Approximating the Coalescent with Recombination

A Thesis submitted for the Degree of Doctor of Philosophy

Niall Cardin

Corpus Christi College, University of Oxford

March 30, 2007

Approximating the Coalescent with Recombination

Niall Cardin, Corpus Christi College

D.Phil. Thesis, Michaelmas Term, 2006

Abstract

Recent advances in sequencing and genotyping technologies have caused an explosion in the availability of DNA data. Recent studies have been concerned with characterising patterns of diversity and measuring variation within populations. Understanding these data will require new methodologies which consider the biological and evolutionary processes that underly the data.

The evolution of DNA is a complex and highly random process and as a result the information contained in our DNA sequences about quantities of interest is difficult to extract. Statistical models provide a framework in which to understand these data. However it is extremely challenging to produce models that capture the critical features of the underlying processes while retaining the simplicity required to perform inference.

The coalescent, introduced by Kingman [1], provides a model of the ge-

nealogical process under which simulation of ancestral relationships is straightforward; this was extended to include recombination by Hudson [2]. The coalescent model captures many of the important features of the evolutionary process and has become widely used in population genetics. Unfortunately it is difficult to perform inference under the coalescent: the high dimensional space of genealogies is difficult to explore and when recombination is present the data does not fully inform these ancestral relationships between individuals.

In this thesis I explore approximations to the coalescent under which inference may be more tractable. I investigate models which greatly simplify the ancestral process and provide very efficient computational means of performing inference; I also investigate a new model, the Sequentially Markov Coalescent, which closely mimics the coalescent with recombination.

Using these approximations provides an interesting alternative to full coalescent inference although there may well be considerable improvements that can still be made. I conclude by describing possible approaches to creating new approximate models that capture more of the biological reality of the ancestral process while retaining computational efficiency.

Acknowledgements

I first became interested in mathematical genetics after being inspired when listening to a talk by Gil McVean. Thanks to his efforts I gained this studentship in Oxford with Gil as my supervisor. His guidance, enthusiasm and patience were essential in helping me produce this work. Thank you, for all of your help Gil.

Working in the Mathematical Genetics Group in Oxford has been a hugely rewarding experience because of the fantastic people I've had the chance to work with. I won't attempt to list everyone who I've met here, but it could never have been as fun, rewarding or productive without you. I would especially like to thank the people who helped me find my feet when I arrived: Graham Coop, Christopher Spencer and Daniel Wilson for their careful and clear explanations of research ideas and programming techniques. Also thanks to Adam Auton, Jo Gay and Ella Chase for many useful and enjoyable discussions and who have put up with sharing an office with me and my sometimes unusual ways.

There have been many further members of the department I would like to thank for discussions and advice, in particular: Daniel Falush, Jonathan Marchini, Peter Donnelly, Simon Myers, Jay Taylor, Colin Freeman and Bob Griffiths.

Thanks also to my examiners, Jotun Hein and Richard Durbin.

I would like to thank the Department of Statistics for their generous financial support throughout this project, on which I have been completely dependent.

Thanks to my parents and my sister Cathy, for their huge support and eternal encouragement, you've always believed in me. Thanks finally to Alex who has been there for me whenever I needed him.

Contents

Chapter 1

Introduction

Genetic data provides a rich source of information about organisms. Our DNA contains clues to age old questions concerning human history, human origins, and the differences between humans and other animals. There is also information about important medical phenomena, such as the co-evolution of host-parasite systems and locations of genes involved in diseases with heritable components. It is therefore important to develop methods which can extract useful information from our DNA sequences. Unfortunately the DNA of organisms is affected by highly complex random processes and the resulting signals for quantities of interest are often difficult to extract. Statistical modelling provides a general tool for understanding the patterns of variation that we observe.

In this thesis I introduce some statistical models that have proved suc-

cessful in performing population genetic inference. I explore both complex genealogical models under which inference is extremely challenging as well as simpler models under which inference is computationally feasible, even for large data sets. One class of models that has proved useful in many statistical scenarios is that of Markov models and of *Hidden Markov Models* (HMMs). Hidden Markov models are used extensively in this thesis, in particular in Chapter 2 and as a key component of methods in Chapter 4. In the current Chapter I introduce Markov processes and the basic theory of Hidden Markov Models. I start with a simple example of a Hidden Markov Model and use this to explain how it can be used to perform efficient and powerful inference. I then describe some further applications and scenarios where Hidden Markov Models have been successfully used to help understand genetic data.

I also introduce a model for the genealogy of a population sample of genetic data, the coalescent with recombination [1, 2]. The coalescent provides a prior distribution on genealogies for a sample of size n and allows the efficient simulation of such genealogies and of population genetic data under certain simplifying assumptions about demography and the mutation and recombination processes. At the end of this Chapter I ask how inference can be performed under the coalescent and explain some of the difficulties with full genealogical inference.

1.1 Markov Processes

Given an ordered sequence of random variables, $X_1, \ldots, X_L$, we say that the the X_i form a Markov chain if

$$P(X_{i+1} \mid X_1, \ldots, X_i) = P(X_{i+1} \mid X_i) \tag{1.1}$$

For an in depth discussion of Markov chains and their properties see, for example, Grimmet and Stirzaker [3]. A simple example of a Markov chain is the allelic state of a particular locus on the genome. The type in any individual is determined by the types of its ancestors back in time. However, given the types of the parents and the mutation and recombination rates, the distribution of types is independent of earlier ancestors.

Consider a single locus, l, on the genome of a simple idealised bacterium b_1. Imagine that every 20 minutes the DNA in the bacterium is replicated and a second bacterium splits from the first. Assume that after each replication the bacterium that retained the original DNA dies. After t generations a very simple lineage of individual bacteria, $(b_1, \ldots, b_t)$, has been generated. Assume that there are four possible allelic states at a locus, A, C, G and T and notice that mutation events at locus l give rise to a Markov Chain $X_1, \ldots, X_t$ of random variables representing the type (at l) in each generation.

It is now possible to construct a matrix, $T^{(i)}$, to describe the pattern of mutations in generation i in the following way. If the allelic state of b_i, X_i, is j then the distribution of possible states in generation $i+1$ given state X_i in i is denoted by a vector $T^{(i)}(j)$. As these are the only possibilities for the states in generation $i+1$ the elements of $T^{(i)}(j)$ sum to 1. These transition vectors can then be collected into a matrix $T^{(i)}$, called the *transition matrix*. When the transition matrix is independent of i the Markov chain is said to be *homogeneous*.

Suppose that rate of mutation between bacterial replications does not change in time. Dropping the, now redundant, conditioning on i define the elements of T, $T(j \to j')$, as the transition rate from state j in any generation to state j' at the next. The transition matrix allows simple calculation of the distribution of allelic states after one generation. Given a row vector, v_i, of probabilities (describing knowledge of the allelic states of the bacterium in generation i) then right multiplication by the transition matrix gives the correct row vector of probabilities in generation $i+1$. Similarly, right multiplication by T on the vector of probabilities for generation $i+1$ gives the distribution of states in generation $i+2$. This pair of right multiplications is equivalent to a single multiplication by the square of the transition matrix. More generally

$$P(X_n = j \mid v_i) = [v_i \times T^{n-i}]_j \tag{1.2}$$

The theory of Markov chains allows many efficient computations such as that described above, however that is not the focus of this chapter. The next section describes the situation where there is an underlying Markov process, but this process cannot be directly observed.

1.2 Hidden Markov Models

Sometimes the observable data are not Markov but it is possible to construct Markov models for the underlying processes. A model which is Markov on these hidden states is referred to as a *Hidden Markov Model*. Examples are given in the following sections, but I first introduce the notation and basic mathematics of Hidden Markov Models.

Consider a Markov Chain $\mathbf{X} = X_1, \ldots, X_L$ Suppose that the X_i's cannot be observed, and that only the sequence $D = \mathbf{O} = O_1, \ldots, O_L$ (these are *emitted* by the underlying states) are observed. Define the i^{th} emission probability, $E^{(i)}(o_i, x_i)$ by

$$E^{(i)}(o_i, x_i) = P(O_i = o_i \mid X_i = x_i) \ \ \forall (x_i \leq K), (i \leq L). \tag{1.3}$$

These can then be collected into a K by L matrix, $\mathcal{E}$. Note that then the probability of each observed datum is dependent only on the underlying state

at the same point. This comprises a Hidden Markov Model: a set of possible underlying states X_i, transition rates between the X_i (given by $\mathcal{T}$) and a set of emission probabilities (given by $\mathcal{E}$).

Assuming that the emission and transition rates of a given Hidden Markov Model are known it is then possible to give a straightforward (albeit computationally intensive) method of calculating the probability of a set of observations given these parameters. Using the partition rule:

$$P(\mathbf{O} = \mathbf{o}) = \sum_{\mathbf{x}} P(\mathbf{O} = \mathbf{o}, \mathbf{X} = \mathbf{x}) \tag{1.4}$$

To calculate each of the terms on the right hand side it is often helpful to break the likelihood for each path into a product over each step in the path. Consider a possible sequence of underlying states, $\mathbf{X} = \mathbf{x}$:

$$\begin{aligned}
&P(\mathbf{O} = \mathbf{o}, \mathbf{X} = \mathbf{x}) \\
&\quad = \quad P((O_1, \ldots, O_L) = (o_1, \ldots, o_L), (X_1, \ldots, X_L) = (x_1, \ldots, x_L)) \\
&\quad = \quad P(O_1 = o_1 \mid X_1 = x_1) P(X_1 = x_1) \times \\
&\qquad \prod_{i=2}^{L} P(O_i = o_i \mid X_i = x_i) P(X_i = x_i \mid X_{i-1} = x_{i-1}) \\
&\quad = \quad E^{(1)}(o_1, x_1) T(x_1) \times \prod_{i=2}^{L} E^{(i)}(o_i, x_i) T(x_{i-1}, x_i)
\end{aligned} \tag{1.5}$$

where $T(x_1)$ denotes the probability that the Markov Chain starts in state x_1.

I now introduce an example Hidden Markov Model and some techniques that can be employed to perform computationally efficient inference using Hidden Markov Models.

1.3 Inferring Population of Origin in Admixed Populations

When a population has recently been derived from the mixing of two previously isolated populations the population is said to be admixed. For example, individuals of both African and European descent have mixed in the Americas giving rise to individuals with both recent African as well as European ancestry.

Given such a population it may be useful to infer the population of origin along the genome of extant individuals. Models for performing inference on admixed populations using unlinked markers were developed in 2000 by Pritchard et. al.[4]. In 2003 Falush et. al. [5] described an extension to this model designed for data from linked loci.

The model attempts to capture the following (simplified) scenario. A set,

$\mathbf{X}$ (with $|X| = K$), of ancestral populations have been genetically separated for enough time for genetic drift and other forces to create significant differences in allele frequencies. At some recent point in the past, t_a generations ago, these populations mixed giving rise to random mating between individuals within this mixed population. Since admixture, recombination events have broken up ancestral haplotypes giving rise to individuals with contiguous stretches of genetic material from each population - with the boundaries of such material lying at the positions of historical recombination events. These breaks between ancestral chunks occur at a Poisson rate of 1 per Morgan in each generation. So after t_a generations the breaks between chunks have a Poisson distribution with t_a breaks expected per Morgan.

Formally, consider n individuals sampled at random from the admixed population and typed at L loci and suppose that it is possible to obtain haplotype data. Denote the probability that individual i has directly inherited ancestral material from population k at locus j by $P(X_j^{(i)} = k)$. Let $q_k^{(i)}$ denote the average expected proportion of ancestry from population k in individual i. The transition rate between state k at site j and state k' at site $j+1$ is denoted by $T^j_{k\to k'}(rd_j)$. Then, where r is the recombination rate per unit distance since the start of admixture and d_j is the (physical) distance

between sites j and $j+1$,

$$T^{j}_{k\to k'}(rd_j) = \begin{cases} \exp(-rd_j) + (1-\exp(-rd_j))q^{(i)}_{k'} & \text{if } k' = k \\ (1-\exp(-rd_j))q^{(i)}_{k'} & \text{otherwise.} \end{cases} \tag{1.6}$$

The term $\exp(-d_j r)$ captures the probability of no recombination events between loci j and $j+1$ while the term $(1-\exp(-d_j r))$ is the probability of at least one such recombination. In the former case ancestry at site $j+1$ is guaranteed to match that at site j while in the latter case the population of ancestry is chosen according to the population frequencies. I now drop explicit conditioning on i for notational simplicity and as the following theory pertains to general Hidden Markov Models.

Note that the model of recombination described above is a first order Markov Chain. The ancestry at each site depends only on the ancestry at the last site (in that individual). However, the ancestral states themselves are not directly observed. Instead the alleles at each locus are observed and do not form a Markov chain. Given the population of origin the probability of observing a particular allele is the population frequency of that allele. This model of admixture can then be phrased in terms of a standard Hidden Markov Model from section 1.2. The underlying states are the unknown

ancestral populations, while the observed states are the alleles at each locus. The transition rate between pairs of sites is defined by the recombination fraction (this chain is not homogeneous, the transition rates depend on the genetic distance between sites, and this is different for each pair of sites). The emission matrix is defined by the population frequencies in the ancestral populations.

Given the sequences of a set of individuals in an admixed population it may be interesting to ask, in each individual, which population of ancestry their DNA is derived from at each site. Using equation 1.5 it is possible to work out the joint probability of the data and a path. Then, using Bayes' Rule, the probability of a given underlying sequence given the data (this could then be used to calculate the most likely sequence of underlying states, for example) can be calculated

$$P(\mathbf{x} \mid D) = \frac{P(D, \mathbf{x})}{\sum_{\mathbf{x}} P(D, \mathbf{x})}. \tag{1.7}$$

Calculating the likelihoods for each possible sequence of ancestral states requires a sum over K^L possibilities. However, it is possible to use the Markov structure of the underlying states to construct an efficient algorithm that finds the most likely sequence of underlying states.

1.3.1 The Viterbi Algorithm

The Viterbi algorithm is a *dynamic programming* algorithm that finds the global maximum likelihood sequence of underlying states in a Hidden Markov Model. To do this it is first necessary to consider the likelihood function for a given sequence of underlying (ancestral) states, or *path*, in the presence of data.

Consider Equation 1.5. It can be seen from this that the likelihood of the data and a specific sequence of underlying states can be written as the product of the likelihood up to some point i and the data and hidden states after point i. That is, in calculating the likelihood for sites $i+1$ to L, information about the hidden states at sites i or earlier is not required. Naïvely we would like to directly use this equation to recursively derive the most likely sequence of underlying states site by site, by maximising the terms in this product. This would provide an algorithm linear in the length of the data. Unfortunately, a direct implementation is not optimal as the likelihood calculations at each site require knowledge of the underlying states at the previous site. Instead, the Viterbi algorithm is used, this recursively calculates the most likely path *conditional* on the underlying state at each site (see Figure 1.1).

Denote the probability of the most likely sequence of underlying states

that ends in state k at site j in individual i by $P_i^*(k, j)$. Then

$$P_i^*(k', j+1) = \max_{k=1}^{K} \left(P_i^*(k, j) T(k, k')\right) \times E(j+1 \mid k'). \tag{1.8}$$

Normally the likelihood of the most likely path itself is not of interest and the sequence of underlying states (as in this case) needs to be estimated. For this purpose a matrix, T^* of most likely transitions from state k at site j to state k' at site $j+1$ is stored. That is:

$$T^*_{k',j} = \text{argmax}_k \left(P_i^*(k, j) T(k, k')\right), \tag{1.9}$$

where 'argmax' denotes the underlying state k which maximises the quantity in parenthesis. The most likely state at the final site is chosen according to:

$$T^*_L = \text{argmax}_k \left(P_i^*(k, L)\right). \tag{1.10}$$

It is then possible to *trace-back* the most likely sequence of underlying states as in Figure 1.1. To summarise, suppose we know the frequencies of each allele in each of the K ancestral populations, haplotype data for n individuals and a recombination map for the regions under study. Then the Viterbi algorithm allows us to efficiently calculate the most likely sequence of underlying ancestry at every site in every individual under this model of admixture and

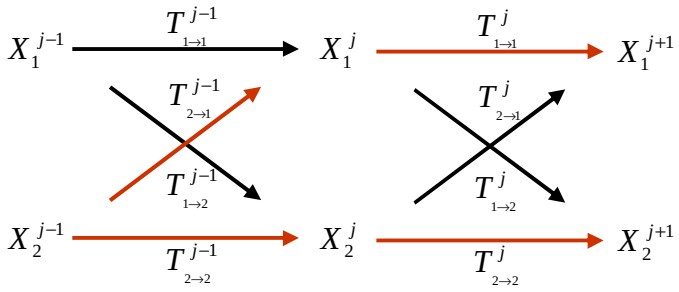

Figure 1.1: This diagram represents the set of possible sequences of underlying ancestry when $K = 2$. The transition rates depend on the recombination fraction between each site (according to Equation 1.6), although this is not shown for notational convenience. Suppose that the most likely path to each state is marked in red. Then if the most likely path in the data is in state 1 at site $j + 1$ then it must also be in state 1 at site j. This is because the path that passes from state 2 at site j to state 1 at site $j + 1$ leads to a smaller value at site $j + 1$ and all subsequent calculations are independent of the path taken to site $j + 1$. In this way, given the knowledge that the most likely path is in state 1 at site $j + 1$ it can be inferred that it is also in state 1 at site j and in state 2 at site $j - 1$ and so on. Note that the most likely state at site j cannot, in general, be inferred until the most likely state at site $j + 1$ is known.

linkage.

1.3.2 Posterior Decoding

The Viterbi algorithm guarantees to find the globally most likely sequence of underlying states, however, often measures of uncertainty in estimates or even the full posterior distribution of underlying states are required. In the case of admixture it may be important to know in which regions ancestry is well defined and where there is little certainty about the parent population. Perhaps an ideal answer would be to give the full posterior probability of

each locus in each individual being derived from each of the K populations. The following section describes how this can be achieved.

The Forward Algorithm The forward algorithm is an efficient algorithm for calculating the likelihood of a set of data under a Hidden Markov Model that takes time quadratic in the number of underlying states and linear in the length of the data. This is often useful in itself, but is also the first step in calculating the posterior distribution of underlying states.

The underlying states are viewed as missing data, using the partition law it is then possible to write an expression for the likelihood, this involves summing over all possible sequences of underlying states,

$$P(D) = \sum_{\mathbf{x} \in \mathcal{X}} P(D \mid \mathbf{X} = \mathbf{x}) P(\mathbf{x}) \tag{1.11}$$

where $\mathcal{X}$ denotes the set of all possible underlying sequences $\mathbf{X}$. However, for K states and data of length L a direct calculation of the quantity requires a sum over K^L terms which is, of course, impractical for most data sets of interest.

Instead of a straightforward sum over all paths, it is preferable to use the *forward algorithm*. This calculates the sum over all paths so that the time taken is linear in the length of the data. Define $P_i(k, j)$ to be the joint probability of the data for individual i when only using the data up to site

j, with underlying state k at j. Then

$$P_i(k, j+1) = E^{(j+1)}(o_{j+1}, k) \times \sum_{k'=1}^{K} P_i(k', j) T_{k',k}. \tag{1.12}$$

These terms are recursively calculated for all $j \leq L$ and the probability of the data, for each individual, is given by $\sum_{k=1}^{K} P_i(k, L)$. Individuals are assumed to be independent so the likelihood of the data is the product of these likelihoods for each individual. Note the similarity between equations 1.8 and 1.12. The reasoning behind both is precisely the same, but the forward algorithm calculates the probability summed over all paths, instead of finding the maximum likelihood path.

The Backwards Algorithm The backwards algorithm does not have such an intuitive interpretation as those of the forward and Viterbi algorithms. Informally the backward algorithm performs exactly the same calculations as the forwards algorithm, but in a different order. More formally, denote the value of the backwards algorithm for individual i at site j in state k by $B_i(k, j)$. Then define $B_i(k, L) = 1, \ \forall k$ and

$$B_i(k', j-1) = \sum_{k=1}^{K} B_i(k, j) T_{k',k} \times E(O_j \mid X_j = k) \tag{1.13}$$

Note then that

$$P(D) = \sum_{k=1}^{K} P_i(k,j) \times B_i(k,j), \ \forall j. \tag{1.14}$$

The backward algorithm greatly increases the range of questions that can be asked of the data, such as: what is the joint probability of the data with state k at site j? To answer this question it is necessary to calculate the backwards quantity to site j in state k and the forwards quantity to site j in state k. The product of these terms gives the probability of the data given state k at site j. That is:

$$P(D, X_j = k) = P_i(k,j) \times B_i(k,j). \tag{1.15}$$

This may be easiest to understand by considering Figure 1.2.

Suppose we now want to know the quantity $P(X_j = k \mid D)$, this can be calculated exactly (under the model) by using Equation 1.15 in conjunction with Bayes' rule:

$$\begin{aligned} P(X_j = k \mid D) &= P(D \mid X_j = k)P(X_j = k)/P(D) \\ &= P(D, X_j = k)/P(D). \end{aligned} \tag{1.16}$$

This procedure is highly efficient as it is necessary to perform the forward and backward algorithms only once each. Then the above equation allows

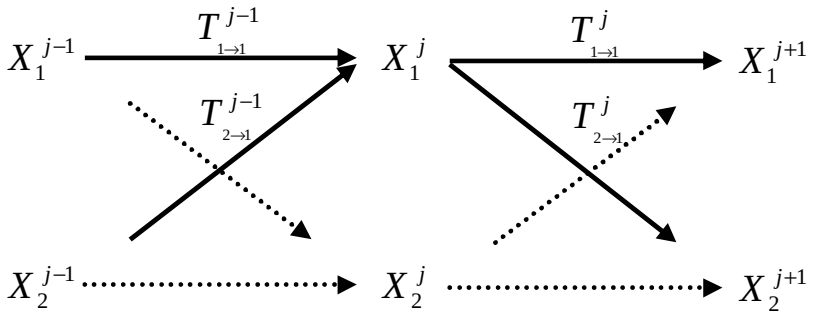

Figure 1.2: This diagram gives a pictorial representation of how the forward and backward algorithms can be combined to produce the probability of the data given state k at site j. The terms preceding site j, calculated by the forward algorithm calculate the probability of the data using only information up to site j. Also, the contribution from state 2 at this site is not included. When the product with the result from the backwards calculation is taken the set of transition and emission probabilities in constructing this value correspond precisely to those that would be subsequently calculated by the forward algorithm if it continued past site j but did not include terms from state 2 at this point.

the straightforward calculation of the vector of all posterior probabilities.

This calculation was performed on high quality human data by Patterson et. al. [6] producing Figure 1.3

1.4 The Lander and Green Hidden Markov Model

In 1987 Lander and Green [8] published a paper introducing one of the first Hidden Markov Models to genetics. The problem concerns constructing a linkage map based on pedigree data. Unlike in the case of admixture map-

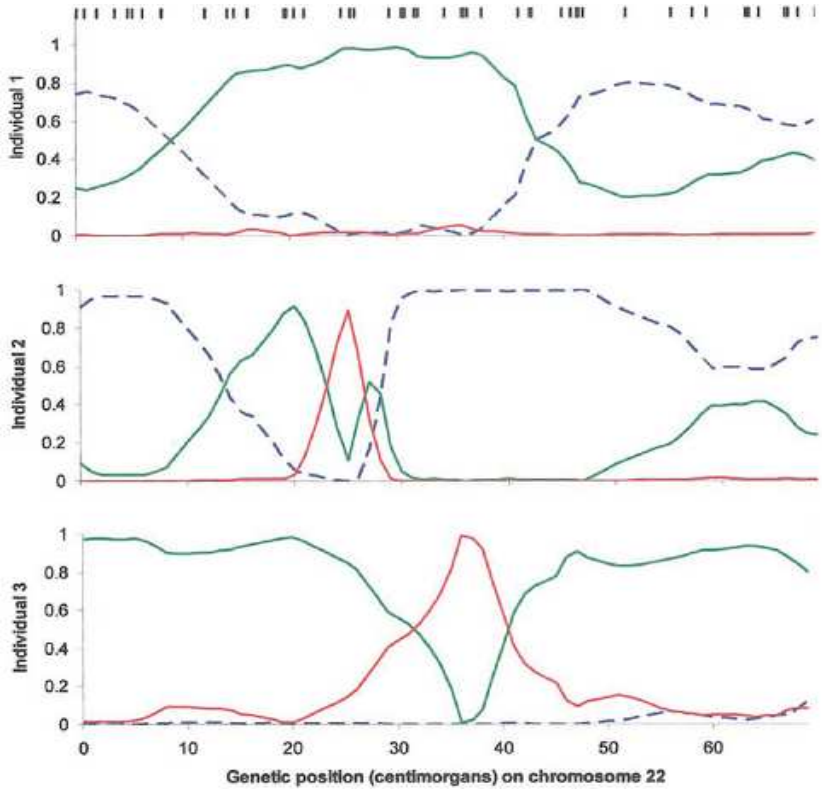

Figure 1.3: This Figure comes from Patterson et. al. [6] and shows the posterior probability of ancestry from the African and European founder populations. The data here is diploid and so there are 3 cases - homozygous for African inheritance, homozygous for European or heterozygous. The data comes from Smith et al [7] and the ancestry is derived from 52 SNPs on chromosome 22. The blue dotted line indicates the posterior probability of no European Ancestry alleles at this locus. The green line shows the probability of one allele of European origin while the red line gives the probability that both alleles are of African Origin.

ping above much of the family structure of the individuals in the sample is known here. The information about recombination is gleaned from inferring specific switches between maternally and paternally inherited loci in a single individual. The recombination fraction is then simply calculated as the expected number of switches inferred at each site.

