Approximating the coalescent with recombination

Gilean A. T. McVean* and Niall J. Cardin

Department of Statistics, 1 South Parks Road, University of Oxford, Oxford, OX1 3TG, UK

The coalescent with recombination describes the distribution of genealogical histories and resulting patterns of genetic variation in samples of DNA sequences from natural populations. However, using the model as the basis for inference is currently severely restricted by the computational challenge of estimating the likelihood. We discuss why the coalescent with recombination is so challenging to work with and explore whether simpler models, under which inference is more tractable, may prove useful for genealogy-based inference. We introduce a simplification of the coalescent process in which coalescence between lineages with no overlapping ancestral material is banned. The resulting process has a simple Markovian structure when generating genealogies sequentially along a sequence, yet has very similar properties to the full model, both in terms of describing patterns of genetic variation and as the basis for statistical inference.

Keywords: coalescent; recombination; linkage disequilibrium; genealogy

1. INTRODUCTION

The coalescent with recombination describes the distribution of genealogies underlying samples of chromosomes from unrelated individuals in idealized natural populations (Hudson 1983; Griffiths & Marjoram 1996). Starting from the present and looking back in time, the ancestral lineages relating to the sampled chromosomes are traced until coalescence (where two ancestral lineages meet in a common ancestor) or recombination (where an ancestral lineage splits in two). The resulting ancestral recombination graph (ARG) has embedded within it the marginal genealogy (or phylogenetic tree describing the ancestry of the chromosomes) at any position along the sequence and, by mapping mutations on to the graph, describes patterns of genetic variation in the sampled chromosomes. Under models with constant population size and random mating, two parameters determine the distribution of variation: the population mutation rate \( \theta = 4N_e \mu \) (where \( N_e \) is the effective population size and \( \mu \) is the per generation mutation rate); and the population recombination rate \( \rho = 4N_e r \) (where \( r \) is the per generation recombination rate).

Stochastic simulation under the model (backwards in time starting from the present) is computationally straightforward because at any point in time the rates of coalescence and recombination are simple functions of the ancestral lineages present (i.e. it has a Markovian structure; Hudson 1983). In contrast, the alternative approach of simulating genealogies while moving along a sequence (Wiufl & Hein 1999) has a complex non-Markovian structure in that the distribution of the next genealogy depends not just on the current genealogy, but also all previous ones. Both approaches, however, can make use of the separation of the genealogical and mutational processes under neutrality (Hudson 1990). Consequently, the ARG may be generated first with mutations subsequently added to the marginal genealogies as a Poisson process.

Efficient inference under the coalescent with recombination is notoriously difficult (Stumpf & McVean 2003). For example, moment estimation of \( \rho \) can be achieved by comparing the sample variance in pairwise differences to the expectation under neutrality (using a point estimate of \( \theta \); Hudson 1987; Wakeley 1997), but the estimator uses only a fraction of the available information about recombination and is both biased and has high variance (Wall 2000).

In contrast, likelihood-based inference (which uses all possible information) is currently restricted because there exists no analytic or numerical expression for the likelihood function and the construction of efficient Monte Carlo methods for estimating the likelihood is technically challenging. Naively, the likelihood could be estimated by simulating ARGs from the coalescent distribution given \( \rho \), adding mutations to the ARGs from the distribution given by \( \theta \) and looking to see if the simulated data matched the data observed. By repeating many times under different values of \( \theta \) and \( \rho \), maximum likelihood estimates of the statistic could be obtained. In practice, the naive approach is infeasible because the vast majority of ARGs contribute nothing to the likelihood. Consequently, sophisticated Monte Carlo methods such as importance sampling (IS; Fearnhead & Donnelly 2001) and Markov Chain Monte Carlo (MCMC; Kuhner et al. 2000; Nielsen 2000) must be used (reviewed in Stumpf & McVean 2003), which create bias towards the simulation of ARGs that make significant contributions to the likelihood.

To date, while Monte Carlo methods can be used to calculate likelihoods for very simple datasets, they are still impractical for most datasets currently being collected. Instead, three alternative approaches to coalescent-based inference have been explored.

* Author for correspondence (mcvean@stats.ox.ac.uk).

