July 5, 1995

Ancestral inference from samples of DNA sequences with recombination

R. C. Griffiths and P. Marjoram,*

Mathematics Department, Monash University,
Clayton, 3168, AUSTRALIA.

Phone: 61 3 9905 4416
Fax: 61 3 9905 4434

*Corresponding author.

Abstract. The sampling distribution of a collection of DNA sequences is studied under a model where recombination can occur in the ancestry of the sequences. The infinitely-many-sites model of mutation is assumed where mutation is supposed to always segregate a new mutant site. Ancestral inference procedures are discussed for : estimating recombination and mutation rates; estimating the times to the most recent common ancestors along the sequences; estimating ages of mutations; and estimating the number of recombination events in the ancestry of the sample. Inferences are made conditional on the configuration of the pattern of mutations at sites in observed sample sequences. A computational algorithm based on Markov Chain Monte-Carlo is developed, implemented, and illustrated with examples for these inference procedures. This algorithm is very computationally intensive.

Keywords. Ancestral inference, Markov Chain Monte-Carlo, Population genetics, Recombination, Samples of DNA.

Running head. Ancestral inference with recombination.

Introduction.

If a sample of genes is taken from a population in a model where the infinitely-many-alleles model is assumed then the sampling distribution for a configuration of alleles is the well known Ewens' sampling formula derived in Ewens (1992). The infinitely-many-sites model is a refinement of the infinitely-many-alleles model where the base structure of the genes is known. If distinct sequences are labelled as different alleles then the allele configuration is distributed as Ewens' distribution. In both these models mutation is assumed to produce a new allele type, equivalent to segregating a new site at the finer level. The distribution of the number of segregating sites was found by Watterson (1975), but an explicit form for the complete sampling distribution of sequences is unknown. Studying the sampling distribution of sequences in this model is equivalent to studying distributions on genealogical trees that can be deduced from the mutation configuration at bases (Ethier et al. 1987, Griffiths 1989, Griffiths et al. 1995a). These trees are derived from a condensation of the coalescent tree with mutations, have vertices as mutations and are without a time scale. It is possible to find a recursion for the sampling distribution in terms of a distribution on trees. Griffiths et al. (1994b) exploit this recursion by finding a representation for the likelihood of a sample of sequences which then allows a Markov Chain Monte-Carlo simulation method to be applied to estimate the likelihood. The technique also generates likelihood curves for the scaled mutation rate θ, allowing maximum likelihood estimation of it from observed data. It is also possible to compute related ancestral distributions using a simulation technique. One of great interest is the distribution of the time to the most recent common ancestor of the sequences, conditional on the data. It is easy enough to compute the unconditional time to the most recent common ancestor distribution, but conditioning on the observed data makes it a much harder problem. Another model of sequences with a finite number of bases where back mutation is possible is studied in Griffiths et al. (1994a) and a similar Markov Chain Monte Carlo method implemented to find the likelihood of samples. Although this model seems more realistic, the infinitely-many-sites model is really better in that it uses the natural ancestral tree structure in a very efficient way. Variable population size models are studied in Griffiths et al. (1994c).

In this paper a model where recombination can occur within sequences is studied, a recursion for the sampling distribution of sequences found, and a Markov Chain Monte-Carlo scheme developed and implemented to obtain the likelihood of a sample of sequences. Joint maximum likelihood estimates of the mutation and recombination rates can then be found from a likelihood surface. We stress that this uses the maximal information in the sample, rather than using summary statistics such as the number of segregating sites. Related ancestral distributions, *conditional on the observed data*, can also be found numerically by using this scheme. Of interest are methods to compute the time to the most recent common ancestor at points on the sequence, and estimate the recombination events occurring to the sample's ancestors.

It is possible to understand the computer program interface and output without fully understanding the theory and algorithm for finding likelihoods, so we hope that readers concerned about the complexity of the theory will skip to details of the program, and the example output.

A gene is represented as a continuous length of DNA, denoted by the unit interval $[0, 1]$. We wish to model the evolution of a population of such genes. The model used here is a neutral one in which recombination and mutation events occur. The population is assumed to be evolving through discrete generations in a Wright-Fisher-like manner, each generation being of constant size $2N$. The model is a haploid one, but a diploid model in a random mating population essentially behaves like this haploid model.

If there is no recombination in such a model then the ancestry of a sample of genes can be described by the *coalescent process* in Kingman (1982). This traces the ancestral tree or *genealogy* of the sample back through time. When there is recombination the analogue of the coalescent process is an *ancestral recombination graph*, described in Griffiths et al. (1995). The ancestry is no longer a tree. When recombination events occur to an individual which is an ancestor of the sample, the genealogy bifurcates (ie. the individual has two ancestors). Thus a graph is obtained rather than a tree.

Specifically, in our model genes choose their parents from the previous generation according to the following scheme:

With probability $1 - r$ a single parent is selected uniformly at random from the previous generation;

With probability r a recombination event occurs and so two parents are chosen uniformly at random.

Each gene in the next generation chooses one or two parents in this manner (independently of all other choices). The collection of these $2N$ offspring genes forms the next generation.

If recombination occurs a position for its location, Z, is chosen (independently from the location of other such events) according to a given distribution, and the offspring gene is formed from the lengths $[0, Z]$ and $[Z, 1]$ from the first and second parents respectively. Both of the parents are regarded as ancestors of the offspring and therefore of any individual in a (forward) line of descent of the offspring. Here Z is taken to have a continuous distribution on $[0, 1]$, where breaks are possible at any point in $[0, 1]$. We may choose to use a Uniform distribution, to model a situation in which the recombination rate is constant along the gene, or use other distributions if we wish to model varying recombination rates, hotspots, or other features.

As is usual, time is measured in units of $2N$ generations and we let $N \to \infty$. The recombination rate per gene per generation r is scaled in the normal way by holding $\rho = 2Nr$ fixed.

Offspring are also subject to mutation events at rate u per gene per generation, which is similarly scaled by setting $\theta = 2Nu$. The infinitely many sites model of mutation is assumed, so that mutation never occurs at the same site twice.

We assume the population is stationary and draw a random sample from it at the present time. An observed sample of sequences is distinguished by its segregating sites, where there are two types of bases, together with the identity of the genes carrying each of the two types at such sites. Thus, assuming that the wild site types (ie. the type of the most recent common ancestors (MRCAs) of the bases) are known, each gene in a sample can be described by positions of mutant types within the scale $[0, 1]$, and thus the sample is described by a collection of such positions. This continuous model has been studied in Hudson (1983), (1991), and Kaplan et al. (1985), Griffiths et al. (1995).

Figure 1 illustrates a recombination graph for a sample of n genes. Dots on the edges indicate mutations occurring to ancestors and such points are labelled with the location of the mutation. Looking back in time, coalescences occur when two edges join to a vertex, and recombinations occur when one edge joins to two. Positions $Z_1, Z_2, \ldots$ where breaks occur are labelled on the graph (as a,b, c and d).

Let $\{\xi(t), t \geq 0, \xi(0) = n\}$ denote the number of ancestors of a sample of n back in time. This is a birth and death process with respective rates $\lambda_k = k\rho/2$ and $\mu_k = k(k-1)/2$. Because of the quadratic death rate compared to the linear birth rate, with probability 1 there is a MRCA in the graph. It is implicit that the process is defined backward in time to negative infinity.

Mutations occur according to a Poisson process along the edges of the graph at rate $\theta/2$, and are chosen at random uniformly within the sequence. Whether a mutation appears in a sample sequence depends on whether the point it falls on is ancestral to the sequence. Not all genes in previous generations carry DNA which is ancestral to the sample. It is clear that genes not in the recombination graph will carry no ancestral DNA. However it is also true that, even for genes in the ancestral graph, only part of their DNA may be ancestral. Whenever we observe a recombination event in the ancestral graph the offspring gene consists of a subset from the DNA each of its two parents. Thus if a mutation carried by the parent is not within the chosen subset it will not be passed on.

Figure 1.

Each point $x \in [0,1]$ has a coalescent tree $\mathcal{T}(x)$ associated with its ancestry. This tree traces the ancestry of the sample at that particular point. These trees are imbedded in the recombination graph. To obtain $\mathcal{T}(x)$ trace from the bottom of the graph upward toward the MRCA in the graph. If there is a recombination vertex with label z, take the left path if $x \leq z$, or right path if $x > z$. The MRCA in $\mathcal{T}(x)$ may occur in the graph before the grand MRCA. Figure 2 shows an example of $\mathcal{T}(x)$ when $x > b$ and $x < c, d$.

Figure 2.

Since recombination events do not affect the ancestry of a tree $\mathcal{T}(x)$ the marginal distribution of the tree is the same as if it were from a single coalescent process. It follows that the time to the most recent common ancestor (TMRCA) in $\mathcal{T}(x)$ is marginally distributed as the time to absorption in a death process with rates $\{\mu_k, k \geq 2\}$ starting at n, the sample size. It also follows that the expected number of segregating sites in the sequences is $\theta \sum_{j=1}^{n-1}(1/j)$, which is independent of the recombination rate (Hudson 1983).

An urn scheme for simulating samples is described in Griffiths et al. (1995), which is adapted from a two-locus method of Ethier et al. (1990). This scheme simulates the order of coalescent, recombination, and mutation events back in time as an imbedded chain, then develops the sample forward as a second step starting from the MRCA.