At each site on each of an individual's haploid chromosomes it is important to know whether the site was maternally or paternally derived. Given this information it is then possible to observe recombination events (see Figure 1.4). However, it is not always possible to distinguish between the maternal and paternal types at each locus. Often the type donated to the child is present in both parents, also genetic data does not include information about the phase of the genotype. Not knowing which type each individual parent had, or which type has been passed to the child can make it impossible to infer recombinations even when the full haplotype data are informative (see Figure 1.5).

Lander and Green introduced a Hidden Markov Model for these unknown patterns of inheritance (given the recombination fractions). Their algorithm for inferring recombination fractions combined this Hidden Markov Model with the expectation maximising (EM) algorithm which iteratively proposes new recombination fractions based on the inferred ancestry given the previous set of proposed fractions. This approach allowed the computation to be

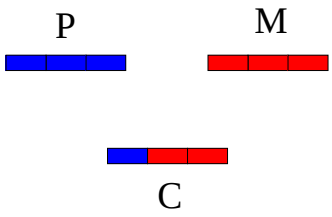

Figure 1.4: In this mock data set the sequences of coloured boxes represent haplotypes, the colours indicate different types. On the left is the paternal chromosome and on the right is the maternal chromosome. If the progeny are observed to have the haplotype displayed underneath then a recombination between the first and second sites can be inferred as the inherited material is paternal to the left of this site and maternal to the right. In most studies there would be many triples, achieved both by typing multiple generations as well as many pairs in each generation.

completed in a short period of time when the pedigree is of small size, even for large numbers of loci. I now consider the problem of calculating the expected number of recombinations given (unphased) genotype data and a vector of recombination fractions. This is a required step in performing the EM algorithm. In their paper [8] Lander and Green use a compact matrix notation to describe the dynamic programming algorithms. I attempt to rewrite the algorithm in the same style as the models described above.

Consider a set of n genotyped individuals with parents who have also been genotyped; such individuals are referred to as *non-originals* in the Lander and Green paper. It is only in these individuals, that have parents in the sample, that recombination events can be inferred. Each individual is typed at L loci in between which are $L-1$ points at which recombination events could have occurred. The recombination fraction between two sites is the probability

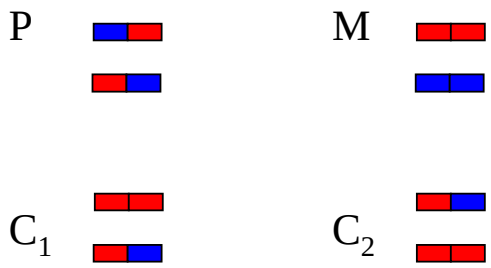

Figure 1.5: In the above diagram there are two parents and two offspring, all typed at two loci. On the left is the paternal parent and the right is the maternal. In this diagram the paternal parent donates the material for the lower of the two sequences and the maternal to the higher. Given child C_1 (left) no recombination need be inferred as the maternal contribution comes from one chromosome in the mother, likewise with the paternal contribution. However, for child C_2 recombination is required as both the maternal contributions and the paternal contributions are recombinant. Without phasing information the difference between C_1 and C_2 is invisible. Both parents have genotypes $\{1, 1\}$ and both children have genotypes $\{2, 1\}$.

of inheriting the genetic information at these sites from different parents in a single generation. Let $\mathbf{r} = \{r_1, \ldots, r_{L-1}\}$ be the vector of recombination fractions between each pair of adjacent sites. Throughout this description I index sites by the letter j and individuals by the letter i. For each site define a binary vector $\mathbf{v_j}$ (of length $2n$) which denotes the inheritance of individual i at site j by $v_{i,j} = 0$ if the gamete inherited its DNA from the parent's paternal chromosome and $v_{i,j} = 1$ otherwise.

In the pedigree, some individuals will be both children and parents of other individuals in the pedigree. For this reason, the paternities at each site are non-independent. Although it would be simpler and faster to take each individual separately this non-independence makes it necessary to consider

the vectors $\mathbf{v_j}$ as a whole, these are the underlying states at each site. The transition rates are independent between individuals and the prior probability of a transition between any two adjacent sites, j and $j+1$, in any individual is precisely the recombination fraction r_j. There are 2^{2n} possible inheritance vectors at any given site and so the formal transition matrix between each site is of size 2^{2n} by 2^{2n}. For a given pair of sites, j and $j+1$, the transition probability between inheritance vectors $\mathbf{v_j} = a$ and $\mathbf{v_{j+1}} = b$ is denoted by $t_j^{a,b}$ and is the product of the individual transition rates in each individual. This leads to the simple formulation that, where d denotes the number of parental differences between a and b,

$$t_j^{(a,b)} = r_j^d \times (1 - r_j)^{2n-d} \tag{1.17}$$

Denote the emission probabilities given that $\mathbf{v_j} = a$ by $q_{j,a}$ so that

$$q_{j,a} = P(D_j \mid \mathbf{v_j} = a) \tag{1.18}$$

Where $P(D_j)$ refers to the probability of the genotypes (at site j) in the non-originals conditional on the types of the individuals with no parents in the sample. Note that it is straightforward to calculate the probability of observing each child's genotype conditional on the inheritance vectors and

phase of the parents. Given probabilities for each phase it is then possible to calculate the likelihood for each child as a sum over all possible phases. Note that by phase I mean here that for each site it is known whether it was maternally or paternally derived - not simply which pairs of sites are inherited from the same individual.

It is now possible to use the forward and backward algorithms to calculate the probabilities $p_{j,a}$ of underlying state a at site j. The forward algorithm recursively describes the joint probability of the data up to site $j+1$ with underlying state $\mathbf{v_{j+1}} = b$ in the standard way:

$$P(D_{j+1} \mid \mathbf{v_{j+1}} = b) = q_{j+1,b} \sum_a P(D_j \mid \mathbf{v_j} = a) t_j^{a,b}. \tag{1.19}$$

For shorthand denote $P(D_j \mid \mathbf{v_j} = a)$ by $F(\mathbf{v_j} = a)$. The 'backward' quantities at each site, $B(\mathbf{v_j} = b)$ are then calculated by defining $B(\mathbf{v_L} = b) = 1$, $\forall b$ and recursively calculating

$$B(\mathbf{v_j} = a) = \sum_b B(\mathbf{v_{j+1}} = b) t_j^{a,b} q_{j+1,b}. \tag{1.20}$$

This then gives

$$p_{j,a} = \frac{F(\mathbf{v_j} = a) \times B(\mathbf{v_j} = a)}{P(D)} \tag{1.21}$$

where $P(D)$ is calculated by $\sum_a F(\mathbf{v_L} = a)$.

To calculate the expected number of recombinations between each pair of sites requires a a further dynamic programming step. Note that, in general, the expectation of a discrete random variable, X, is simply:

$$E(X) = \sum_{i=0}^{\infty} P(X = i)i \tag{1.22}$$

With this in mind define $d(a, b)$ to be the number of parental switches between configurations a and b. Then the expected number of recombinations between sites j and $j + 1$ is

$$E(R_{j,j+1}) = \frac{\sum_{a,b} F(\mathbf{v_j} = a) \times B(\mathbf{v_{j+1}} = b) q_{j+1,b} \times t_j^{a,b} d(a, b)}{P(D)}. \tag{1.23}$$

Informally this is the probability of the data and state a at site j, and state b at site $j+1$ divided by the unconditional probability of the data (which by Bayes rule is the probability of those underlying states given the data). This is then multiplied by the number of recombinations given this transition and summed over all possible underlying states to produce the expected number of recombination events in this interval.

Lander and Green use the EM algorithm to explore the space of possible recombination fractions. In this case the procedure is straightforward: Given a set of recombination fractions $\mathbf{r}_{\text{old}}$, use the HMM to calculate the expected

number of recombinations. Set $\mathbf{r}_{\text{new}}$ as the maximum likelihood estimate of the recombination fractions given the expected number of recombinations (which is straightforward). This procedure is repeated until the likelihood converges to a maximum.

1.5 The likelihood of Data given a Genealogy

One natural method of calculating the likelihood of genetic data involves first constructing a genealogy (or genealogies) that may have given rise to the data and then calculating the likelihood of the data given the genealogy. This approach might involve the calculation of a maximum likelihood tree, as often done in phylogenetics [9]. Alternatively multiple trees might be sampled from a probabilistic distribution usually utilising a model of ancestry and conditioning on the data. It is therefore necessary to be able to calculate the likelihood of the data given a genealogy (see Figure 1.6). In 1973 Felsenstein developed a dynamic programming algorithm to perform this calculation.

The allelic states on a lineage form a Markov Chain and this leads to a natural Hidden Markov model for the probability of genetic data given a genealogy. The observed data is the types of each individual at each site in the present. The hidden states are the types at internal nodes and transition rates are defined by the mutation model. Nothing is observed at the internal

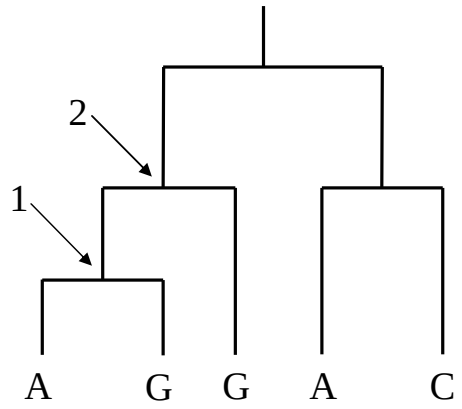

Figure 1.6: This diagram shows an example genealogy where calculating the full likelihood given the genealogy involves a sum over many possible states for the internal nodes. For example, it is not clear what the ancestral states are at nodes 1 and 2. Just as in previous dynamic programming approaches the trick is to calculate the *joint* likelihoods at each node for each possible underlying states.

nodes so the emission probabilities are always 1 at internal nodes. At the leaf nodes the data is directly observed and the emission probabilities are trivial (see Equation 1.24).

Suppose we are given a tree for a given set of n sequences of L sites. Suppose also that we have a mutation model that allows the calculation of the probability of a transition from base a to base b in a given time t along a single branch, $P(a \rightarrow b)_t$. It is also necessary to assume that

1. The transition rates are independent between branches on the tree

2. The transitions are independent between sites

Note that the likelihood for the whole data can be expressed as a product of the likelihoods from each site; I now describe the method for calculating

the likelihood at a single site. Denote a node by ν and the genealogy by G. Given a set of n types at the leaves, x_ν (defined only at the leaf nodes), we wish to calculate the joint likelihood for each node and its state, $s(\nu)$. Denote this likelihood by $P(s(\nu) = a \mid G)$ and note that it can be calculated recursively according to the following scheme: If the node is a leaf node then

$$\begin{aligned} & P(s(\nu) = a \mid G) \\ & = \begin{cases} 1 & \text{if } a = x_\nu \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{1.24}$$

If the node is not a leaf node then it has two daughter nodes ν_1 and ν_2 and associated branch lengths t_1 and t_2 respectively. Then $P(s(\nu) = a \mid G) =$

$$\sum_{b,c} P(s(\nu_1) = b \mid G) P(a \to b)_{t_1} \times P(s(\nu_2) = c \mid G) P(a \to c)_{t_2}. \tag{1.25}$$

The likelihood of the data given the genealogy is then

$$P(D \mid G) = \sum_a P(s(\nu(r)) = a \mid G) \tag{1.26}$$

where $\nu(r)$ is the root node in the tree.

In the same way as before there is an exact analogue between this method for calculating the probability of the data given the genealogy and a method

for calculating a most likely set of underlying types at each node - the Viterbi algorithm. Let $V(s(\nu) = a \mid G)$ denote the likelihood given a most likely set of underlying states up to node ν and with state a at node ν. Then $V(s(\nu) = a \mid G) = P(s(\nu) = a \mid G)$ for all leaf nodes and if ν is not a leaf node then $V(s(\nu) = a \mid G) =$

$$\max_{b,c} V(s(\nu_1) = b \mid G)P(a \to b)_{t_1} \times V(s(\nu_2) = c \mid G)P(a \to c)_{t_2}. \qquad (1.27)$$

In order to *trace-back* a most likely set of underlying states it is necessary to record the types at each daughter node which gave rise to the highest likelihood at each internal node conditional on each state at that node, $\nu(b^*, c^* \mid a)$:

$$\nu(b^*, c^* \mid a) = \text{argmax}_{b,c} V(s(\nu_1) = b \mid G)P(a \to b)_{t_1} \times V(s(\nu_2) = c \mid G)P(a \to c)_{t_2}. \qquad (1.28)$$

It is also possible to sample from the posterior distribution of underlying states using the analogous forward-backward methodology described in Section 1.3.2. The backward algorithm starts at the root node and works towards the leaf nodes and then the product of the terms from the forward and backwards algorithms at each node are used to produce samples at that node.

I now discuss an important population genetic model, the coalescent, which traces the ancestry of a sample backwards in time to the most recent common ancestor. The coalescent, while a considerable simplification of the ancestral process, is unfortunately not amenable to the efficient inference techniques used so far in this chapter. Performing inference under this model is a central topic in the rest of this thesis.

1.6 The Coalescent

Suppose we are given haplotype DNA data from unrelated individuals from a single, neutrally evolving population. There may be many questions we would like to ask: is there population substructure, and if so, what are the levels of migration? Perhaps we wish to know about population history, to identify the geographical origins or to find evidence for recent expansion or decline in a population. Also biological processes, such as mutation and recombination events, may be of interest. Finally it may be important to find links between observed phenotypes (eg. diseases) in individuals and the sequence data, to uncover the underlying genes and hence mechanisms involved.

In analysing such data knowledge of the underlying genealogy would greatly simplify the problem of inference. Population substructure and his-

tory would of course become apparent while estimation of rates would involve merely counting events in the genealogy and measuring the total evolutionary time. Studies designed to find associations between phenotypes and genetic type would also be greatly aided by genealogical information; unfortunately the genealogy of a random set of individuals in a population is usually impossible to observe. It is also usually impossible to directly infer the genealogy using the data, not least because the process of recombination causes the relationships between individuals to change across loci.

One approach to solving this problem is to design statistical models which approximate the ancestral process. Early models such as the Wright-Fisher [10] or Moran [11] traced a finite population of fixed size forwards in time. Such models, while drastic simplifications of real populations, capture many important features of the evolutionary process and provided the basis for theoretical insights and a foundation for population genetics. Many other models have also been proposed, but in this thesis I focus on the Coalescent, introduced by Kingman in 1982 [1].

1.6.1 The Coalescent with Recombination

The Coalescent models the history of a (finite) *sample* of n individuals backwards in time. The simple coalescent assumes a neutral population of size N

(where N is very large compared to n) of constant size with random mating in continuous time. This model can be derived as the limit of the Wright-Fisher or Moran models with an appropriate scaling of time, although it might be viewed as a simple model of evolution in its own right. It is possible to extend the coalescent to include recombination, and in this thesis I follow the method proposed by Hudson in 1983 [2]. For reasons of simplicity it is assumed here that the rate of recombination is constant along the length of the chromosomes.

The ancestry of the sampled chromosomes is traced backwards in time until every site on all of the lineages has reached a common ancestor. At any point in time a pair of lineages can undergo a coalescence event, where the two lineages merge into one. Also a single lineage may recombine, in this case the lineage is split into two separate lineages. To see why recombination events cause lineages to split backwards in time it may be useful to consider the effect of recombination when lineages are traced forwards in time.

Recombination allows the genetic material of descendant chromosomes to be comprised of a combination of two extant chromosomes (see Figure 1.7). To the left of the recombination break point the material is directly descended from one parent chromosome, to the right the material is descended from the other. If genetic information is passed directly from an extant chromosome to a descendant chromosome at a particular locus, then the parent chromosome

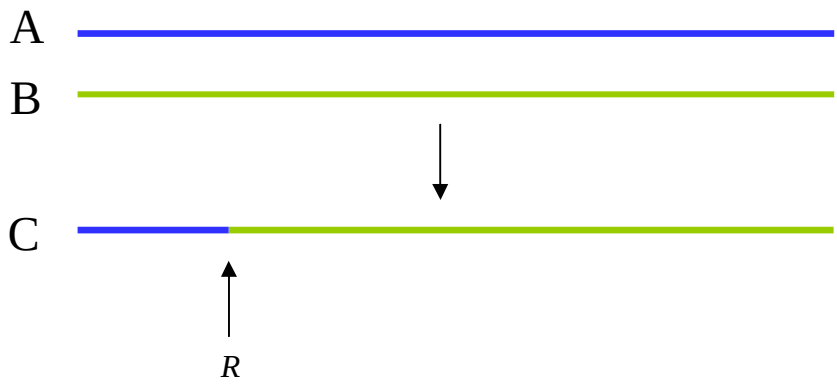

Figure 1.7: Recombination forwards in time: Chromosomes A and B recombine to produce chromosome C. The recombination event occurred at R and A is ancestral to C to the left of R while B is ancestral to C to the right.

is said to be *ancestral* to the descendant chromosome at that locus. For any given lineage we denote the loci at which it is ancestral to a chromosome in the sample as *ancestral material*

When tracing lineages backwards in time recombination events cause chromosomes to be split at a locus. Each resulting chromosome will have ancestral material only on at most one side of this split, see Figure 1.8. It is only important to trace those lineages which hold some material that is ancestral to a chromosome in the original sample. The process keeps track of the ancestral material in each lineage and only events which affect the history of ancestral material are considered. When a coalescent event occurs between two lineages the set of loci on which the new lineage has ancestral material is the union of the set of loci with ancestral material on those lineages. The process stops when every point on the sequence has reached its

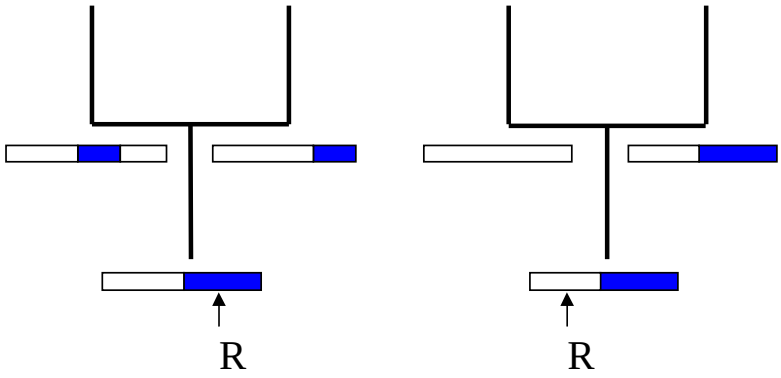

Figure 1.8: Two Recombination events, backwards in time: blue denotes ancestral material. On the left a recombination occurs within ancestral material and this is then split to create two further lineages. The event on the right does not affect the history of ancestral material so is not considered in Hudson's formulation of the process.

most recent common ancestor (MRCA). The time taken to reach the most recent common ancestor ancestor at a site is denoted by 'tMRCA'.

A formal description of the coalescent with recombination follows, it is first necessary to define my notation:

- n : The initial number of chromosomes in the sample
- k : The number of lineages active at a given time
- C_i : The i^{th} lineage
- x_i : The set of intervals on which C_i has ancestral material
- m_i : The number of intervals of ancestral material on the chromosome C_i
- λ_C : The instantaneous rate of coalescence at any given time

- λ_R : The instantaneous rate of recombination at any given time
- $I_{i,j}$: Indicator function indicating whether chromosomes i and j can coalesce
- r : The per generation recombination rate
- N_e : The effective population size
- ρ : $=4N_e r$, the population scaled recombination rate

The process is Markov in time so that only the rate of coalescence, λ_C, and recombination, λ_R, events are required to calculate the distribution of the time to, and type of, the next event. The instantaneous coalescent rate

$$\lambda_C = \sum_{i \neq j} I_{i,j} = k(k-1)/2. \tag{1.29}$$

Write $x_i = \{(x_{i,1}, y_{i,1}), \ldots, (x_{i,m_i}, y_{i,m_i})\}$ where the $x_{i,j}$ and $y_{i,j}$ indicate the upper and lower bounds of the j^{th} interval on the i^{th} chromosome. It is a trivial extension to allow the boundaries of these intervals to lie on a genetic map to allow for recombination rate variation. The instantaneous recombination rate is

$$\lambda_R = \rho/2 \sum_i (y_{i,m_i} - x_{i,1}). \tag{1.30}$$

The time to the next event is exponentially distributed with rate $\lambda_C + \lambda_R$.

The next event is a coalescence with probability $\frac{\lambda_C}{\lambda_C+\lambda_R}$ and is otherwise a recombination event. In the case of a coalescence event two lineages, i, j, are chosen uniformly and at random from those with $I_{i,j} = 1$. The resulting lineage has ancestral material at the union of loci where the lineages i, j had ancestral material. In the case of recombination a point is chosen uniformly on $(x_{i,1}, y_{i,m_i})$ to produce two lineages, each with ancestral material from one side of the split.

At the start of the process $k = n$, $m_i = 1$, $x_{i,1} = 0$ and $y_{i,1} = 1$ for all i. The process stops when a most recent common ancestor has been found at all sites. An example genealogy is given in Figure 1.9

Compared with the biological reality of the evolution of species, the coalescent is a very simple process. The coalescent assumes panmictic mating, in real populations there is complex substructure, due to many factors, acting at both large and small scales. The coalescent with recombination requires a large fixed population size. It is also assumed that the sample size is (very) small compared to the population size (really the 'effective population size' due to Wright [12, 13], although this is not discussed here). Two specific consequences of this, under the coalescent are that

1. No more than two lineages can coalesce at any instant

2. Recombination events never occur between a pair of lineages in the

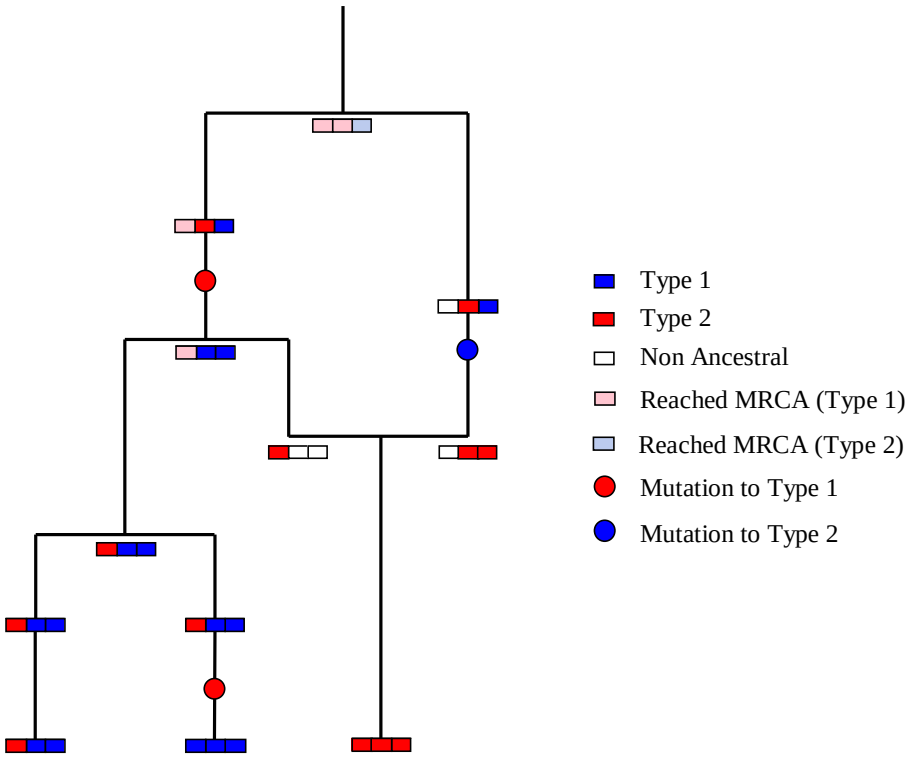

Figure 1.9: This diagram shows an example genealogy with both mutation and recombination events. Lighter colours at a site indicate that the marginal most recent common ancestor has been reached at that site.

sample. Instead the second ancestor of a recombinant chromosome is always assumed to be non ancestral.