One contribution of 16 to a Theme Issue ‘Population genetics, quantitative genetics and animal improvement: papers in honour of William (Bill) Hill’.
First, it is possible to calculate the likelihood of a summary of the data, rather than the data itself. For example, Wall (2000) suggested estimating \( \rho \) by calculating the likelihood of observing the number of haplotypes \( H \) and the minimum number of recombination events \( R_m \) as estimated by the method of Hudson & Kaplan (1985). Importantly, this likelihood may be calculated by naive simulation, potentially aided by regression techniques (Beaumont et al. 2002).

The second approach is to divide the complete data into smaller subsets (pairs of segregating sites; Hudson 2001; McVean et al. 2002 or non-overlapping windows Fearhead & Donnelly 2002), the likelihood of which can be calculated using IS or even naive methods. Combining likelihood calculations across subsets can give accurate estimates (Wall 2000), and can be used to estimate variation in the recombination rate (McVean et al. 2004), but the resulting likelihoods do not have standard properties (e.g. be used to calculate support intervals).

The third approach is to simplify or approximate the coalescent model itself. Building on research into optimal IS proposal distributions, Stephens & Li (2003) proposed a new statistical model for genetic data with recombination that generates patterns of genetic variation similar to the coalescent, but uses an approximation to the genealogical process. Importantly, the approximation means that likelihoods are easy to compute (referred to as product of approximate conditional, or PAC likelihoods), hence the approach generates a true likelihood.

The disadvantage of approximating the coalescent model is that the biological validity of the approximation may be poor. In the PAC approach, chromosomes are no longer exchangeable (i.e. the likelihood depends on the order in which chromosomes are analysed), and the estimated recombination parameter can only be related to that of the coalescent through an empirical bias correction. Even more importantly, the coalescent approximation means that conditional on the set of marginal genealogies embedded within the ARG, and every set of marginal genealogies corresponds to an infinite set of ARGs. In effect, the coalescent with recombination provides an indirect model for the set of marginal genealogies underlying a sample of chromosomes. The problem is that there is no direct way of calculating the coalescent likelihood from a set of marginal genealogies (i.e. integrating over the possible ARGs that could have given the set of marginal genealogies). As a result, inference methods are based on augmenting with the more complex and potentially redundant ARGs (redundant in the sense that two ARGs with identical marginal genealogies may have very different likelihoods).

The coalescent is, of course, just one possible model for the set of marginal genealogies (albeit based on an appreciation of reproduction in natural populations). In the following section we develop the idea that alternative models might be worth considering if they both prove easier for inference and have properties similar to the coalescent. In particular, we introduce a simplification of the coalescent that generates sets of marginal genealogies with a similar correlation structure to the coalescent but has both a much-reduced state-space and a direct way of calculating the coalescent likelihood for a set of marginal genealogies.

**2. WHAT IS DIFFICULT ABOUT THE COALESCENT WITH RECOMBINATION?**

Before considering novel approximations to the coalescent with recombination it is necessary to appreciate why the basic model is so difficult to estimate likelihoods under. There are three important and inter-related issues.

(i) The state-space of ARGs is huge.

(ii) The data are generally not very informative about the actual ARG.

(iii) Likelihood estimation is a missing-data problem with highly redundant augmentation.

The state-space of ARGs is huge because the number of recombination events is unbounded and when the rate of recombination is high, most coalescent events will be between lineages that have no overlapping ancestral material and which will rapidly get broken up by further recombination. However, the size of the state space in itself does not make the problem of likelihood calculation difficult. More importantly, unless the ratio of mutation to recombination is very high, there is generally little information about the true genealogy underlying the sample at any point. This means that the likelihood function will not be dominated by any single ARG, so that efficient sampling of diverse contributing ARGs is required.

The last issue needs more explanation. Inference under the coalescent with recombination is an example of a missing data problem: if the true ARG were known, calculating the data likelihood would be trivial, hence summing over possible ARGs is achieved by successively augmenting the data with potential ones. However, the data likelihood is actually only dependent on the set of marginal genealogies embedded within the ARG, and every set of marginal genealogies corresponds to an infinite set of ARGs. In effect, the coalescent with recombination provides an indirect model for the set of marginal genealogies underlying a sample of chromosomes. The problem is that there is no direct way of calculating the coalescent likelihood from a set of marginal genealogies (i.e. integrating over the possible ARGs that could have given the set of marginal genealogies). As a result, inference methods are based on augmenting with the more complex and potentially redundant ARGs (redundant in the sense that two ARGs with identical marginal genealogies may have very different likelihoods).