Consider n ancestor genes of d distinct types of a sample in the recombination graph at a fixed time back. Material ancestral to the sample on these genes is a collection of intervals $\mathcal{A}_i = \{A_{i\alpha}; \alpha = 1, 2, \ldots\}, i = 1, \ldots, d$. Mutations in sample genes included in the ancestral material are denoted by $\mathcal{M}_i = \{M_{i\alpha}; \alpha = 1, 2, \ldots\}, i = 1, \ldots, d$. That is, $M_{i\alpha}$ is the collection of mutation points in the ancestral material $A_{i\alpha}$ on ancestor gene i. Multiplicities of genes are denoted by $\mathbf{n}$. The ancestor state is then described by

$$\mathbf{A} = \{\mathcal{A}_i\}, \mathbf{M} = \{\mathcal{M}_i\}, \mathbf{n} = \{n_i\}.$$

$\Big(\mathbf{A}(t), \mathbf{M}(t), \mathbf{n}(t), t \geq 0\Big)$ is a Markov process looking back in time, with entries in $\mathbf{A}(0)$ being complete sequences $\{[0,1]\}$, $\mathbf{M}$ describing the current mutation points and $\mathbf{n}(0)$ multiplicities of the sequences.

Items 1 to 4 in the next proposition detail how transitions are made in the process, according to whether there is a coalescence, mutation or recombination event immediately back in the recombination graph.

As an example consider a sample of four sequences shown on the left in Figure 3, which was generated from the recombination graph in Figure 1, with $a = 0.9$, $b = 0.4$, $c = 0.3$, $d = 0.7$, and has mutations at positions at 0.2, 0.6, 0.8. Five ancestors of this sample taken at a cross section of the graph just below the mutation at 0.2 are also shown on the right of Figure 3. The ancestral fragments of the sequences are, from the top of the diagram, (0.0,0.4); empty sequence; (0.0,1.0); (0.0,1.0); (0.7,1.0). Mutations are only carried by the 3rd and 5th ancestor sequences.

Figure 3.

Let $I_j(x)$ equal 1 if there is a mutation at position $x \in [0,1]$ for sampled gene j, and equal 0 otherwise. If there is no recombination a necessary and sufficient condition that a collection of points represents a sample of sequences with mutations is that in no two locations $x, y \in [0,1]$ do there exist three genes j, k, l such that:

$$\begin{array}{ll} I_j(x) = 1, & I_j(y) = 1; \\ I_k(x) = 1, & I_k(y) = 0; \\ I_l(x) = 0, & I_l(y) = 1. \end{array}$$

The example above violates the condition at the points 0.2, 0.6, and one deduces that there must have been a recombination event between 0.2 and 0.6 in the sample's ancestry.

When recombination is present it is theoretically possible to have any pattern of mutations.

If it is not known which types are wild at segregating sites, then if there is no recombination and we observe s segregating sites there are $s+1$ rooted genealogical trees corresponding to the unique unrooted tree constructed from the data, depending on where the root actually is in the tree. Changing the root from one position to another toggles which bases are wild and mutant between the two potential root positions. This concept is discussed in Griffiths et al. (1995a). If recombination in the model is possible anywhere along the sequences, then theoretically any of the 2^s possibilities for wild types at segregating sites are possible, instead of $s+1$ with no recombination, though some of the configurations may have a relatively small probability.

Sampling distribution of sequences.

In Griffiths et al. (1994b,1995a) a recursion for the probability of a genealogical tree is derived by considering the next event back in time in the coalescent tree.

The next proposition gives a recursion for the sampling distribution of sequences when recombination is possible by considering the next event back in time in the recombination graph, which could be coalescence, recombination or mutation. It is necessary to consider a state space which includes subsets of sequence material because recombination ancestors of a gene only contain part of the gene's material.

Let $(\mathbf{A}, \mathbf{M}, \mathbf{n})$ be a collection of fragments of sequences with distinct mutations $x_1, \ldots, x_m$ in $\mathbf{M}$ and distinct end-points of intervals $a_1, \ldots, a_r$ in $\mathbf{A}$. This will represent a configuration of material ancestral to the sample taken at a cross-section of the ancestral recombination graph. Let $Q(\mathbf{A}, \mathbf{M}, \mathbf{n})$ be the joint density of observing these mutations and end points for a fixed known fragment configuration $\mathbf{A}$, taken from a stationary population. Although $Q(\mathbf{A}, \mathbf{M}, \mathbf{n})$ will represent an ancestral configuration, it is defined just in terms of given fragments, taken from a stationary population, and not relative to an initial sample.

Proposition.

$$\begin{aligned}
&(n(n-1) + a\theta + b\rho)Q(\mathbf{A}, \mathbf{M}, \mathbf{n}) \\
&\quad = n\sum_{1}(n_i - 1)Q(\mathbf{A}, \mathbf{M}, \mathbf{n}_i) \\
&\quad + 2n\sum_{2}(n_k + 1 - \delta_{ik} - \delta_{jk})Q(\mathbf{A}, \mathbf{M}, \mathbf{n}_{ij}^{k}) \\
&\quad + \theta\sum_{3}(n_k + 1)Q(\mathbf{A}, \mathbf{M}_i(m_{i\alpha}), \mathbf{n}_i^{k}) \\
&\quad + \frac{\rho}{n+1}\sum_{4}\int(n_i + 1)(n_j + 1)Q(\mathbf{A}_k^{ij}(x), \mathbf{M}_k^{ij}(x), \mathbf{n}_k^{ij}(x))dx
\end{aligned} \tag{1}$$

The Kronecker delta is denoted by $\delta_{jk} = 1$ if $j = k$ or $\delta_{jk} = 0$ if $j \neq k$. Subscripts on $\mathbf{n}$ denote a decrease in the respective co-ordinates, while superscripts denote an increase. For example $\mathbf{n}_{ij}^{k} = \mathbf{n} - \mathbf{e}_i - \mathbf{e}_j + \mathbf{e}_k$, where $\mathbf{e}_j = (\delta_{jk})$, the jth unit vector. $a = \sum_{i=1}^{d} n_i|\mathcal{A}_i|$, the total fragment lengths, and $b = \sum_{i=1}^{d} n_i\Big(\max\{x; x \in \mathcal{A}_i\} - \min\{y; y \in \mathcal{A}_i\}\Big)$, the total amount of material where a recombination event affects the ancestry of the fragments in $(\mathbf{A}, \mathbf{n})$.

An explanation of the notation in (1), and a description of the summation regions and the immediate event back in time to the ancestors of the fragments from which the terms arise follows.

1. Coalescence of two genes of identical type. The summation is over $\{j; n_j > 1\}$.
2. Coalescence of two different genes. This can only occur if points ancestral in both sequences either both contain a mutation point, or neither does. That is for genes i, j and all α, β

$$\{x_\gamma; x_\gamma \in M_{i\alpha}, x_\gamma \in A_{i\alpha} \bigcap A_{j\beta}\} = \{x_\gamma; x_\gamma \in M_{j\beta}, x_\gamma \in A_{i\alpha} \bigcap A_{j\beta}\}.$$

After coalescence the kth gene is formed by taking $\mathcal{A}_k = \mathcal{A}_i \bigcup \mathcal{A}_j$, and similarly for the mutation points. Notation for $\mathbf{A}$, $\mathbf{M}$ is abused in that a new type may be created, in which case take $n_k = 0$ before creation of the type. It is possible that i or j is equal to k. Summation is over all appropriate unordered pairs (i, j).

3. Mutation in a fragment. Summation is over all singleton mutations (which could have been produced by the immediate event back in time). In the notation $(\mathbf{A}, \mathbf{M}_i(m_{i\alpha}), \mathbf{n}_i^k)$ the ith gene has a mutant point $m_{i\alpha} \in M_{i\alpha}$ removed and then becomes of type k. Take $n_k = 0$ if k is a new type. Type i must be a singleton type, and so is removed from the ancestor list.
4. Recombination to gene k at position x between the minimum and maximum ancestral points in $\mathcal{A}_k$ producing recombination ancestors i and j. n_i and n_j implicitly depend on x. The immediate ancestral configuration is denoted by $Q(\mathbf{A}_k^{ij}(x), \mathbf{M}_k^{ij}(x), \mathbf{n}_k^{ij}(x))$. If x occurs where there is no ancestral material, between $A_{k\alpha}$ and $A_{k\alpha+1}$ then genes i, j are such that $\mathcal{A}_i = \{A_{k\beta}; \beta \leq \alpha\}$, $\mathcal{A}_j = \{A_{k\beta}; \beta > \alpha\}$, and similarly for $\mathcal{M}_i$, $\mathcal{M}_j$. The types i, j may already exist in the ancestors. If x occurs where there is ancestral material, in $A_{k\alpha}$, then this set is split at the point x. Because of the assumption of a continuous distribution of recombination in this case genes i, j are unique in the ancestors, thus $n_i = 0, n_j = 0$.

Boundary conditions in (1) are that $Q(\mathbf{A}^\circ, \mathbf{M}^\circ, \mathbf{n}^\circ) = 1$ for configurations $(\mathbf{A}^\circ, \mathbf{M}^\circ, \mathbf{n}^\circ)$ where all fragments in $\mathbf{A}^\circ$ are disjoint and have multiplicity 1.