Some modern data sets are starting to challenge these assumptions. Case control studies involve thousands of individuals. Furthermore, in the presence of high levels of recombination (as expected over large genomic regions) the number of ancestors at some point in the past may be far in excess of the sample size. This means that multiple simultaneous coalescence events or recombinations between individuals in the sample are no longer unlikely.

Of course, all models must make approximations and the importance of the coalescent in population genetics rests on its ability to simply capture the evolutionary process and to elucidate important information from the data. While it is hard to provide a great deal of evidence that the coalescent model has achieved this goal it has been widely used within population genetics for many years. Coalescent theory has allowed the construction of summary statistics and simplified models that are used to find evidence for population structure, and infer aspects of human history. Perhaps most convincingly: coalescent based recombination rate estimates from McVean et al. [14] show high levels of agreement [15] with previous pedigree analyses and more recent direct experimental approaches such as those of Jeffreys et al. [16]. Although the appropriateness of the coalescent to natural populations

is worth considering I assume for the rest of this thesis that the coalescent provides a reasonable model for inference on population genetic data.

Having decided that the coalescent model is a reasonable model for the ancestry of a sample I turn to the question of performing inference. The coalescent might be described as a simple model, both because of the approximations mentioned above, but also because of the ease and computational efficiency of simulating coalescent genealogies. Unfortunately inference under the coalescent is much less straightforward. The most powerful methods of performing statistical inference involve calculation of the likelihood of parameters of interest. This is problematic under the coalescent because of the enormous state space of genealogies. The likelihood of the parameters, θ, can only be calculated precisely through an integration over all possible genealogies:

$$L(\theta \mid D) = P(D \mid \theta) = \int P(D \mid G, \theta) P(G) dG. \tag{1.31}$$

Direct likelihood calculations based on attempting to evaluate this integral (eg. a recursive method due to Griffiths and Marjoram as part of their paper in 1996 [17]) fail in all but the simplest of scenarios; the summation takes infeasible amounts of computing resources for most data sets of interest. Another approach is to use Monte Carlo estimates of the likelihood, the most

direct approach approximates Equation 1.31 using the formula:

$$\int P(D \mid G, \theta) P(G) dG \approx \frac{1}{M} \sum_{i=1}^{M} P(D \mid G, \theta) \tag{1.32}$$

where the genealogies are simulated from the coalescent prior.

Equation 1.32 gives an extremely poor estimator of the likelihood as most of the genealogies are incompatible with the data. These incompatibilities lead to insignificant contributions to the likelihood from the vast majority of simulations. One method of improving the estimator is to use Markov Chain Monte Carlo (MCMC) techniques (see eg. [18, 19]). Unfortunately these methods remain impractical as it is not known how to construct a chain that converges in reasonable time. Alternatively importance sampling can be used to generate only genealogies compatible with the data. This approach is used in conjunction with the recursion described in Griffiths and Marjoram to provide a more computationally tractable approach. More efficient importance samplers have recently been developed by Stephens and Donnelly [20], in the absence of recombination, and Fearnhead and Donnelly [21], which includes recombination. These methods, while sizeable improvements on previous methods, are also computationally intractable except for very small data. This is discussed in more detail in Chapter 4.

This thesis explores the possibility of using approximations to the coales-

cent process to provide more computationally efficient means to calculate the likelihood of the data. In 2000 Stephens and Donnelly suggested a model of evolution that could be used to approximate the probability of observing a haplotype given a pre-existing set of haplotypes in the absence of recombination. This model is based on the notion that the haplotype can be constructed as an imperfect *copy* of the previous haplotypes and is sometimes known as the 'Look down and Copy' model (of sequence evolution). In 2001 the Look down and Copy approach was extended by Fearnhead and Donnelly to include recombination. They introduced a Hidden Markov Model that was adapted by Li and Stephens to produce a fast algorithm that approximates the likelihood of a set of population genetic data. The likelihood is calculated using a product of these approximate conditionals and is often referred to as a PAC likelihood.

This model has caused widespread interest and many methods (see eg. [22, 23]) have been designed which take advantage of its computational efficiency, although most have not yet been published. Although the scheme can be used to tackle a wide range of questions in population genetics that have previously been computationally impractical, the underlying model suffers from both theoretical and practical drawbacks that affect accuracy. In Chapter 2 I investigate possible alternatives to the scheme proposed by Li and Stephens. I also analyse the strengths and weaknesses of the PAC approach

and the limitations imposed by using models of this form.

In Chapter 3 I propose a new model, the *Sequentially Markov Coalescent* or SMC, for the ancestry of a population in the presence of recombination. This model approximates the coalescent with recombination and involves a simple alteration to the coalescent. I prove that this alteration gives rise to Markovian structure when genealogies are generated along sequences. The state space of genealogies is also much reduced under this model in the presence of high recombination rates. In Chapter 3 I investigate the properties of this model, how well it approximates the coalescent. I look both at the properties of genealogies sampled under the SMC and of data simulated using it.

Having introduced a new approximation to the coalescent I investigate the potential for population genetic inference under the SMC in Chapter 4. While there are many possibilities for using the structure of the SMC to improve computational efficiency, investigating even a handful of these fully is beyond the scope of this thesis. I focus on the method of genealogical sequential importance sampling - following the method of Fearnhead and Donnelly [21] from 2001.

The primary goal of Chapter 4 is to compare inference under the SMC to the coalescent. I also investigate the performance, limitations and potential for improvement of genealogical importance sampling. The performance

improvement that might be expected under the SMC depends broadly on the quantity of data that can be analysed using this technique and so improvements to this method will impact on the ability of the SMC model to improve inference.

Finally, in Chapter 5, I discuss the conclusions I have arrived at after performing the analyses here and suggest possible improvements for future methods designed to perform inference on recombinant population genetic data.

Chapter 2

Non Genealogical Approximations to the Evolutionary Process

2.1 Introduction

Although much can be learned from population genetic data, it is often hard to extract accurate estimates of quantities of interest [20, 24]. In most cases, the full signal for evolutionary parameters or other important quantities, such as recombination rates or the phase of diploid genotype data, can only be extracted by modelling the complex evolutionary history of the data. The highly stochastic nature of the processes involved and the weak information in

recombinant data about the ancestry creates a situation where full inference is extremely computationally intensive, even for small data sets. Approximations are therefore required and in constructing these approximations it is important to capture the key features of the process and the stronger signals in the data, but to avoid modelling those complex processes which contribute little extra information about the parameters of interest.

The best current methods for performing full likelihood inference on population data use a Monte Carlo average from genealogies simulated under the coalescent model of ancestry [20, 21, 18, 19]. Unfortunately, using current methods the number of genealogies required to get an accurate estimate of the likelihood grows extremely fast as the size of the data increases; these methods are prohibitively computationally expensive for anything but the smallest data sets. However, by understanding the properties of more complete models it may be possible to design new approximate models which capture more of the information in the data. Theoretical properties of the ancestral process and the distributions of certain quantities (such as the distribution of the evolutionary time between a pair of sequences) can be calculated under the coalescent and used to inform new approximations. The key is to capture the primary features of the processes leading to genetic variation data while simplifying the mechanism in order to produce computational efficiency.

The models under discussion in this chapter all describe the distribution

of the type of a new sequence conditional on having observed a pre-existing set of sequences. The models try to capture certain features of the coalescent process:

- Under the coalescent a new sequence would show high degrees of local similarity to those already in the sample, however the new sequence might be most closely related to different sequences along its length. The rate that these changes between sequences occur is a function of the recombination rate. The parameter ρ in these models directly affects the probability of switching between the 'inherited' sequences in the sample.

- Under the coalescent process the rate of coalescence rises faster than the rate of recombination or mutation as the sample size increases. To capture this the approximations here all propose sequences more similar to those in the sample when the sample size is large.

- There is a correlation between the density of mutations on a new sequence and the distances between recombination events, this correlation is due to the effect of the evolutionary time separating the sequences. For example sequences separated by a short evolutionary distance will show a low density of allelic differences and share large regions unbroken by historical recombination events.

- Under the infinite sites model it is possible for the data to be incompatible with a single tree, and it is therefore possible to be certain that recombination events have occurred in the history of the sample. Approximate models should capture this strong form of evidence for recombination. It is also important to avoid inferring recombination when little or no evidence for recombination is present.

Not all of these features are captured by all of the models and none of them correctly interpret the signal for recombination given by incompatibilities. These qualities are discussed in more detail later.

2.2 Li and Stephens

In 2003 Li and Stephens [22] developed a model to approximate the process of inheritance. This model allows fast inference on data with recombination, even for large data sets.

Let $h_1, \ldots, h_n$ denote n sampled haplotypes, typed at L bi-allelic loci. The key observation made by Li and Stephens is that:

$$P(h_1, \ldots, h_n) = P(h_1) * P(h_2 \mid h_1) * \ldots * P(h_n \mid h_1, \ldots, h_{n-1}) \qquad (2.1)$$

This expresses the likelihood in terms of a product of conditional proba-

bilities that may be easier to approximate than the full likelihood itself. Unfortunately this observation in itself doesn't allow efficient calculation of the likelihood as no practical exact methods are known to calculate these conditionals. However approximate models for generating a new haplotype given a set of known haplotypes have been developed [21, 22, 25]. In this chapter I present four schemes which use the approach of calculating a 'product of approximate conditional' (or PAC) likelihoods. I analyse the strengths and weaknesses of the overall approach and also the relative merits of the individual approximations. Informally, all the processes for generating a new haplotype, h_{k+1}, from k pre-existing haplotypes, $h_1, \ldots, h_k$, can be described as follows:

1. As haplotypes share a common ancestor they should be related; this is done by allowing h_{k+1} to 'copy' previously considered haplotypes.

2. Some pairs of sequences show regions of similarity due to a high degree of relatedness but recombination allows different relationships at different points along the genome, so h_{k+1} may copy a different h_j at different loci.

3. Mutation creates SNPs which are novel, so that haplotypes derived from the same ancestor at a locus may differ.

In this way the schemes create haplotypes which are an imperfect mosaic of

the haplotypes already in the sample. That is, for $k \geq 1$, at each SNP h_{k+1} is an imperfect copy of h_j for some j. To calculate $P(h_1)$ they assume that the first haplotype is equally likely to be any of those in the sample.

The following is a formal description of the Li and Stephens model including a method for efficient likelihood calculations under this model. I will refer to this model as $\pi_{L\&S}$.

Description of $\pi_{L\&S}$

Let X_j denote which haplotype h_{k+1} copies from at site j. To approximate recombination they model the X_j as a Markov chain on $\{1, \ldots, k\}$ with $P(X_1 = x) = 1/k$ $(\forall x \in \{1, \ldots, k\})$. Let d_j denote the physical distance between markers j and $j + 1$. Define the recombination parameter by $\rho_j = 4Nc_j$ where N is the effective population size, and c_j is the average rate of crossover per unit physical distance, per meiosis, between sites j and $j+1$. The transition rate is constructed assuming that recombination events occur according to a Poisson process of rate $\rho_j d_j/k$, the division by k here reflects the shorter copying time when there are more sequences in the sample. The probability of no recombination events between sites j and $j + 1$ is therefore $e^{-\rho_j d_j/k}$ and the probability of at least one recombination is $(1 - e^{-\rho_j d_j/k})$. Given a recombination there is probability $1/k$ of recombining to any one of the k sequences already in the sample (including that copied at the previous

site). In

$$
\begin{aligned}
&Pr(X_{j+1} = x' \mid X_j = x) \\
&= \begin{cases} \exp(-\rho_j d_j/k) + (1 - \exp(-\rho_j d_j/k))(1/k) & \text{if } x' = x \\ (1 - \exp(-\rho_j d_j/k))(1/k) & \text{otherwise} \end{cases}
\end{aligned}
$$

To capture the mutation process they allow 'imperfections' in the copying scheme, that is, with probability $k/(k+\theta)$ the copy is exact and with probability $\theta/(k+\theta)$ a 'mutation' occurs. These approximate the probability that the next event back in time on the new sequence is a coalescence event (rate k) or mutation (rate θ). Specifically, if $h_{i,j}$ denotes the allele at site j in haplotype i then, given the copying process $X_1, \ldots, X_S$ the alleles $h_{k+1,1}, h_{k+1,2}, \ldots, h_{k+1,S}$ are independent, with

$$
\begin{aligned}
&P(h_{k+1,j} = a \mid X_j = x, h_1, \ldots, h_k) \\
&= \begin{cases} k/(k+\tilde{\theta}) + \frac{1}{2} \times \tilde{\theta}/(k+\tilde{\theta}) & h_{x,j} = a \\ \frac{1}{2} \times \tilde{\theta}/(k+\tilde{\theta}) & h_{x,j} \neq a. \end{cases}
\end{aligned} \tag{2.2}
$$

The term here corresponding to when h_{k+1} copies x at site j and retains the same type is a sum over two possible terms: the probability that no mutation events have occurred since the common ancestor of h_{k+1} and x added to the probability of a sequence of mutations that leave h_{k+1} and x with the same

type. The term $\tilde{\theta}/(k+\tilde{\theta})$ is an approximation to the probability that the first event backwards in time on h_{k+1} is a mutation event, $k/(k+\tilde{\theta})$, approximates the probability that the first event is a coalescence event. Li and Stephens use the quantity $\tilde{\theta} = \left(\sum_{i=1}^{n-1} \frac{1}{i}\right)^{-1}$ as their mutation parameter.

The above formulation of $\pi_{L\&S}$ provides a description of how to simulate data under their new model but the problem of inference is more difficult. The likelihood can be viewed as a sum over all possible histories that could give rise to the data. The advantage of the Li and Stephens model is that this now becomes tractable; the forward algorithm (see Chapter 1) can be used to sum over all possible states.

To do this, let $h_{k+1,\leq j}$ denote the types of the first j sites of haplotype h_{k+1}. For ease of notation let $\alpha_j(x) = P(h_{k+1,\leq j}, X_j = x)$ and let $\gamma_{j+1}(x) = P(h_{k+1,j} = a \mid X_j = x, h_1, \ldots, h_k)$, (given in (2.2)). Then $\alpha_1(x)$ can be calculated directly for $x = 1, \ldots, k$ as $1/k$ multiplied by the appropriate formula from Equation 2.2 at site 1. To compute $\alpha_2(x) \ldots, \alpha_L(x)$ the forwards algorithm (see Equation 1.12 in Chapter 1) can be used:

$$\begin{aligned} \alpha_{j+1}(x) &= \gamma_{j+1}(x) \sum_{x'=1}^{k} \alpha_j(x') P(X_{j+1} = x \mid X_j = x') \\ &= \gamma_{j+1}(x) \left(p_j \alpha_j(x) + (1 - p_j) \frac{1}{k} \sum_{x'=1}^{k} \alpha_j(x') \right), \end{aligned} \tag{2.3}$$

where $p_j = \exp(-\rho_j d_j/k)$. The value of $\pi_{L\&S}(h_{k+1} \mid h_1, \ldots, h_k)$ is then calculated as $\sum_{s=1}^{k} \alpha_L(s)$. This formula precisely mimics that given in Equation 1.12 in Chapter 1.

The advantage of this scheme is that it can be calculated very fast. The dynamic programming algorithm takes $O(Ln^2)$ time to compute which is possible even for large data sets.

Li and Stephens note that their scheme captures certain aspects of coalescent data, when simulating a new haplotype given a set of previously simulated haplotypes:

1. The new haplotype is likely to approximate haplotypes with high frequency in the sample.

2. The probability of seeing new haplotypes decreases as the sample size increases.

3. The probability of seeing a new haplotype increases as the mutation rate increases.

4. When a haplotype is not an exact copy of a previously simulated haplotype it will typically differ by only a small number of mutations. It is unlikely to be completely different to all existing haplotypes.

5. The new haplotype may be very similar to one previously sampled

haplotype in one location while being more closely related to other haplotypes at other locations due to recombination.

Unfortunately, there are problems with this method, both theoretical and practical. Due to the approximations made in calculating $P(h_{k+1} \mid h_1, \ldots h_k)$ the ordering in which samples are considered influences the likelihood of the data strongly. To attempt to compensate for this it is necessary to average over orderings. However the number of possible orderings is so high as to make this infeasible even for a small sample, so only a fraction of these can be used. Li and Stephens use 20 orderings to calculate the likelihood; they claim that although the likelihood can change significantly for different orderings the shape of the curve remains roughly constant and relative likelihoods can be calculated as long as the same set of orderings is used for each value of the parameters. This, at least, should allow estimated maximum likelihood estimates of parameters not to vary greatly between runs. Unfortunately, even the averaged likelihoods produced by this method provide biased estimates of ρ. Furthermore this bias is hard to model and no simple correction is available. In their paper Li and Stephens decided to measure the bias in a number of data sets and adjust their estimates based on these measurements. It is therefore unknown how the bias in $\pi_{L\&S}$ may change when the data does not correspond well to those data sets used to inform their bias

correction. For example, rate estimation may deteriorate in the presence of variable recombination rate estimation.

2.3 Alternatives to $\pi_{L\&S}$

The problems encountered by the scheme of Li and Stephens are a result of the fact that the conditional likelihoods they use are only an approximation to the true likelihoods. Were the true conditional likelihoods known then only one ordering would be required and there would be no bias. With this in mind I propose three alternative approximations to these likelihoods that attempt to capture further features of the data. A natural model to consider is the model on which $\pi_{L\&S}$ is based, that proposed by Fearnhead and Donnelly to generate approximate likelihoods in their importance sampler [21]. This scheme includes an explicit concept of the evolutionary time between the haplotype under examination and the haplotypes it copies at each site. Under this scheme the rate of mutation depends on the time between the haplotypes. I also investigate two novel schemes, the first can be viewed as an extension of the scheme proposed by Fearnhead and Donnelly to allow the time between sequences to also affect the rate of recombination and the final approach is to explicitly model the block like nature of inheritance and calculate the likelihoods for specific blocks analytically.

2.3.1 Fearnhead and Donnelly, $\pi_{F\&D}$

Firstly I investigate the scheme proposed by Fearnhead and Donnelly in 2001 [21]. Although this scheme was not intended as a full model for sequence evolution it is an approximation to the quantity $P(h_n \mid h_1, \ldots, h_{n-1})$, denoted here by $\pi_{F\&D}$. To understand the differences between $\pi_{L\&S}$ and $\pi_{F\&D}$ it is necessary to consider the genealogical time separating a pair of sequences, t. Under the coalescent when t is large the expected density of mutations is higher and the distances between recombination events are shorter on average. When t is smaller the reverse holds with sparse allelic differences and long shared tracts unbroken by recombination.

A natural approach to dealing with the unknown parameter t is to integrate over all possible t. The Li and Stephens model can be viewed as an integration over t at each site, independently. Fearnhead and Donnelly explicitly model t in their conditional likelihood calculations. In regions unbroken by recombination events t should not change, and this is implemented in their model. The rate at which mutations arise between sequences is then dependent on t. In the model of Fearnhead and Donnelly the rate at which recombination occurs is independent of t.

The evolutionary time between sequences is a continuous quantity and so implementing t in a Hidden Markov Model is not directly amenable to effi-

cient likelihood calculation. Instead, using the known distribution of t under the coalescent Fearnhead and Donnelly propose the numerical approximation of *Gaussian Quadrature* (see eg. Kreyszig [26]) to perform their block-wise integration over t. For a given number, a, Gaussian Quadrature is used to define specific values of the times between sequences, $t_1, \ldots, t_a$, and associated weights, $w_1, \ldots, w_a$ so that for a general function f,

$$\int_0^\infty \exp(-t) f(t) dt \approx \sum_{i=1}^{a} w_m f(t_i). \tag{2.4}$$

The weights are chosen so that $\sum_{i=1}^{a} w_i = 1$. Using this approximation it is then possible to use the same site by site dynamic programming approach used in the Li and Stephens scheme, $\pi_{L\&S}$. I use $a = 4$ in this implementation to coincide with the values used by Fearnhead and Donnelly. This corresponds to four times as many possible states in the HMM as in the Li and Stephens approach: at every site there are four times at which the new sequence might copy from each of the already sampled sequences.

The scheme proposed by Fearnhead and Donnelly built on the 'Look down and Copy' model designed by Stephens and Donnelly [20] in 2000. Fearnhead and Donnelly envisage a two stage process by which the new sequence is derived. Firstly recombination events are placed on the length of the sequence and, in between these events, a sequence to 'copy' from is

chosen. In each of these blocks containing no recombination the Look down and Copy model from Stephens and Donnelly is used to determine the allelic states.

To perform inference a Hidden Markov Model is used and I follow the approach given by Fearnhead and Donnelly. To calculate the transition rates Fearnhead and Donnelly consider the next event backwards in time for each pair of adjacent sites. Either the next event is a recombination or a coalescence event. When a coalescence event occurs then the same sequence is copied at both sites. Otherwise a new sequence is chosen independently of the copied sequence at the last site. In this case the weights from Gaussian quadrature are used to determine the probability of each discrete time point. More formally, and using the same notation as before, denote the time that h_{k+1} copies haplotype X_j from at site j by t_j. The transition probabilities are then given by

$$
\begin{aligned}
&Pr(X_{j+1} = x', t_{j+1} = t' \mid X_j = x, t_j = t) \\
&= \begin{cases} \frac{k}{k+\rho_j d_j} + \left(\frac{\rho_j d_j}{k+\rho_j d_j} \times \frac{w_t}{k}\right) & \text{if } x' = x, t' = t \\ \frac{\rho_j d_j}{k+\rho_j d_j} \times \frac{w_{t'}}{k} & \text{otherwise} \end{cases}
\end{aligned}
$$

In Fearnhead and Donnelly 2001 they use a general multi-allelic mutation model. The probability of observing each non-recombinant sequence segment

is defined by a double summation over all sequences in the sample and mutation patterns that give rise to the new sequence from these already known sequences. So as to perform direct comparisons with the $\pi_{L\&S}$ and due to the recent interest in SNP data I have used the bi-allelic mutation model used by Li and Stephens with the same mutation parameter, $\tilde{\theta}$ but now scaled by t to reflect the evolutionary total time between the sequences. Set

$$\begin{aligned} P(h_{k+1,j} = a \mid X_j = x, h_1, \ldots, h_k) \\ = \begin{cases} \frac{k}{k+\tilde{\theta}t} & h_{x,j} = a \\ \frac{\tilde{\theta}t}{k+\tilde{\theta}t} & h_{x,j} \neq a \end{cases} \end{aligned} \tag{2.5}$$

where k is the number of sequences already in the sample. An alternative formulation would be to set the mutation probability to $(1 - \exp(-\theta/k))$ as though the time t were known exactly. If many time points are used this seems more appropriate, of course, for only 1 time point (and so perhaps also for a small number) then the current formulation is more comparable with Li and Stephens. The forward algorithm (from Chapter 1) can be used to calculate the likelihood of the complete data in $O(n^2 L)$ time (remember that n is the number of individuals in the whole sample).

Let t_j denote the time that h_{k+1} copies from haplotype X_j at site j. Define $\alpha_j(x,t) = P(h_{k+1,\leq j} \mid X_j = x, t_j = t)$ and $\gamma_{j+1}(x,t) = P(h_{k+1,j} = a \mid$

$X_j = x, t_j = t, h_1, \ldots, h_k)$, (given in (2.5)). Then

$$\begin{aligned}
&\alpha_{j+1}(x,t) = \\
&= \gamma_{j+1}(x,t) \qquad \sum_{x'=1}^{k}\sum_{t'=1}^{4} \alpha_j(x',t') P(X_{j+1} = x, t_{j+1} = t \mid X_j = x', t_j = t') \\
&= \gamma_{j+1}(x,t) \quad \times \quad \left(p_j \alpha_j(x,t) + (1-p_j)\frac{1}{k}\sum_{x'=1}^{k}\sum_{t'=1}^{4} w_{t'}\alpha_j(x',t') \right), \qquad (2.6)
\end{aligned}$$

where $p_j = \frac{\rho_j d_j}{k+\rho_j d_j}$. The value of $\pi_{F\&D}(h_{k+1} \mid h_1, \ldots, h_k)$ is then given by

$$\sum_{s=1}^{k}\sum_{t=1}^{4} \alpha_L(x,t). \qquad (2.7)$$

This scheme is slower than Li and Stephens but the fixed time within non-recombinant blocks may be more appropriate to population genetic data. However a simplifying assumption that the recombination process is independent of the evolutionary time between sequences is made. The next scheme attempts to incorporate the effect of evolutionary time into the recombination process.

