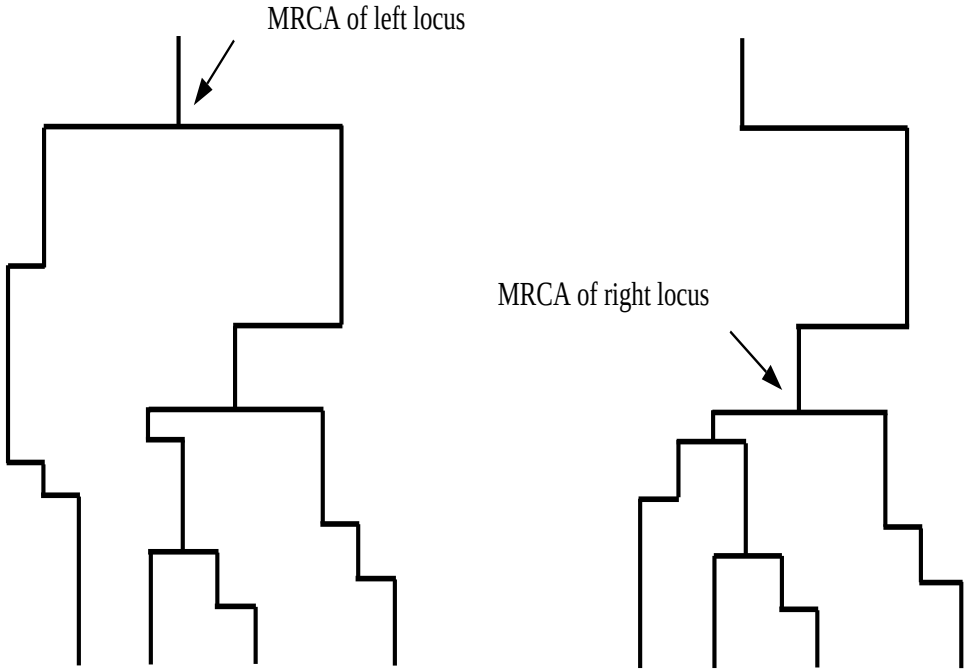


exponential random variable with parameter $k\rho/2 + k(k-1)/2$. Recombination is modeled as follows: In the event of a recombination a chromosome is picked, uniformly at random from among those currently present, and the sequence corresponding to the chromosome is split into two. The left region is inherited from one parent, the right region from another, both parents being randomly chosen from the entire population. After such an event there is now an extra line of ancestry in the graph.

Since the number of lines of ancestry can now increase, it is not immediately obvious that such a process necessarily reaches a single MRCA. However, since coalescence occurs at a rate which is quadratic in the number of lines (*i.e.* $k(k-1)/2$) and the recombination rate grows linearly (*i.e.* as $k\rho/2$), an MRCA is attained with probability 1. For a comprehensive discussion of this and related matters on ancestral recombination graphs see Griffiths and Marjoram (1997). A related approach is given in Wiuf and Hein (1999).

An example of a recombination graph is given in Figure 1. Corresponding to each of the two loci is a *marginal* coalescent tree, implicit in the structure of the graph. The tree corresponds to the ancestry of that particular locus. Such marginal tress are

Figure 2: The two marginal trees corresponding to the graph in Figure1



described by the normal coalescent process. However, the two trees are correlated. To construct the embedded tree for the left (right) locus, trace up from the bottom of the graph, turning left (right) whenever a recombination event is reached, and continuing until there is only a single line of ancestry. The point on the graph at which a single line of ancestry is attained for the given marginal tree corresponding to the left (right) locus, is called the Marginal MRCA for the left (right) locus. In Figure 2 we illustrate the two marginal trees corresponding to the graph in Figure 1. Note that the marginal MRCAs may be attained before reaching the MRCA of the entire recombination graph. While the marginal trees can be recovered from the recombination graph, the converse is not always true. The algorithms discussed in this paper operate on the space of possible ancestral recombination graphs, rather than the space of marginal trees. The introduction of recombination substantially increases the size of the sample space we are exploring. Whereas the number of tree topologies is finite, the number of ancestral recombination graph topologies is infinite because we can add an indefinite number of recombination events. However, graphs with a large number of recombinations are, *a priori* relatively unlikely, so the space

of *reasonable* graphs, is rather smaller. This becomes less true as ρ increases, so the efficiency of the algorithm presented here will decrease as ρ increases. It seems likely that as ρ increases it will be necessary to use methods based on the space of marginal trees (which are independent for $\rho = \infty$.)