The coalescent is, of course, just one possible model for the set of marginal genealogies (albeit based on an appreciation of reproduction in natural populations). In the following section we develop the idea that alternative models might be worth considering if they both prove easier for inference and have properties similar to the coalescent. In particular, we introduce a simplification of the coalescent that generates sets of marginal genealogies with a similar correlation structure to the coalescent but has both a much-reduced state-space and a direct way of calculating the coalescent likelihood for a set of marginal genealogies.
3. THE SEQUENTIALLY MARKOV COALESCENT

We first describe a generalization of the standard coalescent process for a constant population size. At any point in time the state of the coalescent process is described by the set of \( k \) ancestral lineages, the \( i \)th of which contains ancestral material at a set of \( m \) ordered non-overlapping intervals on the unit interval (treating sequences as continuous)

\[
\mathbf{x}_i = \{(x_{i1}, y_{i1}), (x_{i2}, y_{i2}), \ldots, (x_{im}, y_{im})\},
\]

where \( x_{ij} \) and \( y_{ij} \) represent the lower and upper limits respectively of an interval of ancestral material. The instantaneous rate of coalescence is the sum of the rates for all pairs of lineages that can potentially coalesce (note each pair coalesces independently)

\[
\lambda_C = \sum_{i \neq j} I_{ij},
\]

where \( I_{ij} = 1 \) under the standard coalescent for all \( i \) and \( j \) (\( i \neq j \)). The instantaneous rate of recombination is

\[
\lambda_R = \rho \sum_i (y_{im} - x_{i1}).
\]

The time until the next event is distributed as a random exponential variable with rate \( \lambda_C + \lambda_R \), with the type of event being drawn in proportion to its contribution to the summed rates. If a coalescent event is chosen, the resulting lineage acquires the union of the intervals of ancestral material. After coalescence, if any interval is represented by just one ancestral lineage the most recent common ancestor (MRCA) of that interval has been reached and the interval is removed. Recombination events result in a splitting of ancestral material, where the point of splitting is chosen uniformly from the interval \((x_{i1}, y_{im})\) for a constant recombination rate. The process for a sample of size \( n \) is initialized by setting \( k = n \) and \( m_i = 1 \), \( x_{i1} = 0 \), \( y_{11} = 1 \) for all \( i \) and terminated when every point along the sequence has found an MRCA.

The SMC process requires a simple modification to the process. If \( X_i (= \cup x_i) \) is the set of all loci at which chromosome \( i \) has ancestral material

\[
I_{ij} = \begin{cases} 
1 & \text{if } X_i \cap X_j \neq \emptyset \text{ and } i \neq j \\
0 & \text{otherwise}
\end{cases}
\]

In other words, if two ancestral lineages have no interval in common where they share ancestral material they are not allowed to coalesce.

By restricting coalescent events in this way the resulting process has three important differences from the standard coalescent. First, the state-space of ARGs is much reduced (though the state-space of the set of marginal genealogies is unaltered). Second, the SMC model will tend to have many fewer recombination events in its history (figure 1). Third, the resulting process has a Markovian structure in the sequential generation of genealogies along a chromosome.

The Markovian structure along a chromosome is best described in terms of the sequential algorithm for simulating a set of marginal genealogies on the unit interval for \( n \) sequences under the SMC model.
The backwards-in-time and sequential formulations lead to the same distribution on marginal genealogies because restricting coalescence to lineages that have overlapping ancestral material is equivalent to only allowing the floating lineage to coalesce to the remaining part of the previous genealogy. There is a slight difference between the processes in that recombination events in non-ancestral material can occur in the backwards-in-time process (when a site has reached its MRCA but the neighboring sites have not), which are equivalent to recombination events on the stalk of the marginal genealogy. However, such events have no influence on the distribution of marginal genealogies.

4. PROPERTIES OF THE SEQUENTIALLY MARKOV MODEL

The point of introducing the modified coalescent process is that while it has a similar structure to the standard process (and is therefore a reasonable model for patterns of genetic variation), the reduced state-space and sequentially Markovian structure may make inference considerably easier compared with the full coalescent. For example, estimation of likelihoods using the backwards-in-time IS scheme of Fearnhead & Donnelly (2001) may be more efficient because fewer events need to be considered at each step and fewer recombination events will occur in the history of the sequences. In addition, the Markovian sequential nature allows for efficient calculation of the coalescent likelihood for a set of marginal genealogies, which enables the use of MCMC methods in which marginal likelihood for a set of marginal genealogies, which overlapping ancestral material is equivalent to only

of the non-Markovian structure of marginal genealogies under the standard coalescent. Second, we consider expected patterns of linkage disequilibrium under the two models. Finally, we consider inference under both models, using a simple example with two sequences.