Proof.

Consider (1) written in the form

$$\begin{aligned}
Q(&\mathbf{A},\mathbf{M},\mathbf{n}) \\
=&\frac{(n-a)\theta+(n-b)\rho}{(n(n-1)+n\theta+n\rho)}Q(\mathbf{A},\mathbf{M},\mathbf{n}) \\
&+\frac{n(n-1)}{(n(n-1)+n\theta+n\rho)}\sum_1 \frac{(n_i-1)}{n-1}Q(\mathbf{A},\mathbf{M},\mathbf{n}_i) \\
&+\frac{2n(n-1)}{(n(n-1)+n\theta+n\rho)}\sum_2 \frac{(n_k+1-\delta_{ik}-\delta_{jk})}{n-1}Q(\mathbf{A},\mathbf{M},\mathbf{n}_{ij}^k) \qquad (2) \\
&+\frac{n\theta}{(n(n-1)+n\theta+n\rho)}\sum_3 \frac{(n_k+1)}{n}Q(\mathbf{A},\mathbf{M}_i(m_{i\alpha}),\mathbf{n}_i^k) \\
&+\frac{n\rho}{(n(n-1)+n\theta+n\rho)}\sum_4 \int \frac{(n_i+1)(n_j+1)}{n(n+1)}Q(\mathbf{A}_k^{ij}(x),\mathbf{M}_k^{ij}(x),\mathbf{n}_k^{ij}(x))dx
\end{aligned}$$

The previous event back in time in the ancestry of the fragments was coalescence, mutation or recombination at rates of $\frac{1}{2}n(n-1)$, $\frac{1}{2}n\theta$, $\frac{1}{2}n\rho$. Argue then that conditional on these events a configuration $\mathbf{A},\mathbf{M},\mathbf{n}$ occurs depending on the various ancestor configurations and which genes coalesce, mutate or recombine. Although the notation is awkward, the terms on the right hand side of (2) simply represent the probability of the configurations which lead to the configuration $Q(\mathbf{A},\mathbf{M},\mathbf{n})$ one step back in the recombination graph.

The first right hand side term represents the case when a mutation or recombination event does not affect the configuration. Note that the coalescent pair in the 3rd term is unordered, explaining the factor of 2, while the recombinant pair of genes in the 5th term is ordered (by convention in this model).

The argument above relies implicitly on stationarity and the consistency of subgraphs in the recombination graph. That is, the distribution of a subgraph of n_0 ancestors taken from a cross-section of the recombination graph is again distributed as a recombination graph of n_0 individuals.

Notice that the total amount of fragment material in the configurations on the right hand side of (1) is always less than or equal to the amount of material, a, in

$(\mathbf{A}, \mathbf{M}, \mathbf{n})$ on the left hand side of (1). Equation (1) is analogous to equations found by other authors for finite-locus models, for example in a two-locus model equation in with the infinitely-many-alleles model, Golding (1984), Ethier et al. (1990).

If there is no recombination in the model, then taking (1) with $\rho = 0$ and $a = n$ gives a recursion for the likelihood in the infinitely many sites model. This is similar to a recursion in Griffiths et al. (1995a), where there is also a discussion about the combinatorial arrangement of sites and what effect the ordering has on the likelihood. Mutation positions have a uniform distribution on $[0, 1]$ when $\rho = 0$.

Computing the likelihood of a sample.

$Q(\mathbf{A}, \mathbf{M}, \mathbf{n})$ can be computed by a method of Griffiths et al. (1994a b), where it is represented as the expected value of a functional on a Markov chain which moves backward in time to where the MRCA of each point on the sample sequences has been determined. $Q(\mathbf{A}, \mathbf{M}, \mathbf{n})$ is then estimated by taking the average functional value over repeated simulations of the process. A sketch of the representation follows.

Consider a Markov chain which has state space $\{(\mathbf{A}, \mathbf{M}, \mathbf{n})\}$. Transitions in the Markov chain are made to states indicated in the right hand side of equation (1). Denote

$$S = n\sum_{1}(n_i - 1) + 2n\sum_{2}(n_k + 1 - \delta_{ik} - \delta_{jk}) + \theta\sum_{3}(n_k + 1) + \frac{\rho c}{(n+1)},$$

where $c = \sum_{i=1}^{d}\Big(\max\{x; x \in \mathcal{A}_i\} - \min\{y; y \in \mathcal{A}_i\}\Big)$.

In this constructed Markov chain transitions are made from $(\mathbf{A}, \mathbf{M}, \mathbf{n})$ to :

$$\begin{array}{ll}
(\mathbf{A}, \mathbf{M}, \mathbf{n}_i) & \text{with probability } n(n_i - 1)/S; \\
(\mathbf{A}, \mathbf{M}, \mathbf{n}_{ij}^{k}) & \text{with probability } 2n(n_k + 1 - \delta_{ik} - \delta_{jk})/S; \\
(\mathbf{A}, \mathbf{M}_i(m_{i\alpha}), \mathbf{n}_i^{k}) & \text{with probability } \theta(n_k + 1)/S; \\
(\mathbf{A}_k^{ij}(x), \mathbf{M}_k^{ij}(x), \mathbf{n}_k^{ij}(x)) & \text{with probability } \rho c/\Big((n+1)S\Big),
\end{array} \tag{3}$$

where in the last type of transition x is chosen uniformly within $\bigcup_{i=1}^{d}\Big(\min\{x; x \in \mathcal{A}_i\}, \max\{y; y \in \mathcal{A}_i\}\Big)$, without regard to multiplicity of the sequences. Denote $X = (\mathbf{A}, \mathbf{M}, \mathbf{n})$, Y the state the chain moves to, and $f(x, y) = S/(n(n-1) + a\theta + b\rho)$, for transitions not of the last type, $f(x,y) = (n_i+1)(n_j+1)S/(n(n-1)+a\theta+b\rho)$, for transitions which are of the last type.

The process is absorbing at states where there first is a MRCA at all positions on the sample chromosomes. For such a configuration $(\mathbf{A}^0, \mathbf{M}^0, \mathbf{n}^0)$, $Q(\mathbf{A}^0, \mathbf{M}^0, \mathbf{n}^0) = 1$. This process is not a genuine genealogical one, though it does however follow up along a recombination graph, with quadratic rates of coalescence compared to a much smaller rate of recombination. The reason for choosing the last transition probability and f combination is for computational efficiency. Before each transition a program implementation must compute transition probabilities for all possible changes of state. If the last transition did follow the pattern of the others then it would be necessary to search for types in the sample which are possibly the same type as recombinant ancestors of the gene which is constructed by recombination. This would have to be done for every position on each gene. The scheme in (3) is much easier to implement.

Suppose that $X(k)$ is the state of the chain at steps $k = 0, \ldots, \tau$, where τ is the absorption time. Then (as in Griffiths et al. 1994 a,b) it is possible to express

$$Q(\mathbf{A}, \mathbf{M}, \mathbf{n}) = E\Big(\prod_{0}^{\tau-1} f(X(k), X(k+1))\Big). \tag{5}$$

$Q(\mathbf{A}, \mathbf{M}, \mathbf{n})$ can be estimated by repeatedly simulating the process and averaging the functional $\Big(\prod_0^{\tau-1} f(X(k), X(k+1))\Big)$ over replicates.

Actually a more detailed argument is required to show that (5) is a valid representation. Let $Q_r(\mathbf{A}, \mathbf{M}, \mathbf{n})$ be defined similarly to $Q(\mathbf{A}, \mathbf{M}, \mathbf{n})$ but with the restriction that the number of recombination events affecting the sample's ancestry before the MRCAs of the sample sequences be at most r. $Q_r(\mathbf{A}, \mathbf{M}, \mathbf{n})$ satisfies a similar equation to (1), with Q replaced by Q_r except for the last right hand side term where the replacement is Q_{r-1} if $r \geq 1$, or zero if $r = 0$. This modified version of (1) is a true recursive set of equations on a degree defined by $r + n + s$, where s is the number of segregating sites. $Q_0(\mathbf{A}, \mathbf{M}, \mathbf{n})$ is zero if $\mathbf{A}$ does not contain complete sequences lengths $[0, 1]$ or $\mathbf{M}$ is inconsistent with there being no recombination. The recursion is terminated when

$r = 0$ at singleton sequences $\mathbf{A}$ where $Q_0(\mathbf{A}, \mathbf{M}, \{1\}) = 1$ if $\mathbf{A} = \{[0,1]\}$, or zero otherwise. Let R be the number of transitions of the last type in (3), before absorption. A modified version of (5) derived from the recursion for Q_r is

$$Q_r(\mathbf{A}, \mathbf{M}, \mathbf{n}) = E\Big(I\{R \leq r\} \prod_0^{\tau-1} f(X(k), X(k+1))\Big), \tag{6}$$

where $I\{\cdot\}$ is the indicator function. Let $r \to \infty$ in (6). Using the monotone convergence theorem on both sides shows that (5) is true. It is not easy to argue that a solution to (1) is unique directly, but the modified version of (1) with Q_r does have a unique solution, so (5) is a valid representation obtained through this route.