We are interested in estimating the genotypes at the disease locus, denoted by $\mathbf{D}$, and the penetrance function for the disease gene, denoted by $\mathbf{P}$, conditional on the observed marker data, $\mathbf{M}$, and phenotypes $\mathbf{Y}$. Without loss of generality, we assume the left-hand locus is the marker, and the right-hand locus is the putative disease gene.

2 A Markov chain Monte Carlo Approach

2.1 The MCMC Method

More formally, we assume a prior distributions for the parameters of interest and employ an MCMC method to generate observations from the conditional density $f(\Lambda, T, \mu, \rho, \mathbf{P}, \mathbf{D} \mid \mathbf{M}, \mathbf{Y})$.

Letting $G = (\Lambda, T, \mu, \rho, \mathbf{P})$, we write

$$\begin{aligned} f(G, \mathbf{D} \mid \mathbf{M}, \mathbf{Y}) &\propto \mathbb{P}(\mathbf{Y} \mid \mathbf{D}, \mathbf{M}, G)\mathbb{P}(\mathbf{D}, \mathbf{M} \mid G)f(G) \\ &= \mathbb{P}(\mathbf{Y} \mid \mathbf{D}, \mathbf{P})\mathbb{P}(\mathbf{D}, \mathbf{M} \mid G)g_1(\Lambda)g_2(T)g_3(\mu)g_4(\rho)g_5(\mathbf{P}). \quad (1) \end{aligned}$$

The first term on the right of (1) is a simple function of the penetrance probabilities. The second term can be computed using a peeling algorithm (cf. Felsenstein 1981) and an appropriate mutation model. The term $g_1(\Lambda)$ is the coalescent tree topology distribution, $g_2(T)$ is the density of the coalescence times. We give details of these later. The terms $g_3(\mu)$, $g_4(\rho)$ and $g_4(\mathbf{P})$ are the prior distributions for the mutation rate, recombination rate and penetrance probabilities, respectively. The normalizing constant $f(\mathbf{M}, \mathbf{Y})$ implicit in (1) is not known, so we use a Metropolis-Hastings algorithm to simulate from the required conditional distribution.

The algorithm we develop uses a Markov process that produces correlated samples from a distribution π of interest, in our case

$$\pi(G, \mathbf{D}) \equiv f(G, \mathbf{D} \mid \mathbf{M}, \mathbf{Y}).$$

Let $X(t)$ denote the state of this chain after t iterations. We initialize the process with an arbitrary choice of Λ, T, μ, $\mathbf{P}$ and $\mathbf{D}$. New realizations of $(G, \mathbf{D})$ are then proposed according to a transition kernel $Q(\cdot)$, details of which are given below. For proposals that do not alter the dimension of the current state, the new realization is accepted with probability

$$h = \min\left\{1, \quad \frac{f(G', \mathbf{D}' \mid \mathbf{M}, \mathbf{Y}) Q(G', \mathbf{D}' \to G, \mathbf{D})}{f(G, \mathbf{D} \mid \mathbf{M}, \mathbf{Y}) Q(G, \mathbf{D} \to G', \mathbf{D}')}\right\}. \tag{2}$$

If the proposed state is not accepted, the previous state is retained. Proposals which do alter the dimension of the current state, i.e. the addition or removal of a recombination, work somewhat differently. An extra term is included in (2) to reflect the Jacobian of the proposed transitions. We discuss this later, when we give specifics of our proposed transition mechanisms

Once $X(t)$ has reached stationarity, samples from it are equivalent to samples from the distribution $\pi(G, \mathbf{D}) = \pi(\Lambda, T, \mu, \rho, \mathbf{P}, \mathbf{D})$. Note that consecutive outputs will be correlated. If independent samples are desired, we sample output from every l^{th} iteration, for an appropriate choice of l. The chain $X(t)$ must be irreducible and aperiodic to ensure that the limiting distribution is indeed $\pi(G, \mathbf{D})$. Informally this means that the chain must be able to visit all possible states in a finite period of time.