2.3.2 A new algorithm, π_R

Allowing the rate of recombination to depend on the time between sequences is a natural extension of the Fearnhead and Donnelly model to try to capture more features of the evolutionary process. However care must be taken:

recombination events are more likely when the evolutionary time is large. In a direct implementation that allowed recombinations to depend on time, this greater rate of recombination would lead to a bias in the distribution of the tMRCA at each site (away from large copying times). Under the coalescent (and in the absence of data) the marginal distribution of the evolutionary time between two haplotypes should be exponential of rate 1. Using the same approach as Fearnhead and Donnelly I approximate this distribution by a number of discrete time states and appropriate weights that are chosen using Gaussian quadrature. It is then possible to remove the bias against large copying times by altering the transition rates between different time states.

More formally, consider a possible states, $t(1), \ldots, t(a)$, which represent the possible times at which copying can occur. Let t_j denote the time at site j and let the event $R_{j,j+1}$ denote a recombination between sites j and $j+1$. To account for bias I introduce transition rates from time t to time t' conditional on recombination: $\lambda_{t,t'}$. Now, the model is defined to be symmetric in the direction that the haplotypes are read. Also the prior distribution of evolutionary time between sequences does not change along the sequence. Hence the probability of a recombination event given time information at either of its flanking sites (but not both) is independent of whether it is the left or right hand site known about. That is: $P(R_{j,j+1} \mid t_j = t) = P(R_{j,j+1} \mid t_{j+1} = t)$.

Further, the model assumes independence in the copying process across a recombination breakpoint so the $\lambda_{t,t'}$'s are independent of t and can be viewed as the probability of entering time state t' given a recombination event - ie. $\lambda_{t,t'} = \lambda_{t^*,t'} = \lambda_{t'}\ \forall t, t^*$. This approximation can be overcome in the two sequence case, however for multiple sequences the situation is more complicated and it seems unlikely to be a major contributing factor to the efficacy of these schemes. The $\lambda_{t'}$ can be derived, taking an arbitrary site j:

$$\begin{aligned}
\lambda_{t'} &= P(t_{j+1} = t' \mid R_{j,j+1}) \\
&= \frac{P(R_{j,j+1} \mid t_{j+1} = t')P(t_{j+1} = t')}{P(R_{j,j+1})} \\
&= \frac{P(R_{j,j+1} \mid t_j = t')P(t_0 = t')}{P(R_{j,j+1})} \\
&= \frac{r_{t'} w_{t'}}{\bar{r}}
\end{aligned} \tag{2.8}$$

where the $w_{t'}$ are the weights assigned to each time point by Gaussian quadrature (and therefore correspond to the approximation to the coalescent time distribution, which we hope to achieve at stationarity) and $\bar{r} = \sum_{i=1}^{k} w_{t(i)} r_{t(i)}$. In the absence of data there should be no net shift along the sequence in the probability of being in each state so, by symmetry in the direction in which the sequence is read, the total transition between the different states should

be equal:

$$w_t r_t \lambda_{t,t'} = w_{t'} r_{t'} \lambda_{t',t} \quad \forall t, t'. \tag{2.9}$$

Note that the solution given in equation 2.8 trivially satisfies the above criterion.

It is now possible to give a mathematical formulation of the model using the same notation as in the previous two cases. We use a transition rate of a similar form to that of Li and Stephens' scheme.

$$\begin{aligned} &Pr(X_{j+1} = x', t_{j+1} = t' \mid X_j = x, t_j = t) \\ &= \begin{cases} \exp(-\rho_j d_j t/k) + (1 - \exp(-\rho_j d_j t/k))(\lambda_{t'}/k) & \text{if } x' = x, t = t' \\ (1 - \exp(-\rho_j d_j t/k))(\lambda_{t'}/k) & \text{otherwise.} \end{cases} \end{aligned} \tag{2.10}$$

The factor $\lambda_{t'}/k$ is the marginal prior probability of copying from any specific sequence at time t'.

I use the same mutation model as in my implementation of the Fearnhead and Donnelly scheme:

$$\begin{aligned} &P(h_{k+1,j} = a \mid X_j = x, t_j = t, h_1, \ldots, h_k) \\ &= \begin{cases} \frac{k}{k+\tilde{\theta}t} & h_{x,j} = a \\ \frac{\tilde{\theta}t}{k+\tilde{\theta}t} & h_{x,j} \neq a \end{cases} \end{aligned} \tag{2.11}$$

where k is the number of sequences already in the sample.

Again the forward algorithm can be used to calculate the likelihood in $O(n^2L)$ time in precisely the same way as used in the Fearnhead and Donnelly scheme.

2.3.3 An Explicit Block-wise approach, π_{L^2}

The previous models for calculating the likelihood of a haplotype given a set of previously sampled sequences have allowed the likelihood to be calculated site by site. This model considers all of the data in regions, bounded by recombination events (and unbroken by recombination). Within these regions the likelihood, given a sequence to copy from, is calculated analytically. Let the term *block* denote a stretch of a single haplotype bounded by two recombination events with no internal recombination. This algorithm is indexed by the *gaps* between sites. For each gap between sites, i, a partial likelihood, $P(D_i \mid \theta, \rho, R(i))$, using only the data up to i and conditional on a recombination at this point, $R(i)$, is calculated. I use $P(D_i)$ as a shorthand for this quantity and $P(D_{i,j})$ to denote the likelihood conditional on recombinations at j (at the left) and i (at the right) and when only the sites from j to i are considered. For ease of description I first consider only two sequences typed only at L segregating loci. I derive an approximation to the joint proba-

bility of observing these sequences in a region unbroken by recombination and bounded on the right by recombination at i. The likelihood is then a recursive sum over all possible block combinations. For $i > j$ the recursion can be described as follows:

$$P(D_i \mid \theta, \rho, R(i)) = \sum_{j=1}^{i} P(D_{i,j} \mid \theta, \rho, (j \to i)) P(j \to i) P(D_j). \tag{2.12}$$

where $j \to i$ denotes an unbroken block ending at i (by recombination, unless $i = L$) given the start of the block is at j. The likelihood of the parameters given all L sites is then $P(D_L \mid \theta, \rho)$. Using (2.12) above it can be seen that the algorithm takes order L^2 time to calculate the likelihood of the next haplotype.

Consider the quantity

$$P(D_{i,j} \mid \theta, \rho, (j \to i)) \times P(j \to i).$$

It is easiest to view this as the joint probability, $P(D_{i,j}, (j \to i) \mid \theta, \rho)$, for this section. Let j, i be the first and last sites of a given block respectively then, for ease of notation, I drop the explicit conditioning on θ and ρ to define

$$P(\mathbf{m}, (j \to i)) = P(D_{i,j}, (j \to i) \mid \theta, \rho) \tag{2.13}$$

where $\mathbf{m}$ denotes the pattern of mutation in this block. I now derive $P(\mathbf{m}, (j \rightarrow i))$ for a block that is bounded on the left by the start of the data and is bounded on the right by a recombination event. To calculate this quantity it is first necessary to derive the distribution of block lengths in the absence of data. I do this by considering two discrete sequences of infinite length in a Wright-Fisher population of N individuals.

Define the time in a generation to be 1. Only segregating sites are explicitly considered, to account for this the distance between the j^{th} and $j + 1^{\text{th}}$ segregating sites is denoted by d_j. The (per generation) recombination rate in this region is denoted by r_j and recombination events are assumed to be independent. The probability of no recombination between the first l segregating sites (hence $l - 1$ gaps between sites) and at least one recombination between sites l and $(l + 1)$ in the first τ generations is therefore:

$$P(l \mid \tau) = \prod_{j=1}^{l-1}(1 - r_j d_j)^{2\tau} \times (1 - (1 - r_l d_l)^{2\tau}). \tag{2.14}$$

For a diploid population with N individuals the Wright Fisher model

gives the prior distribution on τ as

$$P(\tau) = \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{\tau-1}$$

hence (2.15)

$$P(l) = \sum_{\tau=1}^{\infty}\left(\frac{1}{2N}\left(1-\frac{1}{2N}\right)^{\tau-1}\times\prod_{j=1}^{l-1}(1-r_jd_j)^{2\tau}\times(1-(1-r_ld_l)^{2\tau})\right).$$

In order to calculate this quantity under the coalescent a standard limiting argument is followed in the next few stages. Setting $t=\frac{\tau}{2N}$ gives

$$\sum_{\tau=1}^{\infty}\left(\frac{1}{2N}\left(1-\frac{1}{2N}\right)^{2Nt-1}\times\prod_{j=1}^{l-1}(1-r_jd_j)^{4Nt}\times(1-(1-r_ld_l)^{4Nt})\right)$$

and setting $\rho_j=4Nr_j$ this becomes

$$\sum_{\tau=1}^{\infty}\left(\frac{1}{2N}\left(1-\frac{1}{2N}\right)^{2Nt-1}\times\prod_{j=1}^{l-1}(1-\rho_jd_j/4N)^{4Nt}\times(1-(1-\rho_ld_l/4N)^{4Nt})\right)$$

Noting that $\lim_{n\to\infty}(1+x/n)^n=e^x$ and letting $N\to\infty$ in conjunction with transforming our sum over τ into an integral over t gives

$$\begin{aligned}&\int_0^{\infty}e^{-t}\prod_{j=1}^{l-1}e^{-t\rho_jd_j}(1-e^{-t\rho_ld_l})dt\\ =\;&\int_0^{\infty}e^{-t}e^{-t\left(\sum_{j=1}^{l-1}\rho_jd_j\right)}(1-e^{-t\rho_ld_l})dt.\end{aligned}$$

(2.16)

Let L_{left} denote $\sum_{j=1}^{l-1} \rho_j d_j$ and L_{right} denote $\sum_{j=1}^{l} \rho_j d_j$ then this integral becomes:

$$\int_0^\infty e^{-t(1+L_{\text{left}})} \times \left(1 - e^{-t\rho_l d_l}\right) dt \tag{2.17}$$

$$= \int_0^\infty e^{-t(1+L_{\text{left}})} - e^{-t(1+L_{\text{right}})} dt \qquad \text{which is}$$

$$P(l) = \left(\frac{1}{1+L_{\text{left}}}\right) - \left(\frac{1}{1+L_{\text{right}}}\right). \tag{2.18}$$

Note that the integrand in (2.17) is $P(t)P(l \mid t)$, so

$$P(l \mid t) = e^{-tL_{\text{left}}} \times \left(1 - e^{-t\rho_l d_l}\right). \tag{2.19}$$

To calculate $P(\mathbf{m}, (j \to i))$, as defined in 2.13, for a particular recombination block note that:

$$P(\mathbf{m}, (j \to i)) = \int_0^\infty P(t)P((i,j) \mid t)P(\mathbf{m} \mid (j \to i), t)dt \tag{2.20}$$

where $P((i,j) \mid t)$ denotes the probability of a recombination in between sites i and $i+1$ given that the block started with site j.

Equation (2.20) is used for calculating the partial likelihoods at both boundaries and for any block in the middle of the data. Appropriate substi-

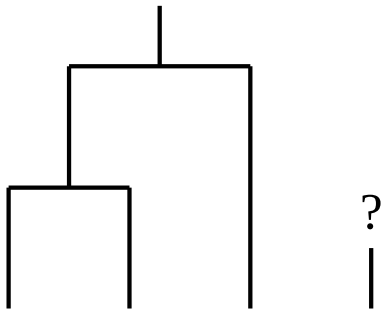

Figure 2.1: This figure a genealogical interpretation for the new sequence in the sample. Under the coalescent model there is an increased rate of coalescence (hence a shorter distribution to the tMRCA) when there are more sequences in the sample. The rate of coalescence of the new sequence to the rest of the tree is proportional to the number of extant sequences in the tree. This is approximated by the initial number of lineages, k, although the number of sequences in the rest of the tree changes through time with coalescence and recombination events on other sequences. I use this simple form to be consistent with the other schemes.

tutions need to be made for $P(t)$ and $P(l \mid t)$ in each case. At the right hand boundary $P(l \mid t)$ is a sum over all possible places for the right hand bound on where the recombination could have occurred.

I first consider a block starting at the left hand edge of the data, with a recombination between sites l and $l+1$. I model mutation as a Poisson process on the block and assume that all mutations lead to bi-allelic sites and that all are observed. Consider now k already observed haplotypes, I use ideas from coalescent theory (see Figure 2.1) to adjust the distribution of evolutionary time to

$$P(t) \approx ke^{-kt}.$$

Mutations at each site are independent and the mutations are assumed

to occur as a Poisson process in time. This gives the following expression for the mutation pattern $\mathbf{m}$ in a block from j to i:

$$P(\mathbf{m} \mid (i,j),t) = (t\theta)^m e^{-t\theta(i-j)}$$

where m is the number of mutations in the mutation pattern $\mathbf{m}$. Continuing the derivation for the specific situation where $j = 0$ and $i = l < L$ this becomes:

$$\begin{aligned} P(\mathbf{m},(0 \rightarrow l)) &= \int_0^\infty P(t)P(l \mid t)P(\mathbf{m} \mid l,t)dt \\ &= \int_0^\infty k e^{-kt} e^{-t(1+L_{\text{left}})} \left(1 - e^{-t\rho_l d_l}\right) (t\theta)^m e^{-t\theta l} dt \\ &= k\theta^m \int_0^\infty t^m e^{-t(k+L_{\text{left}}+\theta l)} dt - \int_0^\infty t^m e^{-t(k+L_{\text{right}}+\theta l)} \\ &= k\theta^m \left(\frac{1}{(k+L_{\text{left}}+\theta l)^{m+1}} - \frac{1}{(k+L_{\text{right}}+\theta l)^{m+1}} \right) \end{aligned} \quad (2.21)$$

The above formula deals with blocks that start at the left hand edge of the data and end with a recombination to the left of the end of the data. There are four scenarios in total for which this quantity must be calculated.

1. From the left hand edge of the data to a site before the right hand edge of the data.

2. From the left hand edge of the data to the right hand edge.

3. From a point to the right of the left edge of the data to the right hand edge.

4. Between two points in the middle of the data.

Note that in situations 3 and 4 the distribution of time between sequences is altered by conditioning on a recombination event between the SNP just before this block begins and the first SNP in the block. Let $P^l(R \mid t)$ be the probability of recombination conditioning on the time to the left of the recombination breakpoint being t and $P^r(R \mid t)$ be the probability of recombination given that the time to the right of the recombination is t. Also use $P^r(t \mid R)$ as the appropriate analogue for the reverse condition. We again assume that blocks separated by recombination are independent, and so the distribution of evolutionary time in one block is independent of the features of the previous. Thus by symmetry in the direction in which the data is read $P^l(R \mid t) = P^r(R \mid t)$ (following the reasoning in the derivation of π_R). Hence

$$\begin{aligned} P^r(t \mid R) &= \frac{P^r(R \mid t)P(t)}{P(R)} \\ &= \frac{P^l(R \mid t)P(t)}{P(R)} \\ &= \frac{\left(1 - e^{-\rho_j d_j t}\right) k e^{-kt}}{\rho_j d_j / (k + \rho_j d_j)} \end{aligned} \tag{2.22}$$

where $P(R)$ is calculated as $\int_{t'=0}^{\infty} P(R \mid t')P(t')dt'$.

It is finally necessary to calculate $P(l \mid t)$ when the right hand edge of the block is the right hand extreme of the data. Let Λ denote the position, in the full unobserved haplotype, where the next recombination event occurs. Then we can rewrite $P(l \mid t)$ as $P(\Lambda \geq l \mid t,)$. Now

$$\begin{aligned} P(\Lambda \geq l \mid t,) &= \sum_{a=0}^{\infty} P(l+a \mid t) \\ &= \sum_{a=0}^{\infty} e^{-t\left(\sum_{j=1}^{l+a-1} \rho_j d_j\right)} - e^{-t\left(\sum_{j=1}^{l+a} \rho_j d_j\right)} \\ &= e^{-t\left(\sum_{j=1}^{l-1} \rho_j d_j\right)}. \end{aligned}$$

Of course the d_j are not specified outside the range of the data, however these terms do not appear in the final result, so they can be chosen arbitrarily.

To save computational resources recombination events are only allowed at the midpoints between segregating sites and the mutation rates for each segregating site are then adjusted to represent to the number of sites in the interval between both midpoints. This then gives rise to the following final formulae:

Let $L_{i,j}$ denote $\sum_{k=j}^{i-1} \rho_k d_k$ and $\prod_{\mathbf{m}} \theta_m$ denote the product of the mutation rates across the intervals at each of the segregating sites in $\mathbf{m}$, this is analagous to θ^m in equation 2.21 but accounts for the distance between SNPs. Finally, for further ease of notation let

$$g(i,j,m,k) = \frac{1}{(k + L_{i,j} + \theta^*)^{m+1}}$$

where θ^* denotes the total mutation rate across the block. Then

$$P(\text{block}) = \begin{cases} n \prod_{\mathbf{m}} \theta_m \left(g(0,i,m,k) - g(0,i+1,m,k) \right) & i < L, j = 0 \\ n \prod_{\mathbf{m}} \theta_m \left(g(0,L,m,k) \right) & i = L, j = 0 \\ \frac{k(k+\rho)\prod_{\mathbf{m}} \theta_m}{\rho} \left(g(j,i,m,k) - g(j-1,i,m,k) \right) & i = L, j > 0 \\ \frac{k(k+\rho)\prod_{\mathbf{m}} \theta_m}{\rho} \left(g(j,i,m,k) - g(j-1,i,m,k) - \right. & \\ \left. g(j,i+1,m,k) + g(j-1,i+1,m,k) \right) & i < L, j > 0. \end{cases} \tag{2.23}$$

Having derived likelihoods for individual blocks it is then possible to use dynamic programming to calculate the probability of observing a haplotype h_{k+1} given a set $\{h_1, \ldots, h_k\}$ of previously observed haplotypes. Similar iterative formulae to those used in equations 2.3 and 2.6 can be employed. Let $X_{j,i}$ denote the haplotype h_{k+1} copies from between sites j and i. Also, let $\alpha_i = P(h_{k+1,\leq i} \mid R_i)$ where event R_i states that there is a recombination event at i. Finally, define $\gamma_{i,j}(x) = P(h_{k+1,(j,i)} \mid X_{j,i} = x, R_i)$. Then

$$\alpha_{i+1} = \sum_{j=0}^{i-1} \sum_{x \in \mathbf{h}} \gamma_{i,j}(x) \times \alpha_j \tag{2.24}$$

where $\alpha_0 = 1$. Then

$$P(h_{k+1} \mid h_1 \dots h_k) = \alpha_L \tag{2.25}$$

2.4 Results

Assessing the strengths and weaknesses of these schemes presents several technical difficulties. Firstly the relative performance of each model will depend on its application - I use ability to estimate a constant ρ as a measure of the model's ability to mimic the coalescent. This is because the Li and Stephens scheme was developed to estimate recombination rates. There are some disadvantages of using this method of assessment. Firstly it is impossible to calculate the true posterior distribution, or even the maximum likelihood estimate of ρ as large amounts of data are required to be at all informative. Such data set sizes are impossible to analyse using current full likelihood methods. Even for the data set sizes used here there is little information about ρ and the data may sometimes be consistent even with $\rho = \infty$. This means that estimators may have infinite variance and standard methods for summarising them may produce misleading results. Nevertheless I try to use straightforward analyses wherever possible - more complicated ap-

proaches also suffer from a number of difficulties and reduce the transparency of the study.

To assess the different PAC schemes I used the program *makesample* by Hudson [27]. This uses the coalescent with recombination to generate data sets under neutrality assuming, in this case, constant population size. I explore various values of the recombination rate, mutation rate and sample size. For each combination of parameters I generated 100 data sets. Likelihoods were calculated for a range of values of ρ under each scheme and then approximate maximum likelihood estimates (MLEs) of ρ were derived from these curves. As an initial summary I present the mean MLEs from each scheme for the different parameter values, see Table 2.1. For small values of ρ this table may well give an appropriate first summary of the performance of the methods. For higher ρ, such as when the data were simulated with $\rho = 1000$, there is a risk that the data contains much evidence for recombination and there is no evidence for correlations between sites. This could cause a true ML estimate of ρ to be extremely large, perhaps infinite. In practice the schemes seemed to underestimate ρ when the true value was very large and only in one case was the maximum tested value ($\rho = 10^8$) not outside the 2 log likelihood range for some of the schemes. However, this problem makes it difficult to draw certain conclusions about the mean estimate of ρ for data sets simulated with $\rho = 1,000$.

Scheme			$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
ρ	θ	n				
0	10	50	1.6	0.41	0.45	0.73
1	10	50	2.7	1.2	1.3	1.9
5	10	50	6.4	3.9	4.2	6.0
10	10	50	11	3.2	5.4	12
50	10	50	38	14	23	48
1000	10	50	468	169	309	784
5	10	5	15	12	11	30
5	10	10	5.3	2.5	2.5	6.9
5	10	20	5.5	3.4	3.3	5.9
5	10	30	6.4	3.8	3.7	6.3
5	10	50	6.4	3.9	4.2	6.0
5	10	100	6.4	1.7	2.7	6.1
5	10	200	5.9	1.6	2.4	5.0
5	2	50	4	2	4	4
5	3	50	3	2	3	3
5	5	50	4.8	2.9	3.8	4.3
5	10	50	6.4	3.9	4.3	6.0
5	15	50	7.9	2.0	3.1	7.6
10	15	50	13	4.0	6.5	15

Table 2.1: Mean ML estimates of ρ under the four schemes. For each set of parameter values 100 data sets were simulated and the schemes were run on a grid of values for ρ. In each case a single set of 20 orderings were used for all 4 schemes. The likelihoods were averaged over these 20 orderings. Some of the simulations gave rise to extremely high variability in estimates of ρ. For smaller values of θ the estimates must be treated with care. Also note that the reported values for $\rho = 1000$ do not include one data set where the likelihoods were extremely flat and all but the L^2 scheme peaked at $\rho \geq 10^8$. There is a noticeable correlation between the value of theta under which the data was simulated and the maximum likelihood estimates of ρ for all 4 schemes.

As well as the mean estimates of ρ it is also interesting to consider the variance of different estimators. As the estimators have significantly different means it is perhaps more consistent to use the coefficient of variation to measure this quantity (see Table 2.2 and Figure 2.2). A complication with measuring the variation in these estimators is that the data itself is not very informative for the underlying value of ρ. Although all of the data sets in a given collection were simulated under the same parameter values, the true maximum likelihood estimates of those parameters, in particular ρ, are probably highly variable between data sets (see eg. [28, 21]).

It is important not only to produce point estimates of parameters but also to give measures of uncertainty and a full picture of the posterior information about the parameter of interest. As a statistic of the peakedness of the likelihood curve I give the range of values of ρ for which the log likelihood is within units of the maximum. I calculated the 2 log likelihood intervals using a linear interpolation of the likelihood curves for the data (See Table 2.3). Figure 2.3 gives a a fuller and more visual presentation of the distribution of estimated likelihoods and the certainty associated for the data sets simulated with $\rho = 5$, $\theta = 10$ and $n = 50$.

Having calculated these approximate intervals and their average width it is interesting to note how often the mean lies within these intervals, the coverage. This is summarised in table 2.4.