If interest is centered on estimating θ, ρ from (5), then an entire likelihood surface can be generated by simulating a process with parameters θ_0, ρ_0, then expressing

$$Q_{(\theta,\rho)}(\mathbf{A}, \mathbf{M}, \mathbf{n}) = E_{(\theta_0,\rho_0)}\Big(\prod_0^{\tau-1} f(X(k), X(k+1); \theta_0, \rho_0, \theta, \rho)\Big), \tag{7}$$

where

$$\begin{aligned} f(X, Y; \theta_0, \rho_0, \theta, \rho) &= f_{\theta,\rho}(X, Y) \frac{p_{X,Y}(\theta, \rho)}{p_{X,Y}(\theta_0, \rho_0)} \\ &= \frac{S(\theta_0, \rho_0)\phi(X, Y)}{n(n-1) + a\theta + b\rho}, \end{aligned} \tag{8}$$

$\{p_{X,Y}(\theta, \rho)\}$ are transition probabilities, and $\phi(X, Y)$ takes values θ/θ_0 for mutation transitions, ρ/ρ_0 for recombination transitions and is unity otherwise. $S(\theta_0, \rho_0)$ is the variable S in (3) with parameters explicit. An entire likelihood surface for θ, ρ is returned for each simulation run of the process, which is generated with parameters θ_0, ρ_0. This technique is standard in Markov Chain Monte-Carlo methods. In practice the algorithm will be most accurate in the neighbourhood of the generating parameters.

The algorithm for computing the likelihood can also be enhanced to compute the distribution of quantities of interest, *conditional on the sample configuration*, such as the number of recombination events in the ancestry of a sample, and the MRCA times along the sequences (TMRCAs).

Recombination events in the ancestry of a sample.

Let R denote the distribution of the number of recombination events in material ancestral to the sample, before the last MRCA along the sequences. Griffiths et al. (1995) study aspects of the distribution.

A particular result is that $E(R) \leq 1 + \rho$.

The emphasis here is on computing the distribution of R, conditional on the observed sample configuration.

Let $(F_1, R_1), \dots, (F_k, R_k)$ denote realizations of $\left(\prod_0^{\tau-1} f(X(k), X(k+1)), R\right)$ over k simulation runs, then an empirical distribution for R is

$$P(R = j \mid \text{Sample configuration}) = \frac{\sum_{\{\ell : R_\ell = j, 1 \leq \ell \leq k\}} F_\ell}{\sum_{\ell=1}^{k} F_\ell}, \ j = 0, 1, \dots \tag{9}$$

There is a distinction between recombination events which fall in ancestral material, and those which do not, but which do influence ancestry of a sample. Both distributions can be estimated as in (9).

At a more detailed level the distribution of the number of recombination events in different regions, conditional on the observed data can be studied in a similar way.

Times to MRCAs along the sequence.

Information about times at which a particular event occurs in the ancestry of a sample such as the time to the last MRCA can be found by considering times between events in the recombination graph. The time between transitions from $(\mathbf{A}, \mathbf{M}, \mathbf{n})$ to a configuration on the right hand side of (1) is an exponentially distributed random variable with rate $\frac{1}{2}(n(n-1) + a\theta + b\rho)$. A single run estimate of the time to a particular event is the sum of exponential random variables with the above rates along the path to the event in a realization of the Markov chain with transitions probabilities given in (3).

There are a finite number of different MRCAs of a sample of sequences at positions along the sequence. Let $\{W(x), 0 \leq x \leq 1\}$ be the MRCA times along the sequence.

A method for computing an empirical finite-dimensional distribution of $\mathbf{W} = \big(W(x_1), \dots, W(x_m)\big)$ for $x_1, \dots, x_m \in [0, 1]$, conditional on the observed sample configuration is the following.

In each Markov chain simulation with functional F, simulate times between events in the Markov process corresponding to the observed imbedded chain with κ replicates and suppose that the κ MRCA times observed on the ith simulation are $\mathbf{w}_{ij}, j = 1, \dots, \kappa$. The empirical (discrete) distribution of $\mathbf{w}$ given the sample configuration is found from the $k\kappa$ simulation replicates. This empirical distribution takes values $\mathbf{w}_{ij}$, with probability $\frac{1}{\kappa} F_i / \sum_1^k F_\ell$, $i = 1, \dots, \kappa$, $j = 1, \dots, k$ corresponding to k Markov chain simulations with κ time replicates. The step of replicating Markov times for the imbedded chain is to improve accuracy in the estimated distribution.

An easier variation is to calculate the mean and variance functions $\{E(W(x)), 0 \leq x \leq 1\}$ and $\{\text{var}(W(x)), 0 \leq x \leq 1\}$, conditional on the data by calculating exactly the expected times and variances in the Markov process corresponding to each of the k simulation runs of the Markov chain. A weighted average is then taken, with respect to the functional values $F_j, j = 1, \dots, k$.

The mean ages of mutations in the sample, conditional on the data are also computed in a similar way.

Implementation of the likelihood algorithm.

The algorithm described above has been implemented as a discrete approximation by taking a large number L of base positions. An infinitely-many-sites model for the mutation process is assumed, thus if there are s segregating sites in the sequences then there are s mutations in material ancestral to the sequences. Recombination is taken to occur in the $L-1$ positions between the L bases. As an approximation to the continuous model recombination is only allowed to occur at most once in any position along the ancestor lines. This can be relaxed to allow multiple recombination at positions as an option.

The state space of the analogue of the process defined by (3) is $(\mathbf{B}, \mathbf{n})$, where $\mathbf{B}$ is a $d \times L$ matrix representing d sequence types, with multiplicities $\mathbf{n}$. A row of $\mathbf{B}$ has the form

$$0\ 0\ 1\ 0\ 0\ 1\ \circ\ 0\ \circ\ \circ\ 1\ 0\ 0\ 1\ \circ\ \circ\ 0\ 1$$

where 0 denotes the MRCA base type of a site, 1 a mutant type since the MRCA, and $\circ$ an undetermined type.

Possible transitions are related to :

Coalescence between like types in row i, where $n_i \to n_i - 1$;

Coalescence between types in rows i and j where no two entries in any column ℓ satisfy $\mathbf{b}_{i\ell} = 0, \mathbf{b}_{j\ell} = 1$ or $\mathbf{b}_{i\ell} = 1, \mathbf{b}_{j\ell} = 0$, coalescing to a sequence $\mathbf{b}'$ such that

$$\mathbf{b}'_\ell = \begin{cases} 0 & \text{if } \mathbf{b}_{i\ell} = 0 \text{ or } \mathbf{b}_{j\ell} = 0, \\ 1 & \text{if } \mathbf{b}_{i\ell} = 1 \text{ or } \mathbf{b}_{j\ell} = 1; \end{cases}$$

Mutation, where a singleton 1 in a column of $\mathbf{B}$ where the multiplicity of the row is 1, is removed; and

Recombination, where a sequence is split into two randomly in one of the $L - 1$ positions with entries to the left of the split replaced by $\circ$ in one parent gene, and similarly for the right in the other.

The transitions described above are analogous to those in (3) and have similar probabilities, taking into account that mutations and recombinations are not allowed to occur at the same site twice. It is possible that the process is absorbed into a state before the common ancestor, where every available site has encountered a recombination event, but coalescence is not possible. This is however very unlikely for large L.

The process is absorbed when the MRCAs of all the L sites have been hit.

As an illustration suppose $\theta = 3.0$, $\rho = 0.5$ are mutation and recombination rates for complete sequences, and the state is

$$\begin{array}{ccccccccccc} 1 & : & \alpha & 1 & \beta & 0 & \gamma & 1 & \delta & 1 & \kappa \\ 1 & : & \alpha & \circ & \beta & \circ & \gamma & 1 & \delta & 1 & \kappa\,. \\ 2 & : & \alpha & 0 & \beta & 1 & \gamma & \circ & \delta & 0 & \kappa \end{array}$$

Numbers before the colon denote multiplicities, $\beta = 20$, $\gamma = 20$, $\delta = 20$ and $\gamma = 20$ represent blocks of non-segregating sites containing entries 0, and $\alpha = 10$, $\kappa = 6$ are blocks of non-ancestral material with entries $\circ$. $n = 4$ with a total base length $L = 100$. Possible transitions and rates corresponding to (3) before scaling are :

coalescence for a pair (3,3)	4.0
coalescence for a pair (1,2)	8.0
mutation at site 1	3.0
recombination in sequence 1	8.3
recombination in sequence 2	8.2

recombination in sequence 3   8.3,

The rate of coalescence for (3,3) is $n(n_1 - 1) = 4$. The pair (1,2) has a coalescence rate of $2n(n_1 + 1 - 1) = 8$, since the coalescent product is the first sequence type. The last sequence cannot coalesce with the first or second because of the respective patterns at the first and last segregating sites. The number of positions between sites in which recombination affects the ancestry is $3 + \beta + \gamma + \delta$ for sequences 1,3 and $2 + \beta + \gamma + \delta$ for sequence 2, a total of $c = 128$. $a = \theta \times$ number of sites not $\circ$, weighted by multiplicity $= 3.0 \times 96 = 288$, and $b = \rho \times$ number of positions for recombination, weighted by multiplicity $= 0.5 \times 251 = 125.5$. The sum of the rates is $S = 39.8$. If X denotes the configuration above, and Y the state moved to, then $f(X, Y) = 0.0935$ for all transitions apart from recombination to the first sequence between segregating sites 2 and 3, when f is multiplied by 2, since then the right recombinant sequence is the same as sequence 2.