Related approaches to MCMC in the context of recombination are given in Kuhner *et al.* (2000) and Nielsen (1999). Griffiths and Marjoram give a method based on importance sampling. A critique of these approaches, in the case of no recombination, is given in Stephens and Donnelly (2000).

Using the expression given in (1) we see that calculation of (2) involves enumeration of the term $\mathbb{P}(\mathbf{D}, \mathbf{M} \mid G)$, where $\mathbf{D}$ denotes the currently accepted genotypes at the disease locus. This is performed in a manner analogous to Markovtsova *et al.* (2000ab). This follows since the recombination graph decomposes into two embedded trees, corresponding to the two loci, and these trees are peeled in the normal way. Conditional on (Λ, T, μ), the data at the two loci are independent. Therefore, we can write $\mathbb{P}(\mathbf{D}, \mathbf{M} \mid G) = \mathbb{P}(\mathbf{D} \mid G)\mathbb{P}(\mathbf{M} \mid G)$. Note also that it is sufficient to stop peeling the tree corresponding to a locus once the MRCA for the locus has been reached, i.e. at the top of its embedded tree, rather than at the MRCA of the recombination graph. Unconditionally on the data, the types of the MRCA for the

two embedded trees are independent and are distributed according to the stationary distribution of the mutation process.

It remains to specify the transition kernel $Q(\cdot)$ that we use to explore the state space. We do this in the appendix. Informally speaking, the kernel makes local rearrangements to the graph including the addition of a recombination event and subsequent new line of ancestry; removal of a randomly chosen existing recombination; small changes to mutation rate, recombination rate, and penetrance probabilities; changes to the genotypes at the disease locus.

3 Discussion

We have extended the approach outlined here to data for multiple markers. Consideration of multiple markers raises the issue of haplotyping, since marker data are normally obtained in the form of genotypes on a locus-by-locus basis. While the phase of the genotypes is not generally known, it is often possible to infer them from the genotypes of related individuals, as is done in pedigree analysis. In our context, two solutions suggest themselves. One, we can construct random samples by taking unrelated individuals from (multiple) pedigrees, for which the haplotypes can be inferred. Two, we can attempt to infer the haplotypes from genotypic data. Such a methodology will need to mix over the space of possible phases. Our early experiences with this suggest that mixing is problematic when the mutation rate is low. We have also adapted our approach to model the case of a continuous phenotype for the disease (see Marjoram et al. (2000) for the flavor of this). We note that our approach can also be used for segregation analysis. Let p_{00}, p_{01}, p_{11} denote the respective probabilities that genotypes 00, 10, 11 at the disease locus exhibit the disease. One can, for example, test whether the disease is recessive by comparing a run in which the penetrances are constrained such that $p_{00} = p_{01} = 0$, with a run in which the penetrances are unconstrained.

4 Literature Cited

Felsenstein, J., 1981 Evolutionary trees from DNA sequence data: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

Green, P., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **86:** 711–732

Griffiths, R. C. and P. Marjoram, 1997 An ancestral recombination graph, in *"Progress in Population Genetics and Human Evolution"* (P. Donnelly and S. Tavaré, Eds.), IMA Proceedings, **87**, Springer-Verlag, Berlin/New York.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Pop. Biol. **23:** 183–201.

Kingman, J. F. C., 1982 On the genealogy of large populations. J. Appl. Prob., **19A:** 27–43.

Kuhner, M. K., J. Yamato and J. Felsenstein, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

Kuhner, M. K., J. Yamato and J. Felsenstein, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393–1401.

Marjoram P., L. Markovtsova and S. Tavaré, 2000 "I see dead people:" gene mapping via ancestral inference. *Genetic Analysis Workshop 12*, submitted.

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

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

Nielsen, R. 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.

Nordborg, M. 2000 Coalescent Theory, in *"Handbook of Statistical Genetics"* (Balding D. J., M. J. Bishop and C. Cannings, Eds.), 179–208. John Wiley & Sons, Inc., New York, New York.

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

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

Wall, J. 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

Wiuf, C. and J. Hein 1999 Recombination as a point process along sequences. Theor. Pop. Biol. **55:** 248–259.