Scheme			$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
ρ	θ	n				
0	10	50	1.14	1.86	1.78	1.68
1	10	50	0.714	0.979	1.06	0.927
5	10	50	0.487	0.630	0.647	0.581
10	10	50	0.324	0.444	0.401	0.373
50	10	50	0.249	0.258	0.251	0.285
1000	10	50	1.43	0.32	0.27	0.26
5	10	5	7	8	9	5
5	10	10	0.9	1.1	1.6	1.2
5	10	20	0.60	0.81	0.89	0.73
5	10	30	0.61	0.71	0.76	0.67
5	10	50	0.487	0.630	0.647	0.581
5	10	100	0.40	0.53	0.47	0.45
5	10	200	0.40	0.57	0.56	0.51
5	2	50	1.70	3.52	2.60	4.60
5	3	50	1.02	1.62	1.37	1.52
5	5	50	0.80	0.92	0.88	0.85
5	10	50	0.487	0.630	0.647	0.581
5	15	50	0.646	0.722	0.696	0.580
10	15	50	0.348	0.431	0.455	0.433

Table 2.2: Coefficient of variation of ML estimates of ρ under the four schemes. The coefficient of variation helps to adjust estimates of variability for changes in mean estimates. However this is not a perfect summary of variation as it can penalise methods which estimate small values of ρ when the true value of ρ is close to zero. It is necessary to place these estimates in the context of the corresponding mean MLE estimates.

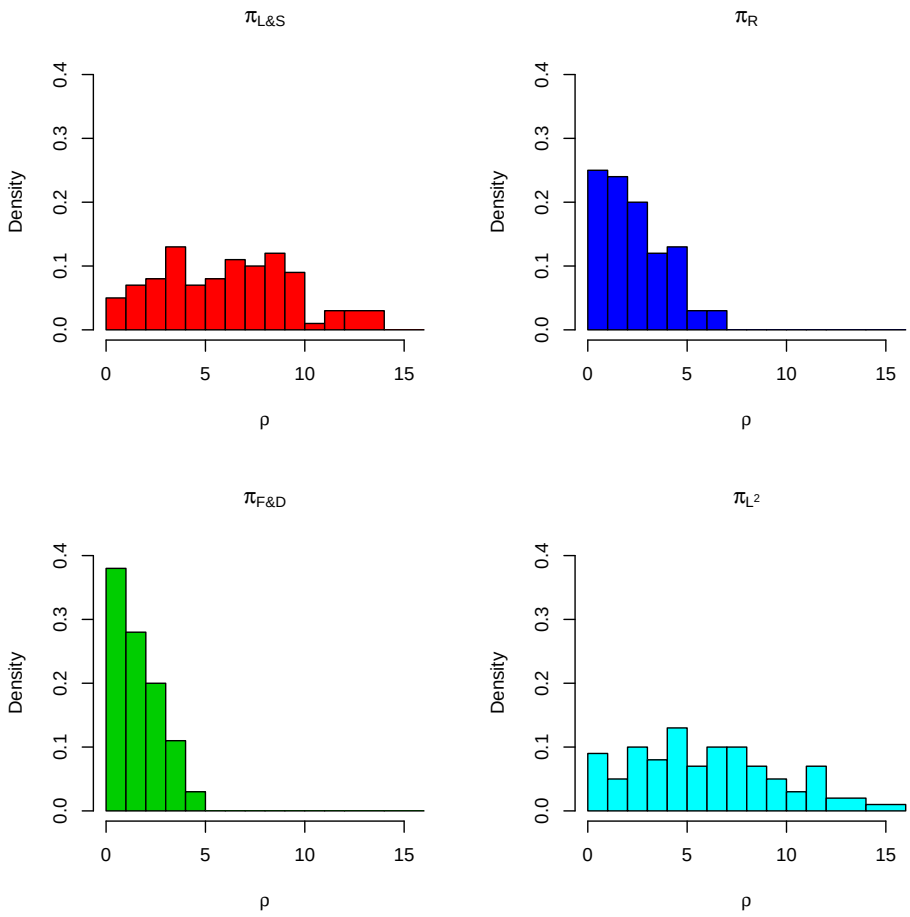

Figure 2.2: These histograms show the distribution of maximum likelihood estimates of ρ under the 4 different schemes. The true values for these simulations were $\rho = 5$, $\theta = 10$. There were 50 sequences in each sample. The results show that there is wide variation in estimates between data sets but also that there are strong systematic differences between the schemes. $\pi_{F\&D}$ and π_R almost never overestimate ρ whereas the estimates from $\pi_{L\&S}$ and π_{L^2} seem to be more evenly distributed around the true value.

Scheme			$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
ρ	θ	n				
0	10	50	(0.30, 5.63)	(0.0118, 3.15)	(0.0104, 2.99)	(0.0168, 5.37)
1	10	50	(0.789, 7.27)	(0.193, 4.83)	(0.202, 4.65)	(0.292, 7.09)
5	10	50	(2.74, 12.9)	(1.33, 9.56)	(1.39, 10.5)	(2.02, 14.9)
10	10	50	(5.79, 19.2)	(1.27, 6.77)	(2.33, 11.0)	(5.24, 24.2)
50	10	50	(19.2, 57.5)	(7.10, 22.3)	(11.4, 37.2)	(20.8, 76.0)
1000	10	50	(288, 10^6)	(111, 10^6)	(196, 10^6)	(440, 2558)
5	10	5	(9, 126)	(0.02, 195)	(0.02, 76)	(0.3, 221)
5	10	10	(1, 17)	(0.3, 24)	(0.3, 25)	(1, 48)
5	10	20	(1.8, 13)	(0.81, 10)	(0.76, 10)	(1.4, 19)
5	10	30	(2.5, 14)	(1.1, 10)	(1.1, 11)	(1.8, 17)
5	10	50	(2.74, 12.9)	(1.33, 9.56)	(1.39, 10.5)	(2.02, 14.9)
5	10	100	(3.09, 12.0)	(0.579, 3.95)	(1.08, 5.93)	(2.35, 13.1)
5	10	200	(2.89, 10.7)	(0.560, 3.53)	(0.97, 4.89)	(2.01, 10.5)
5	2	50	(0.3, 77)	(0.04, 83)	(0.1, 92)	(0.06, 108)
5	3	50	(0.4, 17)	(0.08, 16)	(0.2, 22)	(0.2, 28)
5	5	50	(1.37, 13.4)	(0.518, 10.9)	(0.773, 13.4)	(0.797, 15.8)
5	10	50	(2.74, 12.9)	(1.33, 9.56)	(1.39, 10.5)	(2.02, 14.9)
5	15	50	(3.89, 14.2)	(0.746, 4.47)	(1.29, 6.65)	(3.12, 16.1)
10	15	50	(7.79, 21.4)	(1.93, 7.46)	(3.25, 12.1)	(7.28, 26.9)

Table 2.3: This Table shows the average upper and lower bounds of intervals constructed by taking all values within 2 log likelihood units of the maximum likelihood value. Increasing the size of the data does provide less variable estimates of ρ but even with 200 sequences these schemes showed high variability in maximum likelihood values. The extremely large confidence intervals for $\rho = 100$ amongst the first three schemes were caused by a single data set where the maximum likelihood was finite but $\rho = 10^8$ was still within 2 log likelihood units of the maximum value. In general the width of these confidence intervals is disappointingly large as even the least variable estimates suggest that estimation is accurate to within a factor of 5, or worse.

Scheme			$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
ρ	θ	n				
0	10	50	62	93	89	91
1	10	50	70	99	92	93
5	10	50	83	89	83	91
10	10	50	93	4	60	95
50	10	50	70	0	7	96
1000	10	50	12	1.1	4.4	89
5	10	5	89	96	65	100
5	10	10	85	87	74	94
5	10	20	84	83	76	91
5	10	30	84	87	85	88
5	10	50	83	89	83	91
5	10	100	86	19	66	92
5	10	200	88	8.7	42	87
5	2	50	92	92	95	98
5	3	50	92	86	93	94
5	5	50	90	81	87	93
5	15	50	77	30	71	86
10	15	50	81	14	70	81

Table 2.4: Coverage of the four schemes, as a percentage, for 2 log likelihood intervals (given in Table 2.3). Although coverage is quite variable between the schemes there are a range of reasons for this. The estimates given by $\pi_{F\&D}$ and π_R are strongly biased downwards. While the scheme $\pi_{L\&S}$ is, in general, less biased the approximate 2 log likelihood intervals constructed in this way are narrower than those for the scheme π_{L^2}.

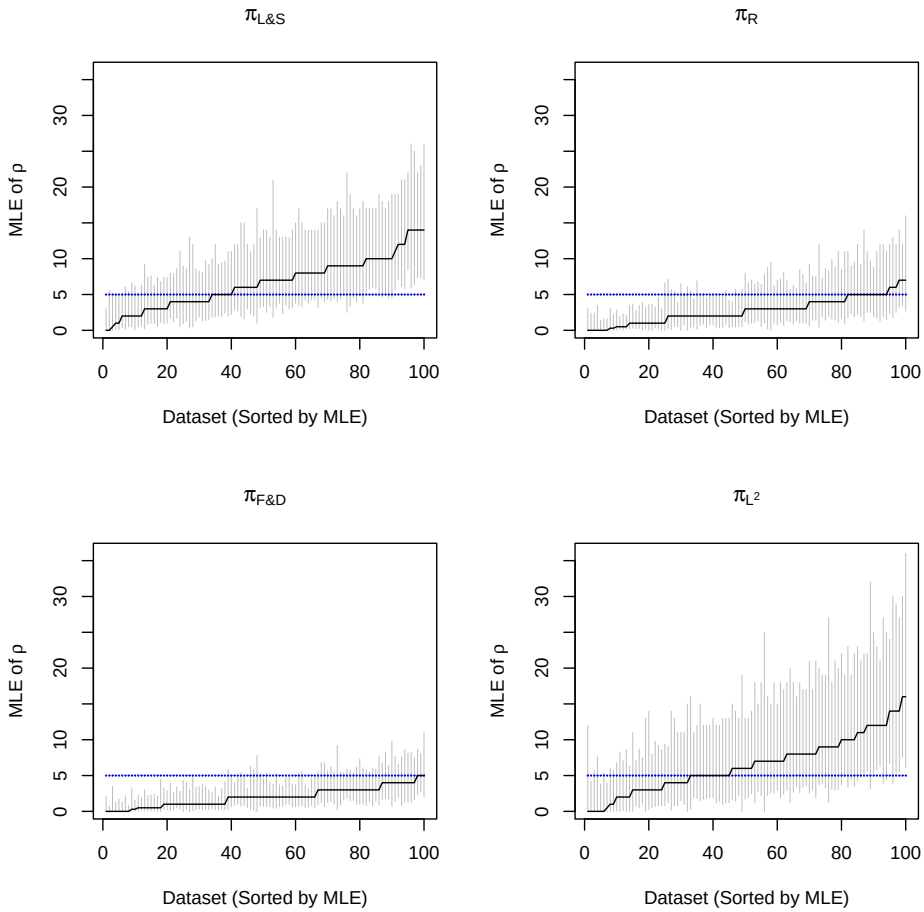

Figure 2.3: Maximum Likelihood Estimates of ρ for 100 data sets (sorted by MLE) with 2 log likelihood unit error bars. The data was simulated with 50 sequences $\theta = 10$ and the true value of $\rho = 5$, indicated by the dotted blue line.

The Variation in Estimates is Large

Having given a broad overview of the schemes it is worth taking a closer look at various aspects of these analyses. First of all it is worth noting that the numbers given here must be treated with care. Variation in estimates appears to be very high between data sets, even for 200 sequences and $\theta = 10$. Both the coefficient of variation in estimates is high and the widths of 2 log likelihood intervals are large. I have made some effort to reflect the accuracy of estimates by reducing the precision to which some estimates are given although I have decided to give more information if there is a chance that it is meaningful and always give at least integer precision unless using scientific notation. However, when assessment of the various methods is carried out it is important to take care to appreciate the accuracy of these estimates.

The Schemes always infer some Recombination

One clearly observable feature of all of the schemes examined here is bias. First of all I examine the tendency of the schemes to estimate $\rho > 0$ even when the true value is 0. It is hard, at first, to be certain that there is real evidence of model misspecification here however. Given lack of information in the data about ρ we would expect some variation and, as it is not possible to underestimate ρ when the true value is zero, a positive bias will always be

observed - even for full likelihood methods. As it is impractical to compare these results with an accurate full likelihood method other means of assessing whether these estimates are appropriate to the data are required. It is worth noting that under $\pi_{F\&D}$, π_R and π_{L^2} just over 1/3 of the MLEs were greater than 0 while for $\pi_{L\&S}$ more than 2/3 of estimates were greater than 0. Perhaps more telling is to look at the proportion of the time that $L(D \mid \rho = 0)$ was more than 2 units of log likelihood from the MLE. Using table 2.4) it can be seen that roughly 10% of the confidence intervals from $\pi_{F\&D}$, π_R and π_{L^2} do not include 0, over 30% of those created using $\pi_{L\&S}$ exclude $\rho = 0$. It is, unfortunately, hard to be certain how excessive this proportion is. However, note that the estimates $\pi_{F\&D}$ and π_R have a downwards bias when ρ is not close to zero. This means that their perhaps impressive performance in the case of $\rho = 0$ is less useful as it merely reflects this general tendency to underestimate ρ. Another subtlety with this analysis is that the 2 log likelihood intervals vary considerably in size. Considering coverage in Table 2.4 it seems that those confidence intervals using the Li and Stephens scheme are somewhat too narrow (the values are low, even when taking into account the fact that some estimates are not centered on the true value). The intervals from π_{L^2} are possibly too large. Taking this into account it still seems that π_{L^2} is less prone to estimating recombination rates far from zero.

Smaller sample sizes produce very poor estimates

It is important to know how the estimates change as various parameters of the investigation change. The main ancillary aspects that I have explored have been the sample size and the value of θ under which the data were simulated. The most noticeable effect of reducing the sample size is to reduce the information in the data to such a degree that reliable estimation becomes impossible. The analysis of data sets with 20 sequences or fewer can give at best an indication of the underlying parameters, and the variance of estimators is in practice infinite. A similar problem is found when θ is low. It is hard to say precisely at what point θ becomes too small for effective estimation because the number of segregating sites observed for any given θ is highly variable. One way to gain greater consistency in the information contained in the data would be to fix the number of segregating sites. However simulation with a fixed number of segregating sites does not correspond to a proper evolutionary model and may well lead to bias - especially when methods that make use of information about the evolutionary time between sequences are used. Another complicating factor is that the minor allele frequency at each site has a large effect on whether that site is useful for recombination rate inference. Many sites are largely uninformative as this frequency is too small, or even singleton.

The Bias of the schemes changes with ρ and θ

For those values explored it seems that the value of θ under which the data was simulated has a positive correlation with the estimates of ρ under all of these schemes. In Li and Stephens' paper [22] they reported a negative linear correlation between the log of the average distance between sites and the bias in ρ estimation, and a similar result may hold for all of the models. Also shown in their paper was the tendency for the method to overestimate when ρ was considerably less than 25 and underestimate when ρ was considerably more than 25 - when the data comprised of 50 sites. Although bias is clearly present in all of the algorithms it seems that π_{L^2} is less susceptible to being 'dragged towards' a particular value and is more sensitive to the true value under which the data were simulated.

There are Problems due to the Order Dependency of the Schemes

The PAC approach requires an arbitrary ordering to be assigned to the sequences in the sample. In order to circumvent this problem an average over twenty such orderings has been used to calculate the likelihood curves for this analysis. For each data set a single set of twenty orderings is used for each of the schemes and for each value of ρ; in this way it is hoped that the shape of likelihood curves will be retained and hence the ML estimates

will not be overly influenced by the particular (random) choice of orderings. I now investigate how effective this strategy is at removing the problem of order dependency from the estimates of ρ. In the following analysis I took a single data set, the first of those simulated under $\rho = 10$, $\theta = 10$ and with 50 sequences. I analyse the variation in the likelihoods estimated at the true value of ρ and also the variation in the maximum likelihood estimates achieved from different numbers of orderings. The first observation is that there is extreme variation in the likelihoods achieved from different orderings. In particular there are a small number of likelihoods which far exceed the values of the majority. Figure 2.4 shows that with a likelihood curve generated using 2,000 independent orderings only a small proportion of these make a significant contribution under any of the schemes. This seems to be most extreme in the case of $\pi_{L\&S}$ where the 5 biggest likelihoods appear to account for more than 80% of the height of the curve, so that the vast majority of the 2,000 orderings make almost no difference at all.

Perhaps a cleaner method of describing the variation in likelihoods is to examine the log likelihoods. Figure 2.5 shows the distribution of log likelihoods for the four schemes. At first glance these distributions look similar to a normal distribution. To check how well this distribution fits I used q-q plots in Figure 2.6. The fit is not very good for $\pi_{F\&D}$, π_R and π_{L^2} which show significantly less variation at upper tail than a normal distribution.

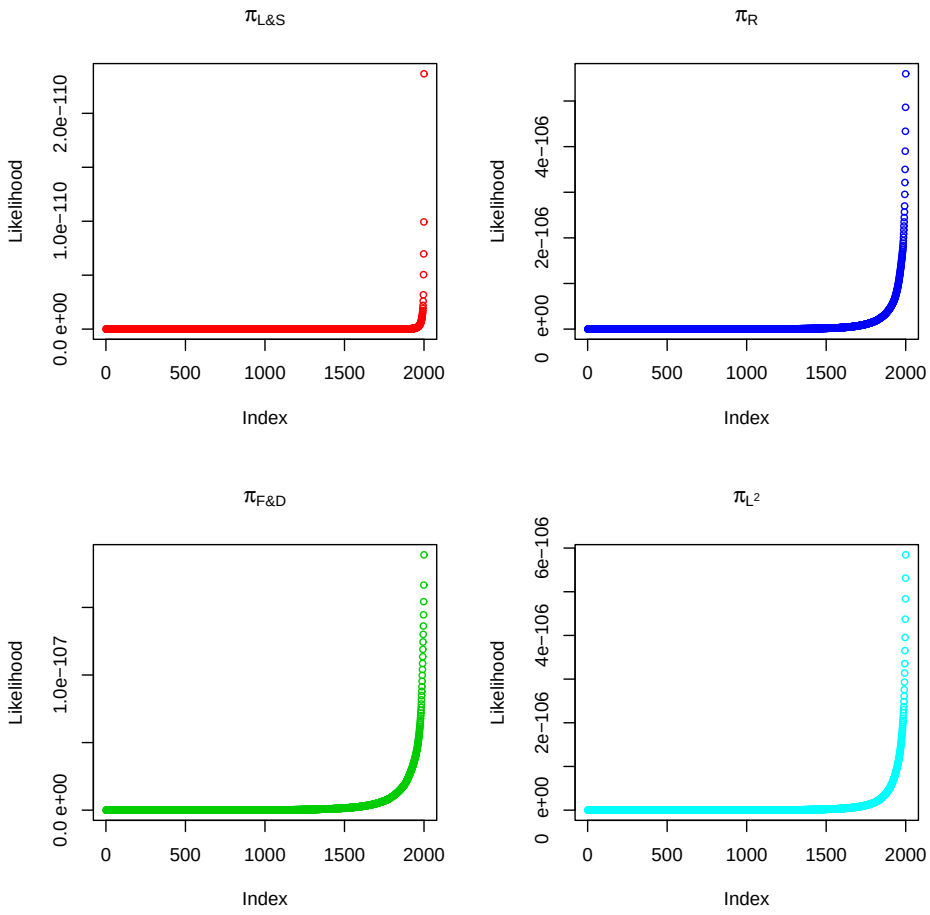

Figure 2.4: The cumulative contribution to the total likelihood sorted in ascending order when the likelihood is averaged over 2000 independent orderings. The same 2,000 orderings were used in all four schemes. Although all of the schemes suffer from extreme variation in likelihood estimates the plot for the scheme $\pi_{L\&S}$ seems to indicate that even for 2,000 orderings variation in likelihood estimates could be particularly significant.

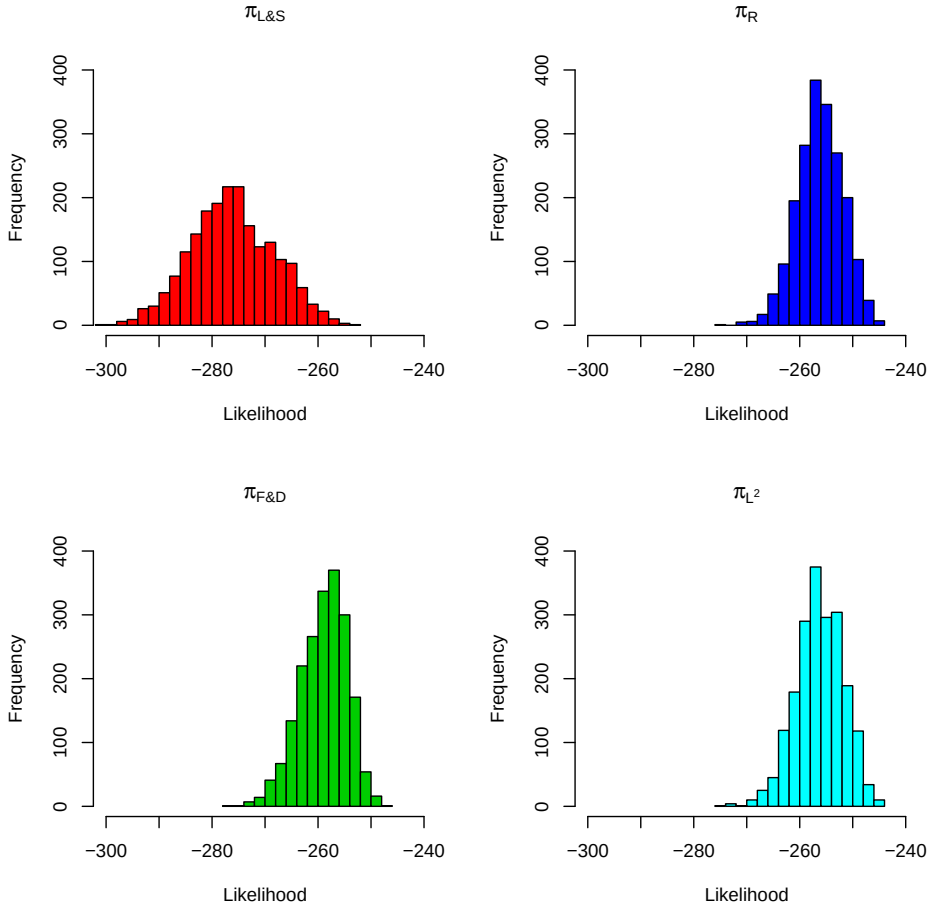

Figure 2.5: Histograms showing the distribution of the log likelihoods calculated for $\rho = 10$ for a single data set with 50 sequences.

This indicates that the variance due to order dependency for these schemes should be somewhat reduced as the likelihood will be effectively comprised of a larger number of samples from the space of all orderings. The means and variances of the log likelihoods shown in Figure 2.5 are given in Table 2.5.

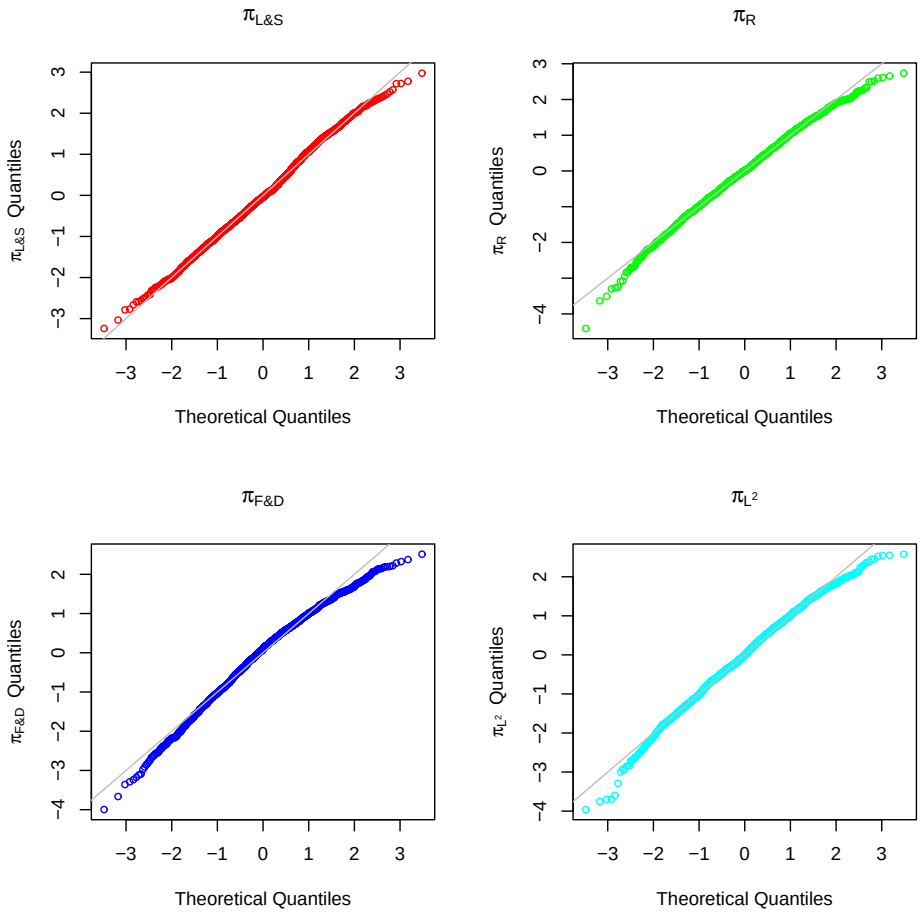

Figure 2.6: Normal Q-Q plots of the log likelihoods at $\rho = 10$ for a single data set each. The grey line is the line $y = x$. Each point is a log likelihood generated using one of 2,000 independent orderings of the data. The quantiles of the distribution of log likelihoods is compared to those of a normal distribution of the same mean and variance.