Program details.

Major options available in the program are shown below.

usage : recom mutation-file theta rho runs seed [options]

Options

-q distribution of recombination events, given data [outfile]

-p estimate recombination hits at each site, given data [outfile]

-t estimate time to mrca at each site, given data [outfile]

-c distribution of time to last mrca, given data [outfile]

-w estimate time to mutations at sites, given data [outfile]

-f likelihood surface [theta0 theta1 points rho0 rho1 points outfile]

-m allow multiple recombination between sites

-b [recombination bound for events to mrcas of sample]

-r [input file of non-homogeneous recombination relative rates]

The examples below show output from the program and discuss some of the options. A variation to the model allowed in the program with switch -m is to allow multiple recombination between sites. This is possible because the implementation is a discrete approximation. Variable recombination rates are allowed along the sequence with switch -r. A bound on the number of recombination events which affect the ancestry can be set. If the bound is r recombination events, then the program computes $Q_r(\mathbf{A}, \mathbf{M}, \mathbf{n})$, the joint likelihood of the sample configuration and the event that $R \leq r$ occurs. The interpretation of other estimates is then conditional on $R \leq r$. The recombination rate ρ can be set to zero, in which case the model is the infinitely-many-sites model with no recombination. Likelihood computations in this model can be done with a program **ptreesim**; theory and examples are in Griffiths et al. (1994b). Output will differ slightly because of the (long) finite locus approximation in **recom**, and there will be a combinatorial factor difference.

There is a large amount of variation in replicates of an evolutionary process with recombination, so the characteristics of the distributions need to be taken as exploratory, rather than as very precise estimates. As well as evolutionary variance the estimated likelihood produced by **recom** is a simulation estimate based on independent runs of the Markov chain described earlier. Estimates will be normally distributed, by standard theory, and have a standard deviation proportional to the inverse square root of the number of runs. Simulation variances are output by the program. The theoretical variance of the functional $\prod_0^{\tau-1} f(X(k), X(k+1))$ for a single realization can be large because many of the paths that the Markov process can visit can have a high probability but return a functional value of effectively zero. There is a system switch in **recom** to abort these paths early, and take the functional as zero for efficiency. We however don't wish to 'force' the process along particular paths, or abandon independent runs, as desirable estimation properties would be lost. TMRCA estimates and other quantities which depend on ratios of means will also be normally distributed. A large number of runs are required to achieve accuracy for the TMRCA estimates, and a substantial computer is needed for speed.

In applying the program to real data there is also a question of how close nature and the basic assumptions of the model really are, and how robust the model is. These are not easy questions to address and we do not try to do so here.

The user interface to the program is quite straightforward, but (as usual) the output should be interpreted carefully. Gnuplot commands are written to files to allow graphical output, such as a curve of the TMRCAs along the sequences.

Internally sequences are represented as bit arrays to minimize memory usage. The program is available in portable C source code on request from the 1st author. A program which simulates samples of sequences under the continuous recombination model is also available. This does not need to use a discrete locus approximation. The algorithm used for **recom** and the implementation is much more complex than the sample simulation program which is relatively easy to code.

The following three examples illustrate some aspects of ancestral inference that can be made from samples of sequences when recombination may have occurred in the samples ancestry. These inferences include estimating TMRCAs along the sequences, estimating the number of recombination events in ancestral material and maximum likelihood estimation of θ, ρ. **recom** is used as a computational tool. Emphasis is on making inferences conditional on the data observed.

Example 1.

In the sample of sequences shown in Figure 3, a simple moment estimate of θ based on 3 segregating sites in a sample of 4 is $\hat{\theta} = 3/(1 + \frac{1}{2} + \frac{1}{3}) = 1.64$.

Characteristics of ancestral distributions, conditional on the data were found by running **recom** for 500,000 replicates for each value of ρ, with a discrete approximation of 100 loci.

Estimates of the mean and standard deviation of the number of recombination events, conditional on the observed data, for $\theta = 1.64$, and illustrative values of ρ are shown in Table 1. The mean is monotonic increasing with ρ, with a standard deviation that is not extremely large. Recall that there must be at least one recombination event in this data set.

Table 1.

The expected number of recombination events, and the average number per interval length, occurring in the regions between mutations are shown in Table 2.

The average number of recombination events per length is higher in the (0.2,0.6) interval, since there must have been recombination there. $\rho = 5.0$ is a very large rate, and it is possible that the algorithm is performing poorly there.

Estimates of the ages of the mutations, with standard deviations are shown in Table 3.

Table 2.

Table 3.

A graph of the expected TMRCAs along the sequence (scaled to length 100), conditional on the data, with $\theta = 1.64$, $\rho = 2.0$ is shown in Figure 4. The expected coalescent time to the TMRCA if there was no recombination, and the data is ignored is 1.5.

Figure 4.

The graph in Figure 4 shows the characteristics of a higher TMRCA around the mutations at 0.2, 0.6, 0.8, and a higher TMRCA in the interval (0.2,0.6) where recombination must occur. The times at mutations are not as large as the pointwise prediction 2.167. Even though 500,000 runs were used to obtain the graph, there is still some simulation variance about a true curve. The TMRCA range in Figure 4 is (1.50,1.65). The small variation along the sequences is difficult to estimate accurately. Another TMRCA graph was generated by **recom** with 15 million runs and is also shown in Figure 4. The curve is much smoother than the curve with 500,000 runs. Mutations at 0.2, 0.6 produce the shape of the curve. The peak in the first curve at 0.8 is missing, but the 2nd curve more accurately reflects the fact that there are two sequences with mutations at 0.2, 0.6 and only one at 0.8, hence the mutation at 0.8 will have occurred more recently. A long period random number generator **ran2** from Press et. al. (1992) was used in **recom**.

It is possible to explicitly calculate the expected TMRCA at a single point, conditional on observing a mutation at that point. This is not the same as conditional on the *whole* data set, but it is of interest. Let $T_n, \ldots, T_2$ be the times while $n, \ldots, 2$ ancestors of a fragment of length δx of n sequences. These are distributed as exponential random variables with rates $\mu_n, \ldots, \mu_2$. Recombination does not have an effect in the computation because for small δx the probability of both recombination and mutation in an interval is $o(\delta x^2)$. (A quantity is $o(z)$ as $z \to 0$ if $o(z)/z$ converges.) Let $\pi_{\delta x}$

denote the number of mutations in the interval of width δx. As $\delta x \to 0$, conditional on there being at least one mutation in the interval there can only be one in the sense that $\lim_{\delta x \to 0} P(\pi_{\delta x} > 1 \mid \pi_{\delta x} \geq 1) = 0$. Then

$$\begin{aligned}
&E(\text{TMRCA at } x \mid \text{mutation at } x) \\
&= \lim_{\delta x \to 0} \frac{E\Big(\sum_{j=2}^{n} T_j \mid \pi_{\delta x} = 1\Big)}{P(\pi_{\delta x} = 1)} \\
&= \lim_{\delta x \to 0} \frac{EE\Big(\sum_{j=2}^{n} T_j I\{\pi_{\delta x} = 1\} \mid T_n, \dots, T_2\Big)}{P(\pi_{\delta x} = 1)} \\
&= \lim_{\delta x \to 0} \frac{E\Big(\sum_{j=2}^{n} T_j (\frac{\theta \delta x}{2} \sum_{j=2}^{n} jT_j) \exp(-\frac{\theta \delta x}{2} \sum_{j=2}^{n} jT_j)\Big)}{E(\frac{\theta \delta x}{2} \sum_{j=2}^{n} jT_j) \exp(-\frac{\theta \delta x}{2} \sum_{j=2}^{n} jT_j)\Big)} \qquad (10) \\
&= E\Big(\sum_{j=2}^{n} T_j\Big) + \frac{\operatorname{cov}\Big(\sum_{j=2}^{n} T_j, \sum_{j=2}^{n} jT_j\Big)}{E\Big(\sum_{j=2}^{n} jT_j\Big)} \\
&= 2\Big(1 - \frac{1}{n}\Big) + \frac{2\Big(\frac{1}{n} + \sum_{j=2}^{n-1} \frac{1}{j^2}\Big)}{\sum_{j=1}^{n-1} \frac{1}{j}}.
\end{aligned}$$

The same formula (10) holds for the expected TMRCA, given a recombination event at that point, but recombination events at points are not visible in the sample sequences. If $n = 4$ then (10) evaluates to 2.167. This is larger than the unconditional expected time of 1.5.

The expected TMRCA in an interval of width δx, conditional on no mutation is

$$2\Big(1 - \frac{1}{n}\Big) - 2\theta\Big(\frac{1}{n} + \sum_{j=2}^{n-1} \frac{1}{j^2}\Big)\delta x + o(\delta x^2) \text{ as } \delta x \to 0, \qquad (11)$$

but this is not such an appropriate quantity to work with because of linkage.