5 Appendix - Details of the Metropolis-Hastings Scheme

We now give details of the transition kernel $Q(\cdot)$ we use for MCMC on recombination graphs. At each iteration of the algorithm, the proposed transition is chosen in two steps: firstly we decide what type of transition to attempt; secondly we propose a transition of the appropriate type. There are seven possible types of transitions. According to the context of the scenario of interest, new states are proposed using a subset of these seven possible transitions. For example we might not allow the mutation rate to vary if this is not of interest. At a given iteration the proposal kernel chooses the type of transition according to an arbitrary distribution. So, a transition of type i is chosen with probability p_i, where $\sum p_i = 1$. The possible types are:

1. Perform a local rearrangement of the graph;
2. Add a recombination to the graph;
3. Remove a recombination from the graph;
4. Change the mutation rate;
5. Change the recombination rate;
6. Change the penetrance probabilities;
7. Change the genotypes at the disease locus.

Note that not all transitions are always allowed. For example, transitions 3 and 4 are not possible if there are no recombinations on the current graph, so these probabilities are set to zero in this case. We now give details of each of the possible transitions. Note that in all cases in which new times are proposed, we generate the new times according to exponential distributions with parameter $k(k-1+\rho)/2$, where k is the number of lines of ancestry on the level we are altering, and ρ is the currently accepted value of the recombination parameter. As an alternative we could use some other distribution, such as a truncated Normal centered around the currently accepted time. Such an alternative will work better when the times supported by the data are very different from the times given by the unconditional coalescent prior.

Transition 1. Local rearrangement of graph. Define an 'event' on the graph to be a coalescence or a recombination, and suppose there are L events on the current graph. For convenience we label the events in time order, moving down the graph. (So event L is the last event to occur before the sample is drawn, $L-1$ is the penultimate event, etc.) Let T_l denote the time between events $l+1$ and l. We begin by picking a event, l, according to an arbitrary distribution, typically uniform. We then propose a rearrangement to events l and $l-1$. Having done this we propose new times, T_l' and T_{l-1}', for the events. If there are k lines of ancestry on the graph just below l, then T_l' has an exponential distribution with parameter $k(k-1)/2$. If events l and $l-1$ are both formed by coalescence events we make changes exactly as given in Markovtsova *et al.* (2000). If the events are both formed by recombination there are two possible cases, according to whether the recombination at event $l-1$ involves one of the lines created by the recombination at event l. Suppose we label the two lines created by the recombination at l as A and B. If the recombination at $l-1$ occurs to line A then we simply move it so that it occurs to line B instead, keeping the recombination breakpoint unchanged. We make the converse change if the recombination originally occurred to line B. This is illustrated in Figure 3. If the recombination at $l-1$ involves neither of lines A or B we simply switch the order of the two recombinations. This is illustrated in Figure 4. The remaining cases involve situations in which events l and $l-1$ are created by one coalescence and one recombination. If event $l-1$ does not involve a line created by event l we again switch the order of the two events. This is shown in Figure 5. If this is not the case, so that