Scheme	$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
Mean	-276	-259	-256	-256
Variance	7.8	4.4	4.3	4.5

Table 2.5: The means and variances of the log likelihoods from Figure 2.5. There is greater variation in the likelihoods generated by $\pi_{L\&S}$ than under the other schemes. This means that the likelihood curves under $\pi_{L\&S}$ will be more greatly influenced by one large likelihood curve than under the other schemes. This means that the maximum likelihood value will be strongly influenced by one very likely ordering. Notice that, as shown in Table 2.2, the scheme does not show greater variation in likelihood estimates.

There is Significant Variation even in the ML estimates of ρ

It may be that estimation of the likelihood itself is not of primary importance. In this chapter the primary method of assessing these schemes is to estimate the level of recombination in the data. To asses variation in likelihoods I give the coefficient of variation in the ML estimates of ρ for a range of numbers of orderings averaged over to create the likelihood curves (Table 2.6). To give a visual understanding of the effect of this variation Figure 2.7 gives a histogram of the variation in all four schemes over 100 independent orderings of size 20. There is clearly substantial variation in the estimates of ρ for all of the schemes. Table 2.6 suggests also that this cannot be easily removed by increasing the number of orderings. However, increasing the number of orderings does have a significant effect on the mean ML estimate of ρ. Table 2.7 shows that for larger numbers of orderings the maximum likelihood estimate of ρ is likely to fall. This is likely to be because most

Scheme	$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
No of Orderings				
1	0.26	0.23	0.20	0.20
2	0.25	0.22	0.19	0.18
5	0.24	0.20	0.18	0.17
10	0.23	0.19	0.17	0.17
20	0.22	0.17	0.17	0.16
50	0.22	0.15	0.16	0.15

Table 2.6: Coefficient of Variation of Maximum Likelihood estimates of ρ using different numbers of orderings. 2000 independent Likelihoods were calculated for dataset 1, which was simulated using 50 sequences and $\rho = \theta = 10$. 500 subsamples of sizes between 1 and 50 were taken (with replacement) and likelihood curves generated from an arithmetic average of the individual values.

orderings infer spurious recombination events and only a very small number will avoid this (and those will have significantly higher likelihoods). This is explained in section 2.4.

A Geometric Likelihood Average can significantly reduce MLE Variation

When the ordering of sequences is viewed as missing data, analogous to a genealogy in the coalescent setting, it makes sense to take the arithmetic mean of the likelihoods over orderings. However, it is possible to introduce attractive properties for estimators using alternative approaches. For example the variation in ML estimates of ρ can be substantially reduced by taking a geometric average of the likelihoods from each ordering. This can be seen in Figure 2.8. The reason for this is that when the arithmetic average is taken

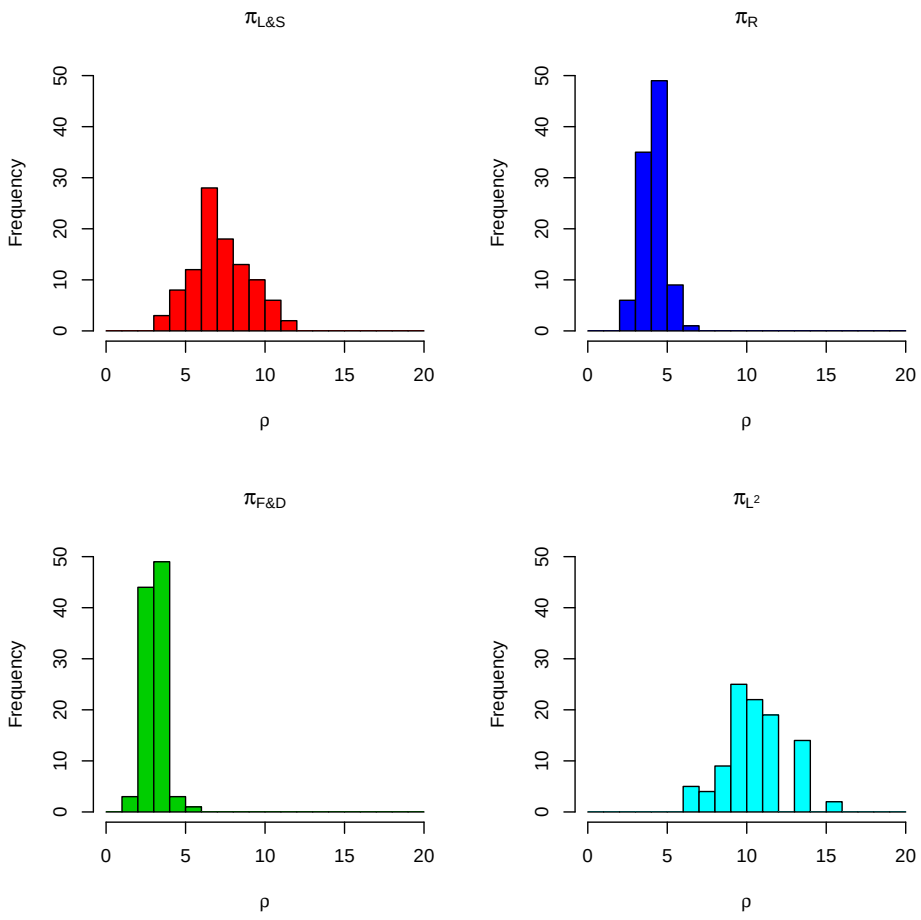

Figure 2.7: Variation in Maximum Likelihood estimates given 100 independent runs of 20 orderings on Data Set 1. These likelihood estimates were generated by arithmetically averaging the likelihoods from each ordering.

the likelihood curve is often overwhelmed by one ordering that resulted in a significantly higher likelihood than any of the others (see Figure 2.4). This weights the estimator heavily towards the maximum likelihood value of ρ for that ordering, and that increases the variance in the estimate. When geometric averages are taken this hugely flattens the difference in likelihoods for different orderings and so each ordering contributes a much more significant proportion of the total. This then leads to a significant reduction in the variance of the estimators. The coefficient of variation for geometrically averaged likelihoods can be seen in Table 2.8. It is also worth noting that the mean maximum likelihood estimators do not change significantly with higher numbers of orderings when geometric averages are taken (Table 2.9). This can probably be explained as most orderings produce lower likelihoods and higher estimates of ρ with some orderings producing significantly higher likelihoods and significantly lower estimates of ρ (as explained in section 2.4). The overall effect of this on arithmetically averaged likelihoods is that for higher numbers of orderings the maximum likelihood estimate of ρ decreases but the variance of the estimator decreases slowly. When geometric averaging is used the likelihood makes little difference and those orderings with low likelihood overwhelm the results from more likely orderings.

The reduced variation produced by taking the geometric mean of the likelihoods produced by different orderings of the samples motivates an analysis

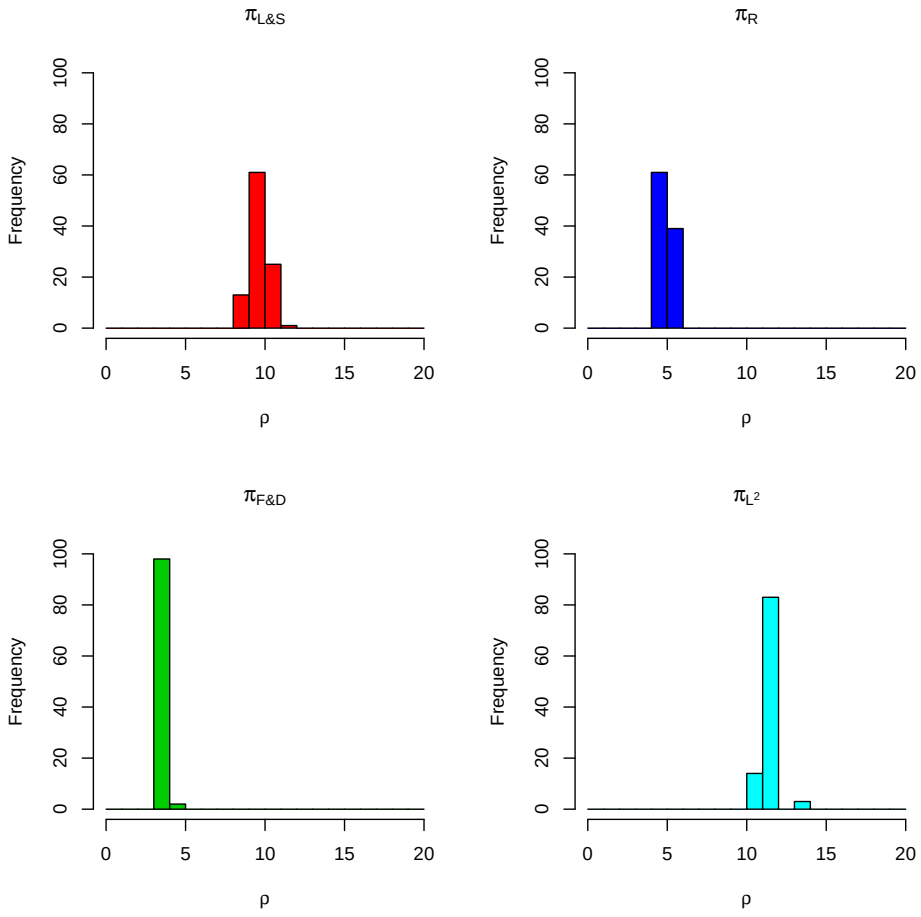

Figure 2.8: Variation in Maximum Likelihood estimates given 100 independent runs of 20 geometrically averaged orderings on Data Set 1.

Scheme	$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
No of Orderings				
1	10	4.1	5.5	12
2	9.7	4.0	5.3	12
5	8.9	3.9	5.0	11
10	8.7	3.7	4.9	11
20	7.8	3.5	4.7	11
50	7.3	3.4	4.6	11

Table 2.7: Mean Maximum Likelihood Estimates achieved using different numbers of orderings to calculate the likelihood surface (with arithmetic averaging of likelihoods) for data set 1.

Scheme	$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
No of Orderings				
1	0.26	0.23	0.20	0.20
2	0.19	0.17	0.15	0.14
5	0.12	0.11	0.10	0.096
10	0.087	0.076	0.092	0.070
20	0.064	0.037	0.091	0.046
50	0.043	0.0056	0.089	0.019

Table 2.8: Coefficient of Variation of Maximum Likelihood estimates of ρ for the first data set simulated under $\rho = \theta = 10$ using different numbers of (geometrically averaged) orderings.

Scheme	$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
No of Orderings				
1	10	4.1	5.5	12
2	10	4.1	5.5	12
5	10	4.1	5.5	12
10	10	4.1	5.4	12
20	10	4.0	5.4	12
50	10	4.0	5.3	12

Table 2.9: Mean Maximum Likelihood Estimates achieved using different numbers of orderings to calculate the likelihood surface (with geometric averaging of likelihoods). Values calculated for the first data set simulated under $\rho = \theta = 10$. Note that there is great variation in the estimates *between* data sets and that average estimates can be found in table 2.10

of the estimators produced in this way. Tables 2.10 - 2.13 show the basic properties of this estimator in the same format used for those estimates acquired using an arithmetic average.

Different methods of averaging over orderings provide the possibility to reduce the variance in estimates of ρ caused by the order dependency of the PAC likelihood. I have explored the possibility of using a geometric average of the likelihoods. This is highly effective at removing the variance in maximum likelihood estimates of ρ due to order dependency. The relative performance of the two averages is assessed through summaries of these estimates and there are reasons to be skeptical about taking a geometric average.

There is a notable upwards shift in ML estimates of ρ. This is because orderings which require more recombinations are often less likely and under an arithmetic ordering these have little impact on the overall estimate; under a geometric average this effect is vastly reduced. This shift in estimates results in a reduction in the coverage of the estimators and this indicates that these estimates are of lower quality. This effect is most severe when the data were simulated under small values of ρ and in these cases the estimates, especially under $\pi_{L\&S}$ are very high. Finally, if the PAC approach is taken as a model for evolution then the orderings represent missing data. In this case an arithmetic average approximates the sum over all of the missing data and is thus easily justified. It is much harder to provide a theoretical basis

Scheme			$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
ρ	θ	n				
0	10	50	3.43	0.36	0.49	1.19
1	10	50	4.7	0.90	1.1	2.8
5	10	50	8.6	2.5	3.2	7.8
10	10	50	13	4.3	5.9	14
50	10	50	42	17	25	56
1000	10	50	514	723	414	851
5	10	100	8.2	2.5	3.2	7.2
5	10	200	7.65	2.5	3.0	6.6
5	5	50	6.9	2.0	2.6	6.1
5	15	50	9.7	2.7	3.4	8.3

Table 2.10: Mean ML estimates of ρ when geometric averaging is used. These values are the means of those estimated over all 100 data sets for each of the parameter values shown.

for taking a geometric average over the sampled orderings.

The Scheme $\pi_{L\&S}$ is significantly faster than the alternatives

Methods which approximate the likelihood of recombinant population data have a wide range of potential applications. In some applications the computational efficiency of the algorithm will be of primary importance. Assessing the processing time required for these Hidden Markov Model based methods is straightforward as, for a given number of sequences and number of segregating sites, the same set of calculations is performed regardless of the data itself. Some features of the computational burden of each of these algorithms are theoretically straightforward: The Li and Stephens algorithm ($\pi_{L\&S}$) is linear in the number of segregating sites but quadratic in the num-

Scheme			$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
ρ	θ	n				
0	10	50	0.688	1.56	1.67	1.50
1	10	50	0.547	0.837	0.959	0.827
5	10	50	0.436	0.557	0.598	0.559
10	10	50	0.301	0.360	0.368	0.351
50	10	50	0.232	0.245	0.243	0.266
1000	10	50	1.30	6.88	2.35	0.27
5	10	100	0.31	0.40	0.45	0.39
5	10	200	0.318	0.382	0.397	0.400
5	5	50	0.74	0.93	0.89	0.85
5	15	50	0.286	0.436	0.467	0.423

Table 2.11: Coefficient of variation of ML estimates of ρ using geometric averaging

Scheme			$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
ρ	θ	n				
0	10	50	(0.82, 8.99)	(0.0171, 2.09)	(0.0317, 2.54)	(0.0709, 6.124)
1	10	50	(1.53, 10.9)	(0.136, 3.01)	(0.171, 3.79)	(0.460, 9.21)
5	10	50	(3.97, 16.6)	(0.87, 5.51)	(1.12, 7.63)	(2.74, 18.0)
10	10	50	(5.18, 20.8)	(1.21, 7.43)	(1.54, 10.2)	(4.02, 23.0)
50	10	50	(22.5, 61.4)	(11.0, 26.4)	(14.6, 39.2)	(33.1, 92.5)
1000	10	50	(316, 10^6)	(150, 10^6)	(201, 10^6)	(477, 2678)
5	10	100	(3.03, 13.1)	(0.508, 4.58)	(0.334, 6.03)	(1.53, 12.8)
5	10	200	(2.78, 12.5)	(0.588, 4.39)	(0.62, 5.42)	(1.33, 11.9)
5	5	50	(2.25, 17.6)	(0.39, 6.58)	(0.529, 8.78)	(1.29, 20.0)
5	15	50	(5.00, 17.1)	(0.343, 5.05)	(0.221, 6.59)	(3.47, 17.4)

Table 2.12: Width of 2 log likelihood intervals when geometric averaging used

Scheme			$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
ρ	θ	n				
0	10	50	27	92	90	86
1	10	50	48	95	94	83
5	10	50	69	60	75	83
10	10	50	90	16	69	93
50	10	50	82	0	11	95
1000	10	50	13	4	4	95
5	10	100	74	48	76	86
5	5	50	90	52	70	93
5	15	50	52	58	79	73

Table 2.13: Coverage of the various estimators, given by percentage.

ber of sequences. The same holds for the scheme of Fearnhead and Donnelly ($\pi_{F\&D}$) and our scheme π_R. However there are four times as many states under $\pi_{F\&D}$ and so it takes four times as long. Under π_R four calculations must be made for each of those under $\pi_{F\&D}$ so π_R roughly a factor of four slower again. The explicit block-wise approach π_{L^2} takes time quadratic in the number of segregating sites, hence its name. I used a Windows XP dual Pentium 4 (1.8GHz) PC to create the empirical times in Table 2.14. These give a summary of the time that each of the schemes takes to calculate the likelihood of a single data set for a selection of data sizes.

The Causes of Order Dependency and aspects of Bias

Unfortunately all of the models suffer from two major problems. Firstly, the likelihoods depend on an arbitrary ordering assigned to the haplotypes. Secondly the estimates of ρ obtained are biased. In order to further develop

	Scheme	$\pi_{L\&S}$	$\pi_{F\&D}$	π_R	π_{L^2}
n	Length				
5	5	0.005	0.0095	0.018	0.013
5	20	0.011	0.038	0.084	0.13
5	50	0.016	0.056	0.15	0.83
5	200	0.063	0.23	0.63	18
20	5	0.042	0.22	0.35	0.19
20	20	0.087	0.40	1.0	2.6
20	50	0.19	0.97	2.7	18
20	200	0.79	4.1	11	334
50	5	0.164	0.67	1.6	1.8
50	20	0.50	2.8	6.8	16
50	50	1.3	6.6	18	105
50	200	4.5	26	69	213
200	5	2.5	12	28	19
200	20	7.6	49	119	279
200	50	18	109	284	1771
200	200	65.9	400	1074	34964

Table 2.14: Time taken, in seconds, for each of the schemes to calculate a single likelihood using 20 orderings.

these models, it is important to understand how the approximations made lead to these symptoms.

I start by illustrating where the order dependency in haplotypes originates, although the likelihoods are comprised of a dynamic sum over all possible mosaics, much can be gained by considering the minimum number of events required to produce a specific configuration. Observe that in Figure 2.9 the ordering on the right requires a greater number of mutations or the same number of mutations and a recombination event. This ordering will then have a lower likelihood but one which improves as ρ increases, as the recombination event becomes more likely. Under the coalescent there is no ordering of sequences and the data set in Figure 2.9 requires only two events - two mutations. Figure 2.10 gives an example genealogy which achieves this, the tree gives rise to all three types with only two mutation events. Unfortunately the PAC model contains no information about tree structure and the parameters of these models cannot capture the subtleties that different topological situations can create. The lack of genealogical information in the PAC approach leads to the need to allow for many repeat or back 'mutations' at a single site. This leads to the unfortunate situation that the PAC scheme cannot recognise when, under the infinite sites model, the data are incompatible with a single tree, as in Figure 2.11 as such configurations can be explained by repeat mutation. Finally, there exist data sets which are

compatible with a tree but for which there is no ordering under the PAC approach which does not require either repeat mutation or recombination. A simple data set consisting of 3 sequences each with a singleton mutation has this property, under the PAC approach the data always requires either four mutations or three mutations and a recombination event. In contrast, under a genealogical model three mutation events would normally be the most likely explanation.

Another consequence of the PAC approach being unable to distinguish between situations such as in Figures 2.11 and 2.9 is that it makes the parameter θ, which mimics mutation by allowing the copying process to change types, crucial to the ML estimate of ρ. If θ is set very low then the copying process will require too many recombinations, as in Figure 2.9. However if θ is set too high then the PAC model will not infer recombination even in cases such as in Figure 2.11. This effect explains the strong downwards bias of the $\pi_{F\&D}$ and π_R schemes: examining the likelihoods from these schemes suggests that they effectively have a higher θ. Evidence for this is that likelihoods for data sets with many mutation events are higher under $\pi_{F\&D}$ and π_R than under $\pi_{L\&S}$ and π_{L^2}, however likelihoods for data sets where most sequences are identical at most of the observed sites are lower under $\pi_{F\&D}$ and π_R. The higher mutation rate leads to the situation that these schemes give more weight to repeat mutation events and less to recombination events,

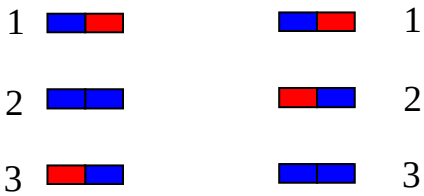

Figure 2.9: These represent two orderings of a simple data set. The colours represent allelic state. On the left hand side it is possible to obtain the second haplotype by copying the first sequence and employing a single mutation event. The third can then be obtained in the same way from the second haplotype. However, on the right hand side, the same data set in a different order, the second sequence requires two mutations when copying from the first but the third sequence is not identical to either of the first two and so either a recombination event or mutation event is required to produce it. When the recombination rate is not much smaller than the mutation rate this can lead to an inflated number of recombination events.

leading to a lower ML estimate of ρ.

Haplotypes are Unordered in the True Ancestral Process

There have been attempts to improve the method by somehow choosing 'better' orderings for the sequences (various private communications), although this work has not been published. Although such approaches may seem appealing at first there are a number of reasons to believe that they cannot provide us with a solution to the current problems with the PAC approach.

Firstly, consider large data, such that across the region there are likely to have been many recombination events on each sequence in the past. In this case, even if a natural ordering could be found in small regions of the data, the evolutionary relationships between the sequences will change drastically

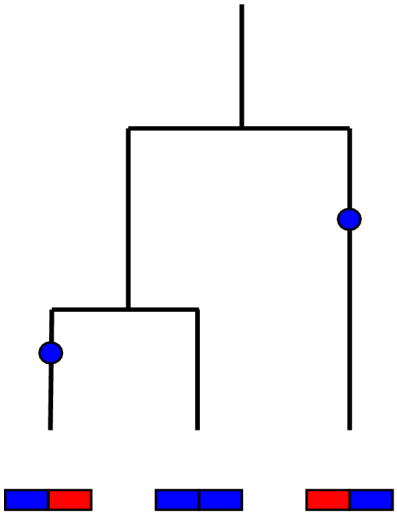

Figure 2.10: A simple coalescent history which is consistent with the data set in Figure 2.9. The blue discs represent mutation events to the blue type.

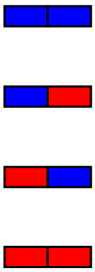

Figure 2.11: This simple data set depicts an *incompatibility* between two sites. A pair of sites is incompatible (with a tree) if, under the infinite sites assumption, it is impossible to construct a joint evolutionary history for the sites without at least one recombination event. In this case, if we were to trace these lineages back to a common ancestor, the next event backwards could not be be a coalescence event as no pair of sequences is identical. Under the infinite sites model the next event cannot be a mutation as there are no singleton alleles. Therefore the next event, backwards in time, in the history of this sample must be a recombination event.

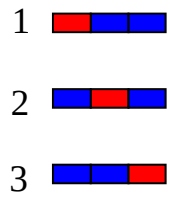

Figure 2.12: This data set is compatible with a tree under the infinite sites model, the data arises as the result of 3 mutations, one on each lineage before any coalescence events have occurred. Under the PAC model 4 events are required to explain this data. Two mutation events are needed to derive the second sequence from the first. Also either 2 further mutations or a mutation and a recombination are required to derive the third sequence from the first 2. This is independent of the ordering of sequences used.

across the data, so that any such relationship would be likely to break down over distance. Secondly there exist simple data sets for which all orderings give precisely the same results but in which recombination may be falsely inferred, an example is shown in Figure 2.12.

2.4.1 Discussion

The PAC approach has provided an extremely powerful new tool for analysing population genetic data in the presence of homologous recombination [29, 30, 23, 31, 32]. By approximating the ancestral process and imposing a Markov structure along the sequence, calculating the likelihood of large data sets has become computationally tractable. Unfortunately the lack of a genealogical structure causes certain configurations in the data to be incorrectly interpreted. This leads to an order dependency in the likelihoods and the problem

that estimates of parameters depend on which orderings of the sequences are used. Furthermore the schemes produce biased estimators and this bias is complex, depending on a range of factors. The schemes cannot distinguish incompatibilities from high frequency minor allele frequencies and this leads to the inability to accurately estimate ρ, especially when there are extreme amounts of recombination in the sample.