Arguing in a similar asymptotic way as in (10) the expected age of a mutation, given that one has occurred at a point on the sequences is

$$
\begin{aligned}
E(\text{Age of a mutation}) &= \frac{E\Big(\sum_{j=2}^{n} jT_j\big(\sum_{k=j+1}^{n} T_k + \frac{1}{2}T_j\big)\Big)}{E\big(\sum_{j=2}^{n} jT_j\big)} \\
&= \frac{2\Big(1 - \frac{1}{n}\sum_{j=1}^{n-1}\frac{1}{j} + \sum_{j=2}^{n-1}\frac{1}{j^2}\Big)}{\sum_{j=1}^{n-1}\frac{1}{j}}.
\end{aligned}
\tag{12}
$$

The ratio on the left hand side of (12) is obtained by noting that the rate of mutation while there are j ancestors is $\frac{\theta\delta x}{2}jT_j$, then if a mutation occurs the expected age of it is $\sum_{k=j+1}^{n} T_k + \frac{1}{2}T_j$. If $n = 4$ then (12) evaluates to 0.9848. This is comparable to the ages shown in Table 3. The age ranking of the mutation sites, oldest to youngest, conditional on the data is 0.6, 0.2, 0.8.

Let $N(x), x \in [0,1]$ denote the number of sequences containing a mutant type at position x. As a variation on the formulae (10), (12),

$$
E(\text{TMRCA} \mid N(x) = m) = 2\Big(1 - \frac{1}{n}\Big) + 2\binom{n-1}{m}^{-1}\sum_{j=2}^{n}\binom{n-j}{m-1}\frac{1}{j(j-1)}, \tag{13}
$$

and

$$
E(\text{Age of a mutation} \mid N(x) = m) = 2\binom{n-1}{m}^{-1}\sum_{j=2}^{n}\binom{n-j}{m-1}\frac{n-j+1}{n(j-1)}, \tag{14}
$$

for $m = 2, \ldots, n-1$.

These formulae are derived by using results about the ancestral partition in the coalescent in Kingman (1982). When there are j equivalence classes of ancestors of a sample of n, then the distribution of the class sizes is the same as the distribution of the numbers of n balls placed in j cells, uniformly at random, with no cell empty. There are $\binom{n-1}{j-1}$ ordered arrangements with equal probability. The probability that a particular class has size m is

$$
p_{n,j}(m) = \frac{\binom{n-m-1}{j-2}}{\binom{n-1}{j-1}},
$$

since fixing m there are $n-m$ balls left to arrange into $j-1$ cells. The rate of mutations while j ancestors that produce a segregating site with m mutations, given $T_n, \dots T_2$ is

$$\frac{\theta}{2} p_{n,j}(m) j T_j = \frac{\theta}{2} j(j-1) \binom{n-j}{m-1} \frac{(m-1)!(n-m-1)!}{(m-1)!} T_j.$$

It is also true that $P(N(x) = m \mid N(x) > 0) = m^{-1} / \sum_{j=1}^{n-1} j^{-1}$, $m = 1, \dots, n-1$. Details of the proof of (13) and (14) are left for the interested reader to fill in.

The distribution function of the last TMRCA of the sequence, generated by the same parameters as Figure 4, for 500,000 and 15 million runs, is shown in Figure 5. The unconditional distribution function of the TMRCA at any fixed point in [0,1], $F(t) = 1 - 1.8e^{-t} - 0.2e^{-6t} + e^{-3t}, t > 0$, is also plotted. F(t) is the distribution function of $T_2 + T_3 + T_4$. The conditional distribution, given the data, has a larger mean, and smaller variance.

Figure 5.

With just four sequences maximum likelihood estimates of θ and ρ would have a large variance, and this is reflected in the surfaces being very flat with respect to ρ. Because of the flatness and large variances between runs it was impossible to estimate θ and ρ accurately as could be done with an exact analytical expression for a likelihood surface. Even so the surfaces suggested that $\hat{\theta}$ is around 1.4-1.6 and $\hat{\rho}$ is around 2.0-3.0.

Example 2.

If a mutation occurs at a specific site then this leads one to expect a greater TMRCA there than would otherwise be the case (cf. equation (10)). Because of the presence of recombination it follows that, while nearby sites are not completely linked, they will still have a correlated genealogy. Clearly if there has not been a recombination event between the two sites their genealogies will be identical. However, even if there has been one or more such recombinations large parts of the their genealogies will be the same and the TMRCA (for example) may still be the same. Thus if a region is observed to contain many sites at which mutation has occurred, one expects to find even greater TMRCAs than would be suggested by a single such site. Conversely, the absence of mutation suggests an early TMRCA (cf. (11)). Similarly, a large number of sites with no mutations within suggest an even earlier TMRCA. To illustrate this effect consider an example data set of just two individuals of 50 bases, neither of which contains any mutations.

The command line of **recom** was:

```
recom test.dat 1.0 0.5 5000000 4867 -t mrcatimes +x 10
```

The command line asks for generating values of $\theta = 1.0$ and $\rho = 0.5$. 5 million runs are used since, for this data set, the program runs relatively quickly. The estimated TMRCAs are output to a file *mrcatimes*. The +x option is a system option requesting a memory allocation of 10 times the initial number of sequences for ancestral sequences. This file contains data to produce a graphical representation of the output, shown in Figure 6.

Figure 6

The effects noted earlier are displayed in this figure. Arguing heuristically bases in the middle of the sequence, which are surrounded by other bases which also show no mutations, have relatively lower TMRCA due to the influence of correlated genealogies. The bases towards the end of the sequence have fewer such bases in their neighbourhood, (since the sequence ends nearby), and so have less reduced TMRCA. Because of the way bases are labelled from 0 to 49 the graph should be symmetric about 24.5. The minimum and maximum of the TMRCA times are 0.507 and 0.526. The vertical scale in the graph does not represent a large range. As a comparison the expected TMRCA with two sequences ignoring the data information is 1.0, and if there was no recombination ($\rho = 0$) then the expected TMRCA, given no segregating sites would be $(1+\theta)^{-1} = 0.5$. The TMRCAs in Figure 6 lie between these values. Having recombination in the model increases the TMRCAs from 0.5. The TMRCA curve conditional on the data and that $R \leq 1$ can be computed by using **recom** with the switch +b 1. This curve is shown in PF1, together with a theoretical curve calculated from (19) below. The option +m was also used to make the rates in the discrete approximation and the continuous model agree as closely as possible. The theoretical and **recom** curves are quite close. It is interesting to note that with the conditioning $R \leq 1$ that they are both below 0.5. The derivation of the theoretical curve is as follows. Let R be the number of recombination events that affect ancestry in a sample of two sequences, S the number of segregating sites, and $W(x)$ the TMRCA at position x on the sequences. A formula is derived for $E(W(x) \mid R \leq 1, S = 0)$. Decompose

$$E(W(x)I\{R \leq 1, S = 0\}) = E(W(x)I\{R = 0, S = 0\}) + E(W(x)I\{R = 1, S = 0\}). \tag{15}$$

While there are two sequences of complete length the rates of coalescence, recombination and mutation are 1, ρ, θ so

$$\begin{aligned} P(R = 0, S = 0) &= \frac{1}{1 + \rho + \theta}, \\ E(W(x)I\{R = 0, S = 0\}) &= \int_0^\infty u \exp\big(-(\rho + \theta + 1)u\big) du = \frac{1}{(1 + \rho + \theta)^2}. \end{aligned} \tag{16}$$

If a single recombination event occurs at a point z, then there are three possible different types of ancestral graph:

(a) The left and right recombinant ancestors coalesce first after recombination;

(b) The left recombinant ancestor and the complete sequence coalesce first; and

(c) The right recombinant ancestor and the complete sequence coalesce first.

Each of (a), (b), (c) have probability 1/3 of occurring. Let τ be the time to the recombination event, η the time between recombination and the 1st coalescence, and ξ the time between the 1st and final (2nd) coalescence. Then if $x < z$,

$$W(x) = \begin{cases} \tau + \eta + \xi, & \text{in case (a);} \\ \tau + \eta, & \text{in case (b);} \\ \tau + \eta + \xi, & \text{in case (c),} \end{cases}$$

with the roles of (b) and (c) changed if $x > z$.

Just after recombination the rates of coalescence, recombination and mutation are 3, ρ, θ. After the first coalescence, considering only events in the regions without a MRCA, in case (a) the rates are 1, ρ, θ; in case (b) 1, $\rho(1 - z)$, $\theta(1 - z)$; and in case (c) 1, ρz, θz. These are found from the rates of recombination and mutation being $\theta/2$ and $\rho/2$ per unit length.