Figure 3: Example of proposed transition

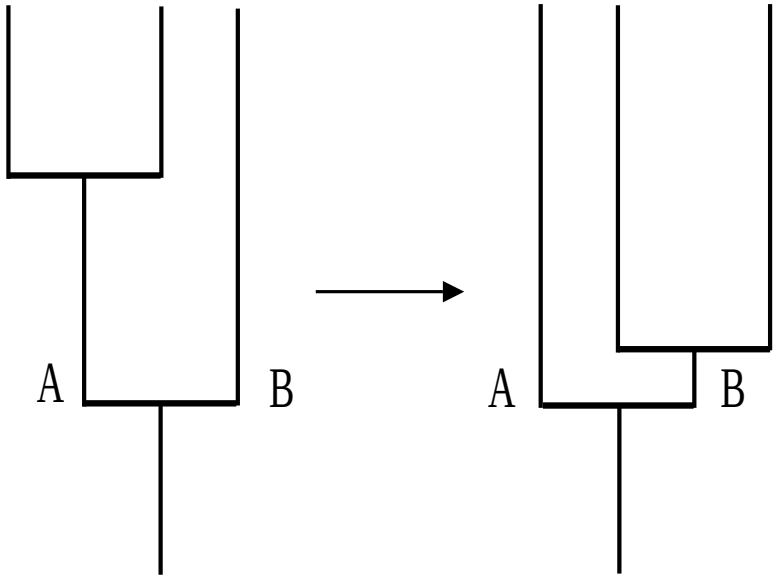


Figure 4: Example of proposed transition

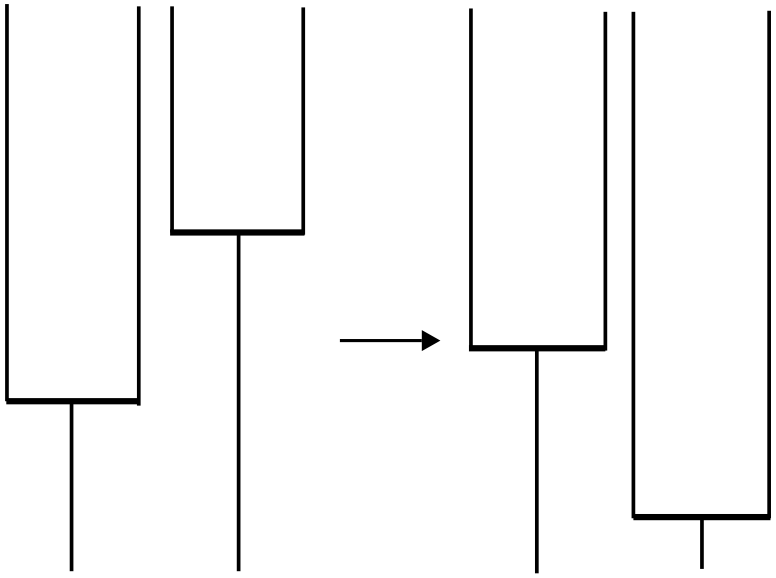
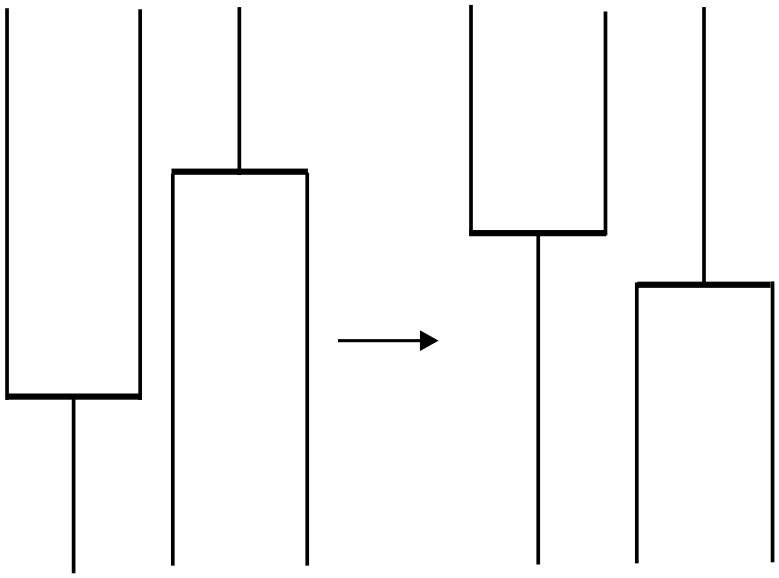


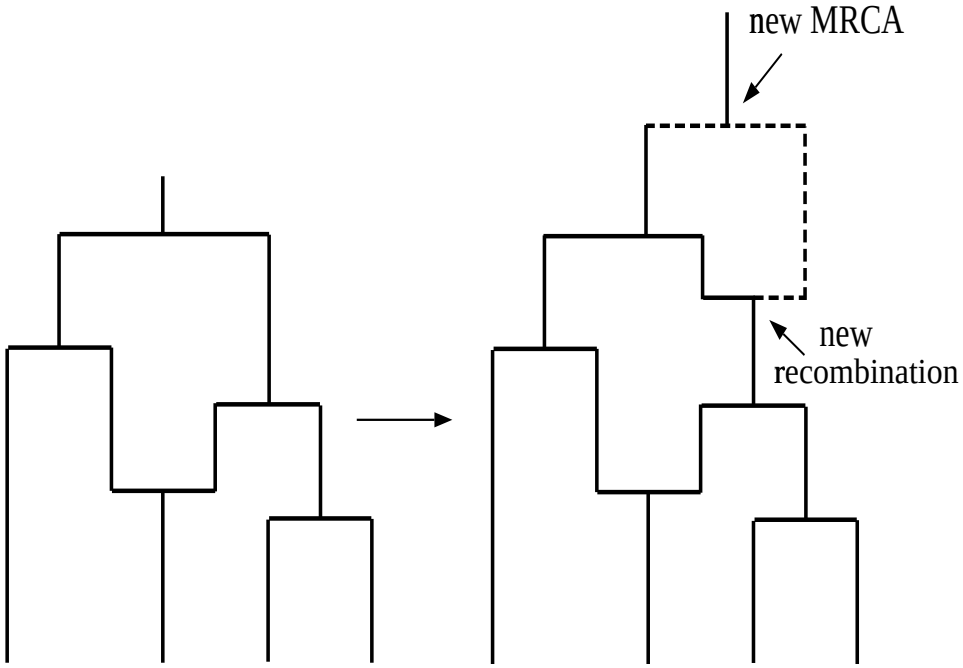
Figure 5: Example of proposed transition