(a) Non-Markovian properties of the standard coalescent process

To quantify the importance of coalescent events between ancestral lineages that share no overlapping material (the events that generate a non-Markovian sequential algorithm) we consider two related probabilities. \( Q(x, \rho) \) is the probability (for a given value of \( \rho \)) that two unrelated sequences that share a single MRCA at points 0 and 1 have a different MRCA at an intervening point \( x \). \( Q^* (\rho) \) is the probability that two sequences that share a single MRCA at points 0 and 1 do not share one in at least one intervening interval. The probability, \( P(\rho) \), that a pair of sequences share a common ancestor at points 0 and 1 is (proof not shown)

\[
P(\rho) = \frac{18 + \rho}{18 + 13\rho + \rho^2}.
\]  

Table 1. \( P(\rho) \), \( Q(0.5, \rho) \) and \( Q^* (\rho) \) as a function of recombination rate.

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( P(\rho) )</th>
<th>( Q(0.5, \rho) )</th>
<th>( Q^* (\rho) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.937</td>
<td>0.0002</td>
<td>0.0003</td>
</tr>
<tr>
<td>1</td>
<td>0.594</td>
<td>0.012</td>
<td>0.023</td>
</tr>
<tr>
<td>2</td>
<td>0.417</td>
<td>0.029</td>
<td>0.054</td>
</tr>
<tr>
<td>5</td>
<td>0.213</td>
<td>0.066</td>
<td>0.106</td>
</tr>
<tr>
<td>10</td>
<td>0.113</td>
<td>0.087</td>
<td>0.123</td>
</tr>
<tr>
<td>20</td>
<td>0.056</td>
<td>0.087</td>
<td>0.112</td>
</tr>
<tr>
<td>50</td>
<td>0.021</td>
<td>0.060</td>
<td>0.065</td>
</tr>
<tr>
<td>100</td>
<td>0.010</td>
<td>0.037</td>
<td>0.045</td>
</tr>
</tbody>
</table>

Whereas \( Q(x, \rho) \) can be obtained numerically (equations not shown), \( Q^* (\rho) \) can only be obtained by Monte Carlo simulation (we used 10^6 samples). Under the sequentially Markov process, \( Q(x, \rho) = Q^*(\rho) = 0 \), so the magnitude of these quantities determine the error in the sequentially Markov process.

Table 1 shows how these probabilities vary as a function of the recombination rate. While the probability that points 0 and 1 share an MRCA decreases monotonically, the conditional probabilities of having a different intervening MRCA peak at intermediate values of \( \rho \). The non-Markovian behaviour is strongest for \( \rho = 10 \), where there is approximately a 12% chance of having a different intervening MRCA, conditional on the two end points sharing an MRCA. Note, however, that there is only an 11% chance of the two points sharing an MRCA at all, so the absolute probability of the non-Markovian event is less than 2%.

(b) Linkage disequilibrium

We can also compare models in terms of the patterns of linkage disequilibrium (LD) generated. Specifically, we can compare the distribution of two-locus summaries of allelic association under the coalescent and sequentially Markov processes.

For two bi-allelic loci, the \( r^2 \) measure of association (Hill & Robertson 1968) is the square of the correlation coefficient between allelic states. Although no simple expression for moments of the distribution of \( r^2 \) can be derived, a related quantity is given by

\[
\sigma_d^2 = \frac{E[D]^2}{E[D^2]}. 
\]  

where \( D \) is the standard disequilibrium coefficient and \( p_i \) and \( q_i \) are the allele frequencies at locus \( i \) (Ohta & Kimura 1971). By conditioning on segregation at both loci and letting the mutation rate tend to zero, McVean (2002) showed how equation (4.2) could be expressed in terms of the covariance in coalescence times at the two loci for different configurations of chromosomes

\[
\sigma_d^2 = \frac{C_{ij,kl} - 2C_{ijk} + C_{ijkl}}{E[t]^2 + C_{ijkl}}. 
\]  

where \( C_{ijkl} \) is the covariance between the coalescence time for two sequences \( i \) and \( j \) sampled at the first locus, and two (possibly identical) sequences \( k \) and \( l \) sampled at the second locus. Expressions for the covariance terms under the coalescent can be derived by solving a system of linear equations. For the coalescent these
two-locus statistics of LD. However, we have also shown that non-Markovian behaviour can have some effect on the distribution of genealogies, so it is not obvious that inferences made under the sequentially Markov process will be similar to those made under the full coalescent.