Of the alternative schemes proposed here all of them suffer similar problems to those of $\pi_{L\&S}$, also $\pi_{F\&D}$ and π_R seem to suffer from a strong downwards bias. However, despite this, these models may in fact be more accurate models of the evolutionary process as the full nature of the bias is not yet understood. The performance of π_{L^2} appears to be slightly better than that of the other schemes, it's performance seems to be most consistent across the range of values of ρ and it rarely rejects the true value of ρ more often than expected. Unfortunately this scheme is computationally more expensive, especially when there are many segregating sites. In a large scale analysis the data would have to be broken into parts to retain efficiency.

To significantly improve these schemes I believe it will be necessary to augment this approach with information about topological constraints. Either direct topological reconstructions or perhaps using incompatibility information, such as estimates of the minimum number of recombination events (eg. $R_{\min}$ [33], R_H [34]). However care must be taken to preserve the computa-

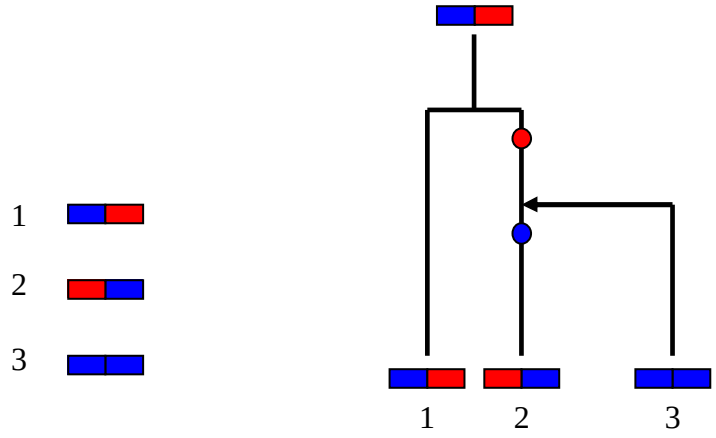

Figure 2.13: A diagram to explain a genealogical subtlety that affects the PAC approach - the 'data' shown is (the second ordering of) that in Figure 2.4. This diagram shows a genealogy constructed by first creating a coalescent tree for the first two sequences, and then augmenting this tree with the third sequence, as indicated by the arrow. The coloured discs on the genealogy indicate the branches and times of mutations *to* the type indicated by that colour. Given the genealogy it can be seen that although the third sequences has a different type to *all* previously sampled sequences no recombination or further mutation need be employed to explain its existence. There exist many algorithms that can propose approximate local genealogies given sequence data (eg. Neighbour joining techniques, or the method of Mailund (unpublished)). Perhaps it is possible to create a copying scheme whereby this information is used and repeat mutations/recombinations could be more definitively identified. The key to improving inference is in utilising the extra information given by the genealogies while maintaining the Markov structure and computational efficiency of the schemes proposed here.

tional efficiency that makes the PAC approach so attractive.

Chapter 3

A New Model for the Ancestry of a Sample

3.1 Introduction

The coalescent, developed by Kingman [1], models the ancestry of a sample of chromosomes from a large randomly mating population of fixed size. In 1983 Hudson [2] extended the model to include recombination and described an Ancestral Recombination Graph, or ARG, which represents the full genealogy of a sample of recombinant chromosomes. These models give rise to straightforward algorithms for generating genealogies backwards in time and hence also for simulating population genetic data. Wiuf and Hein [35] developed an alternative method for simulating coalescent genealogies which

recursively simulates successive coalescent genealogies, separated by single recombination events, moving left to right along a sequence.

While there are straightforward methods for simulating from the coalescent, inference under the coalescent is extremely challenging due to the complexity of the space of ARGs [19, 24]. Of crucial importance is the ability to calculate the likelihood of the data, unfortunately there are no known analytic expressions for calculating the likelihood even under the simplest mutation models. Full likelihood Monte Carlo methods under the coalescent, such as those developed by Fearnhead and Donnelly [21] and Kuhner et. al. [19] could potentially be used for inference under the coalescent with recombination, however, these methods are so computationally expensive as to be intractable for even moderate data sets. While inference under the PAC model (see Chapter 2)) is very fast, much of the biological realism of the coalescent process is lost. When used for recombination rate estimation, the model gives rise to systematic bias. In addition chromosomes are considered in a specific order, or set of orderings, and the likelihood is sensitive to this ordering. The PAC approach is not a genealogical method and so is also inappropriate for inference on the genealogies relating sample chromosomes.

In this chapter I investigate a new approximation to the coalescent: the Sequentially Markov Coalescent, or SMC. This model is designed to closely mimic the coalescent ancestral process. Under the SMC there are fewer

possible full ARG's but there is no change to the state space of marginal genealogies; the purpose of this is to provide a more efficient augmentation of the data that can be used to calculate the likelihood.

3.1.1 Understanding the Difficulties of Inference under the Coalescent

In developing approximations to the coalescent under which inference may be more tractable it is useful to appreciate the reasons why calculation of the likelihood under the coalescent is so difficult. Firstly, there is no known expression for the likelihood without knowledge of the underlying genealogy. It is therefore necessary to first augment the data with an underlying ARG and it is then possible to calculate the likelihood of the data conditional on that ARG. The unconditional likelihood can then be approximated as an average over a large sum of such conditional likelihoods. However, a direct implementation of this approach is not feasible using current computers because the number of samples required for such an estimate to be accurate is extremely large. This is partially because the state space of ARGs is huge and so a thorough exploration requires many simulations. There is no upper bound on the number of recombination events in the history of a sample and this leads to an infinite number of possible topologies. Also the contribution

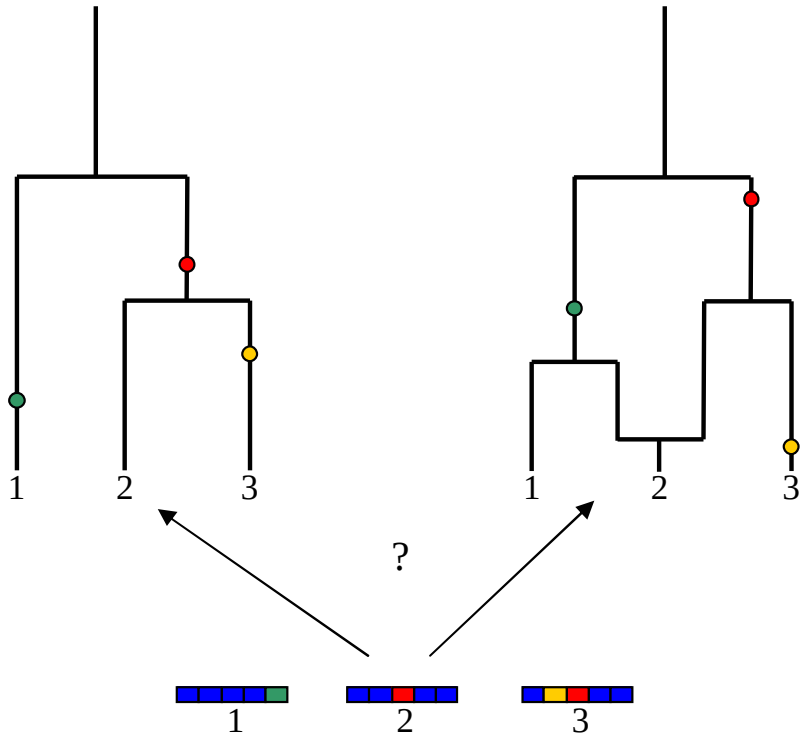

Figure 3.1: This diagram shows an example data set and, on the left a parsimonious coalescent genealogy that explains the data. However, under the coalescent with recombination there are alternative possible genealogies, even under the infinite sites model; one possible alternative is shown on the right. In this case the first event backwards in time is a recombination event which splits sequence 2 into two halves somewhere to the left of site 3. Infinitely many alternative ARGs are possible.

to the likelihood varies enormously between simulated genealogies.

Mutation events provide information about the underlying genealogies. Recombination events reduce our ability to identify the underlying genealogy by breaking up the sequence into small regions each with different genealogies. Unless the rate of mutation far exceeds the rate of recombination it is usually impossible to gain any certainty about the underlying genealogy from the

data. Even in the case of an infinite mutation rate it is impossible to detect all of the ancestral recombination events, and hence elucidate the full topology of the ARG (see Figure 3.1).

This problem can be viewed as the result of a highly redundant augmentation of the data. In calculating the likelihood of the data, given genealogical information, only the underlying marginal genealogies need to be specified. However, there is an infinite number of ARG topologies that correspond to any set of marginal genealogies and these ARGs usually have very different densities under the prior. This means that a thorough likelihood calculation must simulate many ARGs to approximate the likelihood of a single set of marginal genealogies. Unfortunately, performing direct inference on marginal genealogies is currently restricted as it is not yet possible to calculate the density of a set of marginal genealogies under the coalescent.

3.2 The Sequentially Markov Coalescent

I propose a new model of the ancestral process, the Sequentially Markov Coalescent (SMC), which closely approximates the coalescent but causes a reduction in the state space. In particular, the SMC reduces the number of redundant recombination events in each ARG. The model follows exactly the scheme described in section 1.6.1 in Chapter 1with a simple modification.

Using the same notation let $X_i(= \cup x_i)$ be the set of all points at which C_i has ancestral material, then the SMC is then defined by setting

$$I_{i,j} = \begin{cases} 1 & = \text{if } X_i \cap X_j \neq 0 \\ 0 & = \text{otherwise.} \end{cases}$$

That is, only pairs of lineages which both have ancestral material at at least one site can coalesce in this process. The rate of coalescence is the same for all such pairs regardless of the quantity of overlapping material.

The state space of ARG's under the SMC is significantly reduced: there are fewer possible coalescence events and the phenomenon of 'trapped non-ancestral material' does not exist in this setting because the union of two overlapping intervals (of ancestral material) is always itself an unbroken interval (see Figure 3.2). Trapped non-ancestral material increases the instantaneous rate of recombination in the ARG and so under the SMC the number of recombination events is also reduced ([36] contains the results of a simulation study to gauge this reduction in recombination events, see Figure 3.3). Note that recombination events in regions when part of a sequence has reached its MRCA, but is flanked by regions which have not, are still simulated. Despite the reduction in the state space of ARGs the state space of marginal genealogies remains the same.

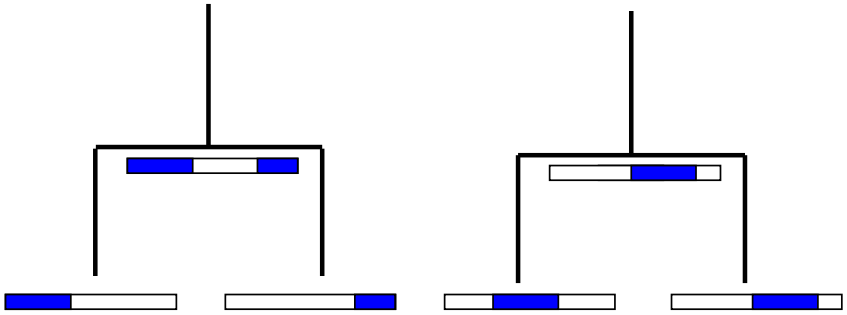

Figure 3.2: The coalescent phenomenon of trapped non ancestral material. Recombinations must be simulated within such material as such recombinations affect the genealogies of ancestral material. On the left two chromosomes coalesce to create trapped material. All coalescence events in the SMC are of the form on the right, where the intervals must overlap, hence trapped non-ancestral material does not occur under the SMC.

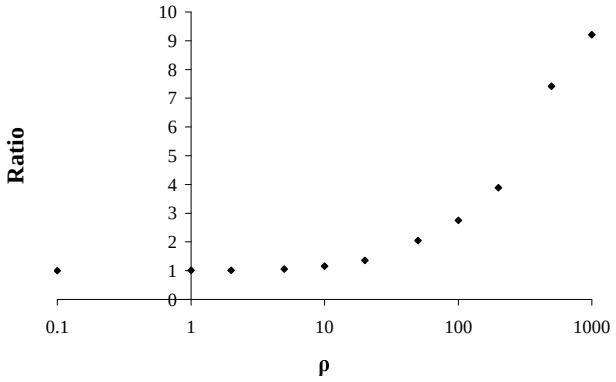

Figure 3.3: This graph shows the ratio of the expected number recombination events in the ARG under the coalescent to the expected number under the SMC for $n = 2$. The results were generated from 10^6 coalescent simulations under both models.

3.3 Simulating Recombinant Genealogies while moving along a Sequence

While the process described for simulating coalescent genealogies backwards in time has a simple Markovian structure, the spatial algorithm of Wiuf and Hein [35] has a complex, non Markovian, structure; the distribution of the next genealogy in the sequence depends on all previous genealogies. However, simulating genealogies under the SMC while moving along a sequence has a simpler, Markov, structure. In particular the following algorithm can be used to simulate genealogies for n individuals while moving along a sequence under the SMC.

Envisage a *continuous* sequence of unit length. Initiate the process by generating a tree from the coalescent at position 0. (This is equivalent to generating a genealogy according to the process in section 1.6.1 where $\rho = 0$). Denote the total branch length in tree i by T_i.

To generate the $i + 1^{\text{th}}$ genealogy from the i^{th}:

- Simulate $d_j \sim \exp(\rho T_i/2)$, the distance to the next recombination event.
- If $\sum_{j=0}^{i} d_j \geq 1$ then stop. Otherwise draw a point on the existing tree uniformly along the total branch length. Erase the portion of this

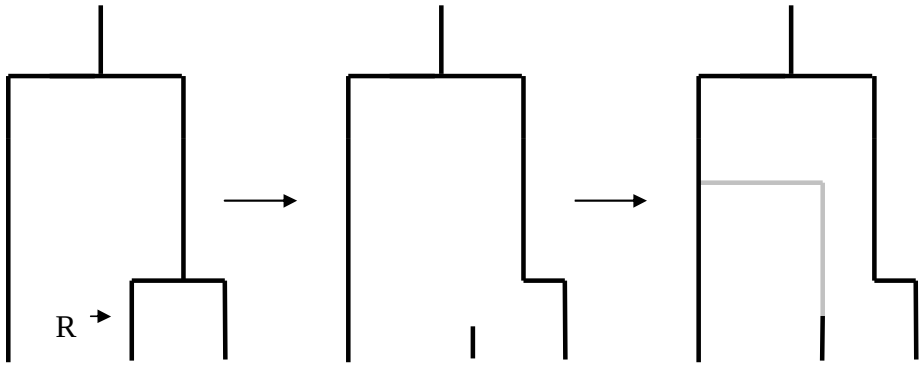

Figure 3.4: Simulating genealogies under the SMC while moving along a sequence. A recombination occurs on the first tree and the portion of the branch above the recombination is erased. This branch then coalesces onto the remaining branches.

branch that is further back in time and denote the remaining branch section by the term 'floating lineage'.

- The floating lineage is then traced backwards in time until it coalesces with one of the remaining branches in the tree. The rate of coalescence at a given time is equal to the number of branches remaining at that time. This time may be further back in time than the MRCA of any previous tree. Having generated the $i+1^{\text{th}}$ tree it is no longer necessary to keep track of the i^{th} tree under this process. Note that under the full coalescent process it is necessary to allow coalescence with a lineage from any of the previous genealogies, not just the most recent.

I now show that the distribution of marginal genealogies is the same under both the backwards in time process and the sequential algorithm. To

do this I first describe a slightly altered version of the backwards in time algorithm, the SMC*, in which recombination events are not simulated in 'trapped' material that has reached its MRCA. By this I mean material that has reached its common ancestor, but is flanked by material which has not. I claim that the distribution of marginal genealogies is the same under the SMC and under SMC*. The basic reasoning is as follows:

1. Suppose material that has reached its MRCA is present in the ARG, consider a single sequence, S, which contains such material. Denote the left edge of this material by T_1 and the right by T_2. On all sequences other than S there is only ancestral material at positions less than T_1 or at positions greater than T_2. This is true whether or not such material has also reached its MRCA.

2. This separation of ancestral material partitions the remaining sequences into disjoint sets α, β and S itself. No sequence in α can coalesce with a sequence in β.

3. The evolution of material at positions less than T_1 is now independent of material at positions greater than T_2. That is, these sequences cannot coalesce below their MRCA and events in one group do not affect the rate of events in the other.

4. This leads to independent marginal genealogies in these two regions

and recombinations in the trapped material between these two regions do not affect these genealogies.

Proof: Point 1 follows directly from the definition of the SMC, (see Figure 3.5). The SMC only allows coalescence events between lineages that share ancestral material, and by 1 it is possible to construct α and β in point 2. As these lineages cannot coalesce the rate of coalescence in one group is unaffected by coalescence in the other. Similarly for all recombinant sequences that are created by further recombination. It is possible for sequences to coalesce with sequence S, but this also does not affect coalescence (or recombination) rates within the subsets, so follows point 3. Point 4 follows directly as recombination events in the trapped material do not alter the rates of coalescence, and coalescence is the only process that affects the marginal genealogies.

I now prove that the simulation of genealogies along a sequence is equivalent to SMC*.

Proof: To prove the result I consider a single abstract ARG and show that the density function of this ARG is the same under both models. To begin the proof I make some straightforward observations:

1. For a given ARG, the tree generated by the sequential process at a particular locus is exactly the genealogy at that locus.

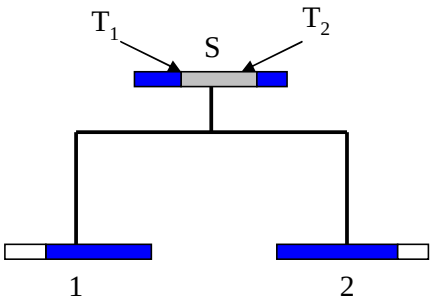

Figure 3.5: This diagram shows an example of material reaching its MRCA. The grey rectangles denote sites which have now reached their MRCA. Note that, as the MRCA is reached here, no other lineages in the sample can have ancestral material at these points. Also under the SMC it is impossible to create trapped non ancestral material, so there are also no lineages which have ancestral material both to the left of T_1 and to the right of T_2. Hence all lineages other than S contain ancestral material only on sites to the left of T_1 or to the right of T_2. Lineage S itself *is* ancestral between T_1 and T_2 and so there is no trapped *non ancestral* material in the ARG.

2. Under the backwards in time process the density of the ARG is the product of the densities in each epoch.

3. The densities calculated in the sequential algorithm can be calculated as the product of densities for each tree.

4. The densities in each tree in the sequential algorithm can be further split into a product of densities for each epoch.

By 3 it is then possible to collect all terms for a given epoch in the sequential algorithm for comparison with the backwards in time process. To prove the result it is sufficient to show that the densities in each epoch are the same by 2 and 4.

To prove this result it is first worth noting some basic figures. Given a coalescence event the probability that it is between any two specific branches is $\frac{1}{p}$, where p is the number of pairs that could coalesce in that epoch. Also, given that an event is a recombination, the probability that it occurs at a particular position, x say, on a particular sequence is $1/\mathbf{L}$, where $\mathbf{L}$ is the total amount of recombinant material in this epoch.

Also, the density function of an ARG (as it is a product of terms) can be separated into two parts, the exponential terms and their coefficients. However, as shown above the coefficients are trivial as, for every coalescence event the rate parameter $\lambda_C = p$ cancels with the probability of choosing that particular pair of sequences. Also, the probability of choosing a recombination at any one position cancels with the rate parameter $\lambda_R = \rho\mathbf{L}$ to give a factor of ρ. This means that the overall coefficient for a particular ARG is ρ^r where r is the total number of recombinations in the ARG. This is also true in the sequential algorithm (shown later).

Backwards in time Algorithm:

Consider an ARG with n leaves. Consider any epoch within this ARG and denote the height from the beginning to the end of this epoch (*not* the time in the tree in this epoch) by t_i.

Denote the rates of coalescence and recombination in this epoch by λ_C^i

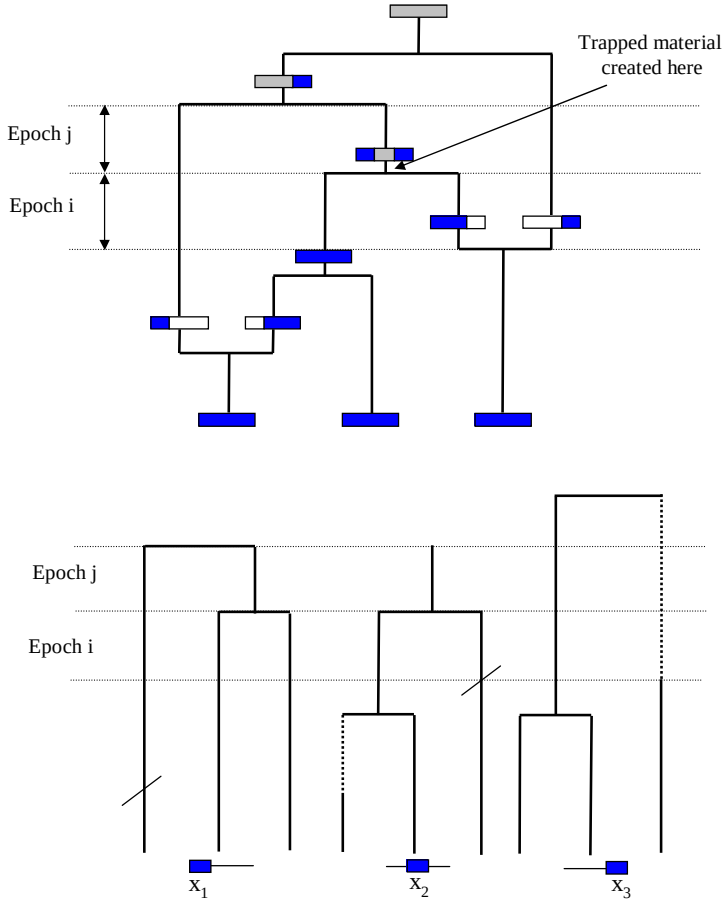

Figure 3.6: Simulating backwards in time (top) and sequentially (below). The strikes through branches represent where recombinations occurred on a tree. The dashed lines indicate the new lineages in each tree. Note that the stem above the marginal MRCA in the middle tree corresponds to material that has reached its MRCA. Recombinations are not simulated in this region under either process. Denote the 'width' of the i^{th} tree by x_i. In epoch i the total amount of recombinant material is, lineage by lineage: $[x_1 + (x_1 + x_2 + x_3) + (x_1 + x_2) + x_3]$ in the backwards in time process. In the spatial process (below) count the branches on each tree giving: $[3x_1 + 2x_2 + 2x_3]$. In epoch j the stem on the middle tree does not contribute to the branch length (this corresponds to not simulating recombinations in the trapped non ancestral material in the ARG above) hence both simulations give a rate of $[2x_1 + 2x_3]$.

and λ_R^i respectively then the backwards in time algorithm has density

$$\frac{\lambda_E^i}{\lambda_C + \lambda_R}.(\lambda_C + \lambda_R)e^{-t_i(\lambda_C+\lambda_R)} = \lambda_E^i e^{-t_i(\lambda_C+\lambda_R)} \tag{3.1}$$

where E is the event at the end of the epoch (at time $\sum_{j=0}^{i} t_j$).