one (or more) of the lines created at event l is also involved in event $l-1$, we make no change to the topology. Note that in cases for which the topology is not changed we still propose new times T_l' and T_{l-1}'.

Transition 2. Add a recombination. We begin by picking a point uniformly at random from the entire graph. With probability 1/2 the left-hand locus follows a new path, whereas the right-hand locus follows the previously existing path, otherwise the converse is the case. The new path is chosen as follows. Suppose the recombination occurs just below event l. We consider events one at a time, starting with event l, to determine at which point the newly created line of ancestry will coalesce with the rest of the graph. Suppose the current event is m, and that there are k lines just below m, we coalesce the new line to one of these lines with probability $2/(\rho + k)$. In such a case we pick one of the k lines uniformly at random to be the one with which the new path coalesces. Otherwise we move to the next highest event. Continue in this way until the new path has coalesced, or until we have moved past the MRCA of the graph. In the latter case we coalesce the new line with the MRCA at a new event, above the MRCA (thereby creating a new MRCA). Denoting the newly created

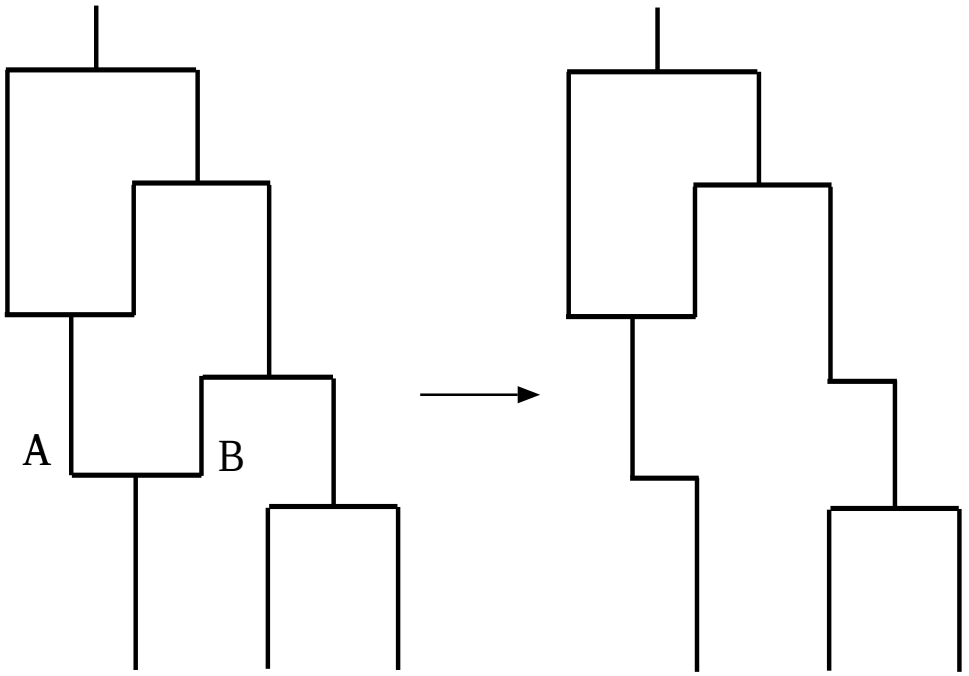
Figure 6: Example of the addition of a recombination