By considering cases (a), (b), (c) it follows that

$$
\begin{aligned}
&P(R=1,S=0)\\
&\quad=\frac{\rho}{1+\rho+\theta}\cdot\frac{3}{3+\rho+\theta}\\
&\quad\cdot\frac{1}{3}\Big(\frac{1}{1+\rho+\theta}+\int_0^1\frac{dz}{1+(\rho+\theta)(1-z)}+\int_0^1\frac{dz}{1+(\rho+\theta)z}\Big)dz \qquad (17)\\
&\quad=\frac{\rho}{(1+\rho+\theta)^2(3+\rho+\theta)}\Big(1+\frac{2(1+\rho+\theta)}{\rho+\theta}\cdot\log(1+\rho+\theta)\Big),
\end{aligned}
$$

and

$$
\begin{aligned}
&E(W(x)I\{R=1,S=0\})=\Big(\frac{1}{1+\rho+\theta}+\frac{1}{3+\rho+\theta}\Big)P(R=1,S=0)\\
&\quad+\frac{\rho}{(1+\rho+\theta)(3+\rho+\theta)}\\
&\quad\cdot\Big(\frac{1}{(1+\rho+\theta)^2}+\int_0^x\frac{dz}{(1+(\rho+\theta)(1-z))^2}+\int_x^1\frac{dz}{(1+(\rho+\theta)z)^2}\Big) \qquad (18)\\
&\quad=\Big(\frac{1}{(1+\rho+\theta)}+\frac{1}{(3+\rho+\theta)}\Big)P(R=1,S=0)\\
&\quad+\frac{\rho(\rho+\theta+2)(4+3\rho+3\theta+4(\rho+\theta)y^2)}{(1+\rho+\theta)^3(3+\rho+\theta)((2+\rho+\theta)^2-4(\rho+\theta)^2y^2)},
\end{aligned}
$$

where $y=|x-0.5|$.

Finally

$$
E(W(x)\mid R\le 1,S=0)=\frac{E(W(x)I\{R=0,S=0\})+E(W(x)I\{R=1,S=0\})}{P(R=0,S=0)+P(R=1,S=0)}
$$

(19).

It is also straightforward to find $P(R\le 1\mid S=0)$ from (16) and (17) which is 0.9442 with $\theta=1.0$, $\rho=0.5$. Because of its functional form $E(W(x)\mid R\le 1,S=0)$ is symmetric about $x=0.5$, with a maximal value at endpoints $x=0,1$. A plot is shown in Figure 6, with the horizontal scaled to be symmetric about 24.5.

Figure 7

The TMRCA curve was computed using **recom** for another data set of two sequences of length fifty bases with the 1st sequence having mutations at sites 10, 30, and the second with a mutation at site 20. An estimate based on the number of segregating sites was $\hat{\theta}_S = 3.0$. **recom** was run with $\theta = 3.0, \rho = 0.5$ for illustration. The effect of mutations lengthening the TMRCA is clearly shown in Figure 7, though the range of the curve is small. The times are of course more than those for two sequences with no mutations shown in Figure 6. If there were no recombination in the model then the expected TMRCA, given three segregating sites is $4/(1+\theta) = 1$. The expected TMRCA at a point, given a mutation there is 1.5, calculated from (13). The whole TMRCA curve in Figure 7 is above 1, and below 1.5 because recombination increases the TMRCAs at mutation sites, but linkage is still tight enough to also increase the TMRCAs at sites nearby. **recom** was rerun with $\theta = 3.0$, $\rho = 1.0$. The TMRCA curve is also shown in Figure 7. The influence of mutations is similar to when $\rho = 0.5$, but because there is less linkage the range of the curve is greater, with times at mutation sites and decreased TMRCAs in long regions with no mutations (cf. positions 35-50).

Example 3.

A sample of 50 sequences was simulated with $\theta = 2.0$, $\rho = 0.5$ using the urn scheme in Griffiths et al. (1995).

Sample sequences are shown in Figure 8, with multiplicities on the left. The input file to **recom**, using a discrete approximation of 100 bases is shown in Table 4. The symbolism @n denotes a block of n zeros, and only the nine segregating sites are shown in full.

The command line of **recom** was

```
recom test.dat 2.0 0.5 2000000 2938984 +b -f 1.0 3.0 41 0.05 2.0 40 test.f -c test.c -q
test.q -t test.t -w test.w -p test.p
```

Program options are shown earlier in this paper. The command line requests **recom** to use the data file test.dat with generating values for the Markov process of $\theta = 2.0$, $\rho = 0.5$, use 2,000,000 runs with random number seed 2938984. Option +b is a system option to return 0 for low probability paths discussed later. Output from the various options is sent to files named *test.**. The surface option -f requests a likelihood surface for ranges θ in [1.0,3.0] and ρ in [0.05,2.0]. There are 41 increment values for θ and 40 for ρ, producing increments of 0.05.

Figure 8.

Table 4.

Recombination must have occurred between segregating sites 4 and 9 at 0.679 and 0.996, but apart from this pair all other pairs are consistent with the infinitely-many-sites-model. Actually in the simulated sample there were three recombination events in material ancestral to the sample, at 0.084, 0.113, 0.976. There does not seem to be evidence of the first two recombinations in the data. If the 4th mutant site is ignored, and it is assumed that recombination has not affected the mutation pattern at other sites, then a genealogical tree in the sense of Griffiths et al. (1994) can be constructed and is shown in Figure 9. In this tree vertices represent mutant sites numbered 1-9 corresponding to data in Table 4. The 4th site is superimposed to show its inconsistency. A site is in an ancestral path from a sample sequence to the root if it appears as mutant on the corresponding sequence in Figure 8. Multiplicities are shown at the tips of the tree. The arrangement of sites within the sets $\{2,5,7\}$ and $\{3,8\}$ in the tree is not unique. It appears that with just one recombination event affecting the site configuration the recombination graph must take the form in Figure 9. The seven sequences with mutation 6 could join on either the left or right above 4 or 9.

Figure 9.

Figure 10.

Characteristics of the likelihood, and ancestral distributions, conditional on the data were explored using **recom** with 2,000,000 replicates, with generating parameters $\theta_0 = 2.0$, $\rho_0 = 0.5$. Likelihood calculations for a sample of this size require a substantial computer and time committment. On a Dec alpha AXP12 the run time was 78.5 hours. **recom** has a switch to abort paths which have a small functional value and return zero for the estimated likelihood for such a run. If $\prod_0^\eta f(X(k), X(k+1) < \epsilon$ for $\eta < \tau$ in (5) and $(\mathbf{A}_1, \mathbf{M}_1, \mathbf{n}_1)$ is the configuration at $\eta + 1$, then the expected single run estimate from this point,

$$\hat{Q}(\mathbf{A}, \mathbf{M}, \mathbf{n}) = \prod_0^{\eta} f(X(k), X(k+1)) Q(\mathbf{A}_1, \mathbf{M}_1, \mathbf{n}_1) < \epsilon.$$

Choosing ϵ appropriately allows paths which have a small functional to be determined early, improving the algorithms speed. With **recom** on this data set 1,781,273 runs returned zero. This seems absurd at first sight, but the reason is that the process with transitions (3) is contrived, and may have quite high probability paths which return a low functional value.

The likelihood surface for (θ, ρ) is shown in Figure 11. The maximum likelihood estimates from the surface are $\hat{\theta} = 1.75$, $\hat{\rho} = 0.4$ with a likelihood of 7.45×10^{-12}. It is difficult to give accurate variance estimates for $\hat{\theta}$ and $\hat{\rho}$. A second replicate computation gave a very similar likelihood surface, with estimates $\hat{\theta} = 1.80$ and $\hat{\rho} = 0.3$. There may be a large evolutionary variance associated with them, but in this example the estimates are quite accurate. The inverse of the information matrix, calculated approximately by finite differences from the maximum point and eight points symmetrically about this point with a step size of 0.05 was

$$\begin{pmatrix} 0.115 & 0.0087 \\ 0.0087 & 0.1096 \end{pmatrix}.$$

If this was a good estimate of variance, then $\mathrm{sd}(\hat{\theta}) = 0.3398$ and $\mathrm{sd}(\hat{\rho}) = 0.3310$ with a correlation between the estimates of 0.0775. The estimates are, however, unlike usual repeated sampling estimates. Even if the total population frequencies were known, then it is not clear if ρ and θ would be determined with probability 1. If there is no recombination then θ can be determined with probability 1 from the entire population, so we suspect that this holds in the more general model with recombination. It is difficult to tell if estimates from the information matrix are reasonable, but they are not outrageous.

To check the effect of the generating values of the process the computation was repeated with $\theta_0 = 2.0$, $\rho_0 = 1.0$. The likelihood surface was similar to that in Figure 11, and the estimated parameters were $\hat{\theta} = 1.90$, $\hat{\rho} = 0.45$ with likelihood 5.81×10^{-12}.

The smaller likelihood suggests that the former estimates are better. A mutation rate estimate just using the fact that there are $S = 9$ segregating sites in a sample of 50 sequences is $\hat{\theta}_S = 9/\sum_{i=1}^{49} i^{-1} = 2.0$.

The distribution of the number of recombination events, affecting ancestry, and in ancestral material, are shown in Table 6. There is little difference between the distributions. The mean and standard deviations are $\mu = 2.228$, $\sigma = 0.824$, the same to three decimal places for both distributions. This is consistent with $\hat{\rho} = 0.4$ since the expected number of recombinations in ancestral material (not conditional on the sample configuration) is estimated by $\hat{\rho}\sum_{i=1}^{49} i^{-1} = 1.79$.