While a full exploration of inference under the sequentially Markov model is beyond the scope of this article, we can consider a simple example for two sequences. For simplicity we will assume a model in which recombination occurs at a finite number of equally spaced points. In our toy example we use 20 sites with single mutations in intervals 2, 3, 4, 5, 16, 17 and 18.

For inference we can calculate a likelihood surface for $\theta$ (per interval) and $\rho$ (for the region) by naive Monte Carlo simulation. Figure 4 shows the joint log likelihood surface under the two models. The maximum likelihood estimates for $\theta$ and $\rho$ are very similar under both models ($\hat{\theta} = 0.325$ for both; $\hat{\rho} = 35$ for the coalescent and $\hat{\rho} = 30$ for the sequentially Markov process).

For the estimated parameter values, we can also calculate the posterior distributions of the number of recombination events between each pair of sites and the marginal TMRCAs (figure 5), using the MCMC method of Griffiths (1999). The distribution of the expected TM RCA along with sequences is almost identical under the standard coalescent and sequentially Markov processes. The distributions of the number of recombination events are also similar in shape, peaking in the regions with mutations. However, the expected number of recombination events under the coalescent is approximately twice as high as under the sequentially Markov process, because coalescence between lineages that share no overlapping ancestral material provide further opportunity for recombination. Note, however, that this does not lead to a large shift in the estimated recombination rate.

5. DISCUSSION

The analysis of patterns of linkage disequilibrium using Wright–Fisher models has long history in populations genetics (Hill & Robertson 1968; Ohta & Kimura 1971; Hill 1975; Weir & Hill 1986). Of particular interest has been the extent to which variation in the extent of linkage disequilibrium along a chromosome reflects underlying variation in the recombination rate (Weir & Hill 1986; Crawford et al. 2004; Fearnhead et al. 2004; McVean et al. 2004). However, recent attempts to use the coalescent with recombination as
a model for genetic variation have used approximations either to the likelihood function or the model itself, because estimating the likelihood function is computationally intractable for all but the smallest datasets.

The problem with the current methodologies is that while they may be useful for estimating the recombination rate, they do not explicitly model the ancestral history of set of sequences and hence cannot be used to estimate genealogies along a chromosome, or the ages of mutations and MRCAs. Knowing such quantities is of interest to many areas of population genetics (such as demographic inference and association mapping), which has generated our interest in developing genealogical models related to the coalescent under which inference may be easier.

There are many potential approaches to approximating the coalescent process so as to result in models that are easier for inference, but retain a notion of genealogy. The sequentially Markov process described might provide one such model, although further research is required into whether inference under the model is sufficiently more efficient than under the standard coalescent. More generally, we have shown that the exclusion of certain types of event in the coalescent with recombination (coalescence between lineages that share no overlapping ancestral material) results in a model for genealogical structure, genealogical correlation, and patterns of genetic variation that do not differ markedly from the standard model. It is therefore unlikely that inferences about genealogical history made under the sequentially Markov model will differ significantly than those made under the more complex model.

For most species, accurate reconstruction of evolutionary histories that include recombination (ARGs) is not possible (indeed it is theoretically impossible for recombination events that can never be detected; Wiuf et al. 2001; Myers & Griffiths 2003). Furthermore, when rates of recombination and mutation are comparable (as in humans), reliable reconstruction of the true genealogy at any given nucleotide position is an impractical aim. Conversely, estimating likelihoods by Monte Carlo methods that sum over possible ARGs given the data will always be difficult when the data are so uninformative. For these reasons, approximations to the coalescent with recombination are likely to be
the only realistic route to genealogical inference from sequence variation in large datasets.

Many thanks to Chris Spencer, Brian Charlesworth and two anonymous referees for discussion and comments on the manuscript.

REFERENCES