event at which recombination was attached by l' and the event at which the new path was re-attached by l'_r, we then propose new times $T_{l'}, T_{l'+1}, T_{l'_r}$ and $T_{l'_r+1}$. The choice of distribution used to propose these times will be reflected in the Jacobian term used in calculating the acceptance probability for the new state. Thus, we see that when we add a recombination to the graph the new topology is identical to the old, except that it has an extra path added. This will influence the way we propose the removal of recombinations, since all transitions must be reversible. The probabilities with which we choose the point at which to coalesce the new line, resulting from the recombination, are chosen to be the unconditional prior probabilities appropriate for recombination graphs, conditioned on the line not undergoing a further recombination before it first coalesces. It is expected this will aid the performance of the algorithm. This transition is illustrated in Figure 6

Transition 3. Remove a recombination. We pick a recombination uniformly, at random, from among those currently on the graph. We then uniformly, at random,

Figure 7: Example of a recombination removal



pick one of the two lines, A and B say, emerging from this recombination event. Without loss of generality suppose we pick line A. We now look at the next event line A experiences, moving up the graph. If it is a coalescence we remove the line (*i.e.* remove all parts of it between the recombination and coalescence events). Otherwise we leave the line unchanged, thus ensuring the reversibility of the process (*cf.* Adding recombination). If we remove the line, all loci now follow the path given by the remaining line, B. In this case there are now two fewer events on the proposed graph. In cases where we remove a recombination we propose new times as follows. Denote the removed events by l_1 and l_2. In the proposed graph we generate new times for T_{l_1} and T_{l_2}. In other words, we propose new times for the events just above the two that have been removed. This is illustrated in Figure 7. Note that it is possible that the MRCA of the proposed graph will be attained more recently than before. Such cases would result in non-reversible transitions, so the recombination, at l say, is not removed, and we simply propose a new time T_l.

Transition 4. Change the mutation rate. We propose changes to the mutation parameter μ in the following way. A new value μ' is proposed according to a Normal distribution centered about the currently accepted value μ, truncated so that μ' is always positive. This distribution has an arbitrary, fixed variance σ^2. Choice of σ^2 affects the behaviour of the chain $X(\cdot)$. If σ is 'small', the chain $X(\cdot)$ is heavily auto-correlated, and will take a long time to mix properly. It σ is 'large', then μ' is liable to be so far removed from μ that it will be rejected very frequently, leading to the same problems. The exact definitions of 'small' and 'large' are data dependent, but we find that a variance which is approximately the square of μ works well. Full details of these points are given in MARKOVTSOVA *et al.* (2000).

Transition 5. Change the recombination rate. We do this in a way that is exactly analogous to changes in the mutation rate, using a Normal distribution centered around the currently accepted value ρ. This leads to exactly the same considerations as those for μ.

Transition 6. Alter the penetrance probabilities. Let p_{00}, p_{01} and p_{11} be the probability that an individual exhibits the disease phenotype, given that its genotype at the disease locus is 00, 01 and 11 respectively. We assume that $p_{00} \leq p_{01} \leq p_{11}$ (although there is no requirement to do so). Particular disease models can be accommodated by restricting the sample space of possible values for p_{00}, p_{01} and p_{11}. For example, to model a gene that is known to be recessive we restrict the state-space so that $p_{00} = p_{01} = 0$. If a gene is known to be dominant, then we restrict the state-space so that $p_{01} = p_{11} = 1$.

Our algorithm makes changes to the penetrance probabilities by picking a single one of the penetrance parameters, p_{ij} and proposing a new value p'_{ij} according to a uniform distribution defined on the interval [0,1], but restricted so that the three penetrance probabilities will still be properly ordered.

Transition 7. Change the genotypes at the disease locus. We use two different moves in order to propose a change of this type. With probability 1/2 we randomly pick one of the $2n$ haploid sequences. Suppose its current type at the disease locus is i. Then we simply propose a new type j at this locus, where $j \sim \text{Unif}(\mathcal{S})$, and $\mathcal{S}$ is the space of possible types at the disease locus. Otherwise, we pick a node on the graph at random among the all internal nodes and change the type of its

ancestors independently, each with probability 0.5 to a different type j, where again $j \sim \text{Unif}(\mathcal{S})$.