A histogram along the sequences of the expected number of recombination hits, given the data, is shown in Figure 12. This shows a higher recombination rate per base around the right end of the sequences. The TMRCAs along the sequences are shown in Figure 13 for two replicated computations, each with 2,000,000 runs. The time is longer at the right end which seems consistent with recombination being in this region. The minimum occurring in one replicate is unreliable. Apart from the minimum the two replicates have a similar shape, indicating some reliability. Unfortunately the TMRCA mean curve given the data cannot be estimated very accurately because the order of magnitude of the range of the curve tends to be of the same order of magnitude as the simulation standard deviation. In this example the simulation standard deviations along the sequences were in the range (0.25,0.45). The standard deviation of the TMRCA distribution is a different concept, and was estimated to be about 0.7 at positions along the sequence. The mean and standard deviation of the last TMRCA are 2.56 and 0.74. At a detailed level the expected TMRCA at mutant sites, the expected age of mutant sites given the data, and a comparison of these times just conditional on the number of mutant sequences are shown in Table 5. Of course the estimates from **recom** should be best, but the comparison is interesting. Mutant sites which were present in higher numbers of sequences have larger TMRCAs and ages in both computations, though (as expected) the variation is more when just conditioning on single sites. As the amount of recombination increases sites behave more independently and (13), (14) will be more accurate. To illustrate this results from a computation with $\rho = 1.0$ are shown in the Table. The TMRCAs are closer to the pointwise values than with $\rho = 0.5$, but this doesn't hold for the ages. Apart from the comparison with the pointwise formula, it is of interest to see where mutations occur in the ancestry relative to the TMRCA at mutant sites.

Overall the expected TMRCA and age conditional on mutation at a site, computed from (10) and (12) are 2.248, 0.685. The larger TMRCA times are consistent with high numbers of mutant sites at positions 0.143, 0.767 0.804, 0.996 in both graphs.

Table 5.

Normally one would rerun **recom** on the data set with new generating parameters equal to the parameter values estimated here. to find the characteristics of ancestral distributions, however in this example the estimated parameters were so close to the generating parameters that the program was not rerun. Mutations are shown in relative age order in Figure 8. The fact that recombination must have occurred in the interval (0.679,0.996) makes the distribution of ages at sites within the set $\{2,5,7\}$ asymmetric because of their positions. If the recombination split in Figure 10 was at z, then another consistent recombination graph could have any of the mutations from $\{2,5,7\}$ at positions less than z appearing in the right recombinant ancestor before mutation 9.

Figure 11.

Table 6.

Figure 12.

Figure 13.

Figure 14.

Another data set shown in Figure 15 was simulated with $\theta = 2.0$, $\rho = 0.0$.

Figure 15.

The likelihood surface generated using **recom** with parameters $\theta_0 = 2.0$, $\rho_0 = 0.5$ is shown in Figure 14. The maximum likelihood estimates were $\hat{\theta} = 2.6$, $\hat{\rho} = 0.0$

consistent with $\rho = 0.0$. A mutation rate estimate just using the fact that there are $S = 11$ segregating sites in a sample of 50 sequences is $\hat{\theta}_S = 2.5$, quite close to $\hat{\theta} = 2.6$.

Acknowledgement.

This research was partially funded by an Australian Research Council Grant.

References.

Ethier, S.N. and Griffiths, R.C. 1987. The infinitely many sites model as a measure valued diffusion. *Ann. Prob.* 15, 515–545.

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

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

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. *Theoret. Popn. Biol.* 3, 87–112.

Golding, G. B. 1984. The sampling distribution of linkage disequilibrium. *Genetics* 108, 257–274.

Griffiths, R.C. 1981. Neutral two-locus multiple allele models with recombination. *Theoret. Popn. Biol.* 19, 169–186.

Griffiths, R.C. 1989. Genealogical–tree probabilities in the infinitely–many–site model. *J. Math. Biol.* 27, 667–680.

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

Griffiths R. C. and Marjoram, P. 1995. An ancestral recombination graph. *In* Donnelly, P. and Tavaré, S. eds., *IMA volume on Mathematical Population genetics.* (To appear.)

Griffiths R. C. and Tavaré, S. 1994a. Simulating probability distributions in the coalescent. *Theoret. Popn. Biol.* 46, 131–159.

Griffiths R. C. and Tavaré, S. 1994b. Ancestral inference in population genetics. *Statistical Science* 9, 307–319.

Griffiths R. C. and Tavaré, S. 1994c. Sampling theory for neutral alleles in a varying environment. *Proc. R. Soc. Lond. B* 344, 403–410.

Griffiths R. C. and Tavaré, S. 1995a. Unrooted genealogical tree probabilities in the infinitely- many-sites model. *Mathematical Biosciences* 127, 77–98.

Griffiths R. C. and Tavaré, S. 1995b. Markov chain Monte carlo in population genetics. *Math. Comput. Modelling* (To appear.)

Griffiths R. C. and Tavaré, S. 1995c. Computational methods for the coalescent. *In* Donnelly, P. and Tavaré, S. eds., *IMA volume on Mathematical Population genetics.* (To appear.)

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

Hudson, R.R., 1991. Gene genealogies and the coalescent process, 7, 1–44, *In* Futuyma, D. and Antonovics, J., eds., *Oxford Surveys in Evolutionary Biology*

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.

Hudson, R.R. and Kaplan, N.L. 1988. The coalescent process in models with selection and recombination, *Genetics* 120, 831–840.

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

Kingman, J.F.C. The coalescent. *Stoch. Proc. Applns.* 13, 235–248.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. 1992. *Numerical Recipes in C, 2nd ed.* Cambridge University Press, Cambridge.

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

Table 1

Number of recombination events.

rho	mean	sd
0.5	2.1043	1.1634
1.0	2.9487	1.4674
1.5	3.8142	1.9719
2.0	4.5078	1.9845
2.5	5.0638	1.8115
5.0	6.7732	1.6983

Table 2

Recombination events in regions;
number, and average per base.

	Sequence Region			
rho	0.0-0.2	0.2-0.6	0.6-0.8	0.8-1.0
0.5	0.3004 0.0150	1.3305 0.0333	0.2983 0.0149	0.1751 0.0092
1.0	0.3422 0.0171	1.8151 0.0454	0.4597 0.0230	0.3317 0.0174
1.5	0.7612 0.0381	2.0089 0.0502	0.5529 0.0276	0.4913 0.0259
2.0	0.7033 0.0351	2.2404 0.0560	0.8544 0.0427	0.7097 0.0374
2.5	0.8492 0.0425	2.5218 0.0630	0.8978 0.0449	0.7950 0.0418
5.0	1.9865 0.0993	2.6752 0.0669	0.6228 0.0311	1.4887 0.0784

Table 3

Mean and sd of ages of mutations.

	Mutation position		
rho	0.2	0.6	0.8
0.5	0.6879 (0.3150)	0.8833 (0.3762)	0.2744 (0.1744)
1.0	0.7764 (0.3356)	0.8582 (0.3469)	0.2908 (0.1697)
1.5	0.8034 (0.3207)	0.9313 (0.3616)	0.2433 (0.1503)
2.0	0.7474 (0.2908)	1.0114 (0.3540)	0.2392 (0.1480)
2.5	0.9501 (0.3354)	1.0285 (0.3558)	0.2916 (0.1471)
5.0	0.4334 (0.1827)	0.7198 (0.2392)	0.2291 (0.1210)

Table 4

Sample sequences with multiplicities.

7	:	@4	0	@9	0	@26	0	@25	0	@8	0	1	@2	0	@10	0	@7	0
32	:	@4	0	@9	1	@26	0	@25	0	@8	1	0	@2	1	@10	0	@7	1
3	:	@4	0	@9	0	@26	0	@25	1	@8	0	0	@2	0	@10	0	@7	0
2	:	@4	0	@9	0	@26	0	@25	1	@8	0	0	@2	0	@10	0	@7	1
2	:	@4	0	@9	1	@26	1	@25	0	@8	1	0	@2	1	@10	1	@7	1
4	:	@4	1	@9	1	@26	0	@25	0	@8	1	0	@2	1	@10	0	@7	1

Table 5

TMRCA and ages of sites.

mutant site	1	2	3	4	5	6	7	8	9
position	0.04	0.14	0.41	0.67	0.76	0.77	0.80	0.91	0.99
mutant sequences	4	38	2	5	38	7	38	2	40
TMRCA, **recom**, $\rho = 0.5$	2.41	2.43	2.36	2.42	2.43	2.43	2.43	2.43	2.44
TMRCA, **recom**, $\rho = 1.0$	2.10	2.07	2.09	2.15	2.19	2.19	2.21	2.21	2.32
TMRCA, given m	2.10	2.80	2.04	2.13	2.80	2.19	2.80	2.04	2.83
age, **recom**, $\rho = 0.5$	0.19	1.73	0.24	0.72	1.47	0.84	1.32	0.26	1.38
age, **recom**, $\rho = 1.0$	1.22	0.12	0.13	0.98	1.05	0.72	1.23	1.45	1.38
age, given m	0.42	1.72	0.25	0.49	1.72	0.62	1.72	0.25	1.76

Table 6

Recombination events distribution, given data,

R, affecting ancestry; R_a in ancestral material.

r	$P(R=r)$	$P(R_a=r)$
1	2.2541×10^{-1}	2.2541×10^{-1}
2	3.4269×10^{-1}	3.4269×10^{-1}
3	4.1494×10^{-1}	4.1494×10^{-1}
4	1.3502×10^{-2}	1.3499×10^{-2}
5	2.6396×10^{-3}	2.6395×10^{-3}
6	7.9004×10^{-4}	7.9005×10^{-4}
7	2.5431×10^{-5}	2.5396×10^{-5}
8	6.1325×10^{-6}	6.1325×10^{-6}