More explicitly: assuming p pairs of sequences are able to coalesce in this epoch and the total amount of recombinant material is $\mathbf{L}$. The contribution from this event (of a coalescence between two particular sequences, or a recombination at a specific locus on a particular sequence) to the total density is then:

$$e^{-t_i(p+\rho\mathbf{L})} \qquad \text{if } E \text{ is a coalescence} \tag{3.2}$$

$$\rho e^{-t_i(p+\rho\mathbf{L})} \qquad \text{if } E \text{ is a recombination} \tag{3.3}$$

Sequential Algorithm:

In the sequential case we break up the terms for each epoch into coalescence and recombination terms. By definition the sequences able to coalesce in the backwards in time process are exactly those sequences which share at least one site at which both sequences have ancestral material. Consider two lineages from the ARG, in the sequential process these two lineages will be visible on a single marginal tree, and hence contribute towards the

overall rate of coalescence, if and only if they share a site on which both have ancestral material. Given that two lineages are visible on at least one tree in the sequential process there is some minimal k such that T_k is the first tree on which both are visible. If $k = 1$ then the in the construction of T_1 these lineages will contribute a factor of e^{-t_i} to the density in this epoch. If $k \geq 2$ then in T_k one of the lineages will have recombined in an earlier tree and will now coalesce back onto tree T_k. In this case also they will contribute a factor of exactly e^{-t_i}. This term appears only once as only lineages that have just recombined contribute towards the coalescence rate in a new tree. So, if the rate of coalescence, in a particular epoch, in the backwards in time process was p then in the spatial process exactly p pairs will have the opportunity to coalesce in the spatial process. Giving an overall rate of

$$1.e^{-t_i(p)}. \tag{3.4}$$

The recombination terms are calculated in a fundamentally different way under the sequential algorithm to the backwards in time algorithm. The distribution of the distance to the next recombination event is exponential of rate ρb_j, where b_j is the total branch length in tree T_j. Therefore the contribution to the density for each tree, T_j, given a recombination distance

x_j along the sequence from the last recombination point is:

$$\rho b_j e^{-\rho b_j x_j}.$$

Given that a recombination has occurred, the density of it occurring at any specific point in the ARG is $1/b_j$. So, the density of a recombination occurring at distance x_j from the last recombination point and at a particular point in the tree is:

$$\rho e^{-\rho b_j x_j}.$$

As I asserted earlier this gives a coefficient of ρ for every recombination. I now suppress this coefficient and consider only the exponential terms. Considering the epoch in question, and a particular tree, T_j, this becomes:

$$e^{-\rho(t_i \beta_j^i) x_j}$$

where β_j^i is the number of extant lineages in tree T_j and in epoch i, see Figure 3.6.

Taking the product of these for all trees T_j gives

$$e^{-\rho t_i (\sum_j \beta_j^i x_j)}. \tag{3.5}$$

The contribution in the last tree is the probability of no recombination in a tree, $P(x_j >$ the remaining length in the sequence); this is $e^{-\rho(t_i\beta^i_j)x_j}$. Note that the root branches are not included in the β^i_j. Informally it may be helpful to imagine that each of the branches in the marginal trees has a 'thickness' equal to the stretch of sequence along to which this marginal tree applies. Using Figure 3.6 as a guide it can be seen that the total amount of material on a branch in the backwards in time process is then the sum of thickness of all of the branches in the sequential process that represent this branch. More formally, by point 1 at the start of the proof, the total amount of recombinant material *per branch* in (3.5) is equal to the total amount of recombinant material extant on the same branch, in this epoch, in the ARG generated by the backwards in time process.

Hence the total amount of recombinant material, being the sum of the amount on each branch, is the total amount of material in this epoch on the ARG. Hence the exponential terms from considering recombination in this epoch are

$$e^{-t_i(\rho\mathbf{L})}. \tag{3.6}$$

Taking the product of the coalescence terms and the recombination terms we get:

$$e^{-t_i(p+\rho\mathbf{L})} \qquad \text{if } E \text{ is a coalescence}$$

$$\rho e^{-t_i(p+\rho\mathbf{L})} \qquad \text{if } E \text{ is a recombination}$$

exactly as in the backwards in time scheme. □ .

An analogous result holds for the corresponding Markov$_d$ process where, instead of marginal trees, ARGs representing the genealogy at the locus and the d sites preceding it are constructed recursively. The lineages created by recombination can then coalesce onto the ARG for this locus instead of the marginal tree as above. This is equivalent to a backwards in time process where lineages can only coalesce according to the following rule: L can coalesce with L' if and only if there exist loci i, j where L has ancestral material at i and L' has ancestral material at j with $\mid i - j \mid \leq d$.

The proof of this result is almost the same as the above. To see that the coalescence terms match consider any epoch and any pair of lineages which do not have material within d sites of each other. Then no two trees in the spatial algorithm will contain both of these lineages in this epoch so there can be no contribution to the coalescence rate associated with this pair. However any other pair of sequences will have a point at which both of them are contained in a marginal tree that the spatial algorithm considers. As in

the first result this then gives a rate contribution of exactly 1.

The recombination terms can be collected in the same way, however note that there is now also the possibility of trapped non ancestral material arising from coalescence events that do not cause any locus to reach its marginal MRCA. Continuous regions of this type of trapped material are restricted to stretches no longer than d sites. The contribution to the recombination rate in the backwards in time process by this trapped material corresponds to the extra time in the tree provided by the spatial algorithm keeping track of lineages up to d sites to the left of the current locus.

3.4 Comparison of the SMC and the Coalescent

In the previous section I defined a new model for the ancestral process under which coalescence events are restricted. In order to assess the value of the SMC it is necessary to compare the properties of it with those of the coalescent. It is useful to consider both the properties of genealogies generated under the SMC and the patterns of variation that these genealogies give rise to.

Due to the Markov nature of the coalescent processes backwards in time

under both the SMC and the full coalescent, the marginal distribution of genealogies for individual loci is unaffected by recombination events and is precisely that of the simple coalescent (in the absence of recombination). Therefore the only differences between the coalescent and the SMC lie in the correlations between genealogies along the sequence and the consequences of these correlations on data. To investigate the changes in these correlations brought about by the approximation proposed here I investigate 3 quantities. First I consider the covariance in the tMRCA across a range of genetic distances (measured in term of ρ). This partially addresses a natural concern that, as the SMC is Markovian along sequences, when simulating genealogies sequentially one would expect reduced correlation between marginal genealogies separated by more than one recombination event. Secondly I investigate the frequency of non Markovian patterns in the tMRCA along the sequence, this attempts to provide a measure of how often the genealogy partially reverts to previous configurations under the coalescent. Finally I investigate correlations between sites for data simulated under both the coalescent and the SMC.

Covariance in the tMRCA: Although it is not yet known how to calculate the joint distribution of the set of tMRCA's under the SMC analytically, accurate results can be obtained by simulation, as in Table 3.4. These results indicate a reasonable relationship between the SMC and the Coalescent co-

variances and the SMC values seem to lie within a factor of 3 of those from the coalescent.

ρ	Coalescent tMRCA	SMC tMRCA
0.1	1.066	1.025
1	0.599	0.458
2	0.377	0.239
5	0.152	0.068
10	0.059	0.023
20	0.022	0.007
50	0.0044	0.0027
100	0.0024	0.0013

Table 3.1: The covariance of the tMRCA at two sites on a chromosome under the Coalescent and the SMC with $n = 5$. The results here are an average over 10^6 coalescent and SMC simulations for each value of ρ. For intermediate and high recombination rates there is a significant reduction in the covariance in the tMRCA. This is because regions separated by significant recombination activity are less likely to coalesce under the SMC - reducing the correlations in coalescence time.

Non-Markovian Behaviour in the Coalescent: The second approach taken here is to try to quantify the scale of non-Markovian behaviour in the coalescent [36]. I start by defining a genealogy as *non Markovian* when the tMRCA of the sample at the loci 0 and 1 are equal but there is some locus, $0 \leq x \leq 1$, such that the tMRCA at x is different. Define $Q^*(n, \rho)$ as the proportion of non Markovian genealogies conditional on the tMRCA at sites 0 and 1 being equal, for a given sample size, n and recombination rate, ρ. Note that, as time is continuous, it is only possible for two loci to share the same tMRCA when the most recent common ancestor is reached in the

same coalescence event at both loci. A non-Markovian event is therefore impossible unless there is a coalescence event between lineages which have no overlapping ancestral material. Hence under the SMC $Q^*(n, \rho)$ is always 0.

ρ	$Q^*(2, \rho)$	$Q^*(3, \rho)$	$Q^*(5, \rho)$	$Q^*(10, \rho)$
0.1	0.0003	0.0006	0.0008	0.0009
1	0.023	0.041	0.057	0.0684
2	0.054	0.111	0.169	0.204
5	0.106	0.295	0.504	0.606
10	0.123	0.447	0.810	0.910
20	0.112	0.554	0.962	0.996
50	0.065	0.609	0.995	1.000

Table 3.2: The proportion of non-Markovian events conditional on the endpoints of the sequence having the same tMRCA. 10^7 coalescent runs were used for each value of ρ. The large difference between $n = 2$ and larger numbers of sequences is probably due to the fact that the majority of simulations which had shared coalescence times at the beginning and end of the sequences when $\rho = 50$ were simply very short genealogies and no recombination had occurred. This is much less likely for greater numbers of sequences. Note that the proportion of simulations in which the tMRCA was the same at both extremes of the sequences is very small (see Table 3.3) for high ρ hence such genealogies are rare as a proportion of all simulations.

The quantity $Q^*(n, \rho)$ cannot be calculated analytically under the coalescent so Monte Carlo simulations were used to generate Table 3.2. Note that care should be taken when comparing the effects for different numbers of sequences. Firstly, there may be non Markovian events in simulations involving more than 2 sequences but that do not affect the tMRCA. Secondly, for a given ρ per sequence, the total rate of recombination increases with

ρ	$P(2,\rho)$	$P(3,\rho)$	$P(5,\rho)$	$P(10,\rho)$
0.1	0.0003	0.922	0.913	0.908
1	0.023	0.508	0.462	0.444
2	0.054	0.309	0.261	0.243
5	0.106	0.114	0.0803	0.0724
10	0.123	0.0403	0.0245	0.0222
20	0.112	0.0119	0.0066	0.0062
50	0.065	0.0020	0.0011	0.0011

Table 3.3: The quantity $P(n,\rho)$ represents the probability of the same tMRCA at the left and right edges of the sequence given recombination rate ρ and with n sequences. To generate this table 10^7 Coalescent simulations were used for each value of ρ.

greater sample sizes as there is a greater quantity of ancestral material in the sample.

The results in Table 3.2 imply that non Markovian behaviour is a significant phenomena in coalescent genealogies. The effects seem to increase with the number of sequences and Q^* increases with ρ for more than 2 sequences. However, Q^* conditions on the end points of the sequences sharing a tMRCA, and the probability of this decreases with ρ. The highest proportion of non Markovian effects was observed for $n = 10$ and $\rho = 2$ when 5% of all simulations showed non this Markovian behaviour. It is likely that with more sequences this number could increase.

Correlation Patterns in Simulated Data: The above results attempt the gauge the effect of using the SMC model on simulated genealogies. It is also important to consider the effect on patterns of variation that generating

ρ	Coal mean r^2	Coal var r^2	SMC mean r^2	SMC var r^2
0.1	0.206	0.1201	0.204	0.1187
1	0.150	0.0759	0.145	0.0721
2	0.119	0.0531	0.114	0.0501
5	0.081	0.0281	0.079	0.0268
10	0.059	0.0165	0.058	0.0161
20	0.044	0.0101	0.043	0.0099
50	0.032	0.0058	0.031	0.0057
100	0.027	0.0043	0.026	0.0043

Table 3.4: The mean and variance of the r^2 statistic under the coalescent and the SMC. The results here are an average over 10^6 coalescent and SMC simulations tracing the ancestry of 50 chromosomes for each value of ρ

ρ	Coal mean D'	Coal var D'	SMC mean D'	SMC var D'
0.1	0.410	0.2187	0.409	0.2183
1	0.375	0.2045	0.372	0.2037
2	0.353	0.1949	0.350	0.1944
5	0.321	0.1800	0.319	0.1797
10	0.296	0.1678	0.296	0.1682
20	0.274	0.1565	0.273	0.1565
50	0.249	0.1432	0.248	0.1428
100	0.235	0.1355	0.236	0.1360

Table 3.5: The mean and variance of the D' statistic under the coalescent and the SMC. The results here are an average over 10^6 coalescent and SMC simulations tracing the ancestry of 50 chromosomes for each value of ρ

genealogies under the SMC has. Note that the prior distribution of marginal genealogies at any given locus is identical under the coalescent to under the SMC. The distribution, in both cases, is exactly that the distribution under coalescent process for a chromosome with recombination rate 0. It is therefore only necessary to consider the correlation structure between sites when assessing the effect of the approximation on data. The co-inheritance

of alleles in a population is referred to as *linkage disequilibrium* and measures of linkage disequilibrium, such as r^2 or D' [37], are designed to quantify the pairwise correlation structure of observed alleles. These distribution of these quantities cannot be derived analytically under the coalescent but simulation results can be found in Tables 3.4 and 3.5.

The distributions of both r^2 and D' under coalescent and under the SMC are very similar which encourages the belief that inferences made under the SMC may reflect well the conclusions that would be drawn under the full coalescent.

3.4.1 Discussion

The coalescent with recombination is a realistic model for the ancestry of a random sample of chromosomes in a freely mating neutral population. However, it is intractable to calculate the full likelihood for modern data sets with current computing technology under the coalescent. I have proposed an alternative model which reduces the ancestral state space and gives rise to a simple Markovian structure when simulating genealogies along a sequence.

For a model to be useful for inference it is necessary that the model provides a good approximation to the biological processes being explored. In order to assess the validity of the SMC I have investigated the relationship of

the SMC to the coalescent, both in terms of the distribution of genealogies and the resulting patterns of variation produced.

These results suggest that the SMC provides a good approximation to the coalescent and hence, if inference under the SMC proves to be more tractable, it could provide a sound alternative to the coalescent for understanding population genetic data. In the next chapter I discuss the possible gains in efficiency that could be made when performing inference under the SMC.

Chapter 4

Using the SMC for Importance Sampling

4.1 Introduction

In the previous Chapter I defined a new model for the genealogy of a sample of genetic data, the Sequentially Markov Coalescent (SMC). In this Chapter I investigate the potential for the SMC as a tool for inference, in particular the relative efficiency of importance sampling under the SMC and the coalescent. I calculate the likelihoods of one hundred data sets under both the coalescent and the SMC given a range of values of ρ. To assess the value of the SMC in this context I examine how closely the SMC likelihoods approximate those of the coalescent and if there is any reduction in the computational burden

of calculating these likelihoods.

4.2 Importance Sampling

In this Chapter I introduce a genealogical importance sampler and use it to perform inference on a collection of simulated data sets. The optimal importance sampler generates samples according to the target distribution - that is proposes genealogies according to their distribution conditional on the data. In 2000, Stephens and Donnelly [20] proposed an importance sampler which approximated this conditional distribution in the case of no recombination. In 2001 this method was extended by Fearnhead and Donnelly [21] to include recombination. I follow the method of Fearnhead and Donnelly in constructing an importance sampler to calculate the likelihood curves under both the SMC and the Coalescent. The method generates a coalescent genealogy backwards in time until the most recent common ancestor is reached at all sites. The genealogy is constructed using the sample and is augmented with mutations so that the probability of the sample configuration given the augmented genealogy is 1 or 0.

I use the infinite sites model in my implementation of the importance sampler. Extending this to a more general mutation model is straightforward. However the infinite sites assumption provides a significant restriction on

the genealogies that can be generated and this reduction in the state space reduces the computational burden of calculating the likelihood. Although the infinite sites assumption will, in general, not hold it has previously been used for inference in the presence of recombination (see eg. [17, 27, 38, 39, 40]). Care must be taken in assessing whether or not this model is appropriate to the organism and type of data in question. For example, given human SNP data, the infinite sites assumption may be realistic - repeat or back mutations will have occurred on only a small fraction of the sites in the data. Note that in practice the importance sampler considers only those sites segregating in the population so that each sequence consists of finitely many (segregating) sites. This does not violate the assumption that the data were generated from sequences with infinitely many sites.

The derivation of the importance sampling scheme that I give here is different in motivation to those given in the population genetics literature so far [17, 21]. In their paper Fearnhead and Donnelly consider generating an ARG backwards in time and say that rates of events *backwards* in time are unknown for an optimal importance sampler. However they note that the *forwards* transition rates are known and use Bayes rule to calculate the backwards transition rates. I consider it more natural to consider only backwards transition rates under the coalescent process and do not define any forwards process for generating ARGs. It is known how to simulate genealogies, back-

wards in time, under the coalescent with recombination in the absence of data. However it is not known how to simulate such genealogies conditional on data. The optimality of generating genealogies conditional on the data is derived below.

Formally, I wish to calculate the probability of observing a particular sample configuration, or data, D, and a set of parameters, ρ, θ: $P(D \mid \rho, \theta)$. I calculate this quantity under the coalescent model and the SMC and use exactly the same methodology to calculate in each case. For the sake of simplicity I now use the shorthand $P(D)$ to denote the probability of the data, given the parameters ρ and θ under the relevant model. Importance sampling can be used to calculate this quantity (see eg [41]) through the approximation

$$\begin{aligned} P(D) &= \int_{t=0}^{\infty} P(D \mid G) P(G) dG \\ &= \int_{t=0}^{\infty} P(D \mid G) P(G) \times \frac{1}{Q(G)} \ (Q(G) dG) \qquad (4.1) \\ &\approx \frac{1}{M} \sum_{i=0}^{M} \frac{P(G_i)}{Q(G_i)} P(D \mid G_i) \qquad (4.2) \end{aligned}$$

where the G_i represent genealogies generated from the *proposal distribution* $Q(.)$. The optimal choice of $Q(.)$, $Q^*(.)$, proposes genealogies according to the conditional distribution of genealogies given the data [20]. In fact, given

$Q^*(.)$ defined by $Q^*(G) = P(G \mid D)$ all genealogies generate exactly the same importance weight, $P(D)$. A simple proof of this follows: Consider one sample genealogy from the distribution $Q^*(.)$, G_1. Then

$$\begin{aligned}
\sum_i \frac{P(G_i)}{Q(G_i)} P(D \mid G_i) &= \frac{P(G_1)}{Q(G_1)} P(D \mid G_1) \\
&= \frac{P(G_1)}{P(G_1 \mid D)} P(D \mid G_1) \\
&= \frac{P(G_1)P(D)}{P(G_1, D)} \frac{P(D, G_1)}{P(G_1)} \\
&= P(D).
\end{aligned} \tag{4.3}$$

Hence under $Q^*(.)$ only one genealogy need be simulated to get a perfect estimate of the probability of the sample, D. I refer to $P(G \mid D)$ as the *target distribution*. Note that the purpose of the SMC is not as a proposal distribution for genealogies (indeed the support of SMC genealogies does not contain the support of coalescent genealogies). Instead the proposal distribution is a further approximation and when the SMC is employed the target distribution is $P(G \mid D)$ under the SMC model.

It is very hard to calculate the conditional density of a genealogy given the data in one step. Fortunately, the coalescent process has a Markov structure when genealogies are simulated backwards in time. This enables sim-

ulation of genealogies conditional on the data by, at each point in time, simulating the next event backwards in time according to $P(E \mid D)$, the distribution of events conditional on the data. This is because a genealogy, G, can be viewed as a sequence of events, $(E_1, \ldots, E_k)$, that act on the sample configuration at certain times in the past, $(T_1, \ldots, T_k)$. That is $G = \{(E_1, \ldots, E_k), (T_1, \ldots, T_k)\}$. However the event sequence is sufficient to define the sample configuration and, given the event sequence, there is no information about the times at which the events occurred contained in the data, so only the event sequence is changed by importance sampling. Now

$$\begin{aligned} P(E_1, \ldots, E_k \mid D) &= P(E_1 \mid D) \times P(E_2, \ldots, E_k \mid E_1, D) \\ &= P(E_1 \mid D) \times \prod_{i=2}^{k} P(E_i \mid (E_1, \ldots, E_{i-1}), D) \\ &= P(E_1 \mid D) \times \prod_{i=2}^{k} P(E_i \mid D_{(E_1,\ldots,E_{i-1})}) \end{aligned} \tag{4.4}$$

where $D_{E_1,\ldots,E_{i-1}}$ denotes the transformed data after events $E_1, \ldots, E_{i-1}$ have acted on it. This final equality holds as the Markov property ensures that the distribution of the next event backwards in time is independent of all previous events in the genealogy. Hence the probability of the data, given the next event backwards in time, $P(D \mid E)$, is equivalent to the probability

of the data as transformed by event E, $P(D_E)$. That is

$$P(D \mid E) = P(D_E). \tag{4.5}$$

It is therefore possible to calculate the likelihood of the data by simulating a sequence of events, each conditional on their distribution given the set of haplotypes as altered by the previous events simulated. Genealogies can then be generated iteratively according to:

$$\begin{aligned} P(E \mid D) &= \frac{P(D \mid E)P(E)}{P(D)} \\ &= \frac{P(D_E)P(E)}{P(D)}. \end{aligned} \tag{4.6}$$

To calculate the quantities $P(D)$ or $P(D_E)$ it is necessary to simplify Equation 4.6, this is done in turn for each of the three event types: coalescence, recombination and mutation. Let $D = \{h_1, \ldots, h_n\}$ and first consider a coalescence event between (without loss of generality) haplotypes h_1 and h_2 in the sample at the current epoch in the simulation. Then

$$P(D) = P(h_1, \ldots, h_n) = P(h_1, h_2 \mid h_3, \ldots, h_n)P(h_3, \ldots, h_n)$$

and

$$P(D_E) = P(h_C \mid h_3, \ldots, h_n)P(h_3, \ldots, h_n)$$

where h_C is the haplotype created from the coalescence of h_1 and h_2 (note that this need not be equivalent to either h_1 or h_2). It is then possible to write

$$\begin{aligned} \frac{P(D_E)P(E)}{P(D)} &= \frac{P(h_C \mid h_3, \ldots, h_n)P(E)}{P(h_1, h_2 \mid h_3, \ldots, h_n)} \\ &= \frac{P(h_C \mid h_3, \ldots, h_n)P(E)}{P(h_1 \mid h_2, h_3, \ldots, h_n) \times P(h_2 \mid h_3, \ldots, h_n)}. \end{aligned} \tag{4.7}$$

Now suppose that E is a mutation event on haplotype h_1 creating haplotype h_M. Then

$$\frac{P(D_E)P(E)}{P(D)} = \frac{P(h_M \mid h_2, \ldots, h_n)P(E)}{P(h_1 \mid h_2, \ldots, h_n)}. \tag{4.8}$$

Consider a recombination on haplotype h_1 event between sites j and $j+1$. Then let $h_{1,<j}$ denote the haplotype identical to h_1 at site j and all sites to the left of j, $h_{1,<j}$ is given non ancestral sites (which have no allelic state) to the right of site j. Define $h_{1,>j}$ as the complementary sequence: non ancestral at site j and those to the left of j and taking the values of h_1 at site $j+1$

and those to the right of $j+1$. Then

$$\frac{P(D_E)P(E)}{P(D)} = \frac{P(h_{1,<j} \mid h_2, \ldots, h_n)P(h_{1,>j} \mid h_{1,<j}, h_2, \ldots, h_n)P(E)}{P(h_1 \mid h_2, \ldots, h_n)}. \quad (4.9)$$

The complete collection of weights must be normalised as the relative likelihoods calculated are approximations.

Under the coalescent even the quantities above cannot be calculated efficiently, so approximations must be used. In their paper Fearnhead and Donnelly used a Hidden Markov model for the data with a finite sites model. This model is the progenitor of the models used in Chapter 2 and is the scheme on which $\pi_{F\&D}$ from that Chapter is closely based. In order to approximate the quantities in Equations 4.7-4.9 I take a similar approach employing the scheme $\pi_{L\&S}$, created by Li and Stephens and described in Chapter 2. The full implementation of this scheme is presented in the next section.

4.2.1 Implementing the Method

Having outlined the Importance Sampling approach I now describe the specific implementation used here and discuss some of the various difficulties and the decisions that must be taken in producing a genealogical importance sampler. I lead the reader through the simulation of a simple genealogy and describe the calculations that must be performed at each stage.

Given a data set it is first necessary to decide on the parameters of the simulation. The first important parameter is the mutation rate, θ. Although it is possible to co-estimate θ using the importance sampler such an approach would greatly increase the parameter space. Instead, as computational efficiency is of such importance, I use Watterson's estimate of θ [42] as it is extremely fast to calculate and the estimates of ρ did not seem sensitive to small perturbations in θ. Secondly, a recombination parameter, ρ, must be specified. This will normally be specified by the user. For simplicity I assume constant recombination rate here, although the extension of the *model* to an arbitrary recombination map is trivial (inference could become much harder). Note that exploring the state space of all maps would be extremely computationally expensive and without some severe restrictions on the type of variation would be impractical with these methods at present.

Next the (prior) instantaneous rates of each type of event, and hence the total event rate in the epoch must be calculated.

1. Coalescence: The coalescence rate is simply the number of pairs of sequences that can coalesce

$$\lambda_C = \sum_{i,j \leq k,\ i \neq j} I_C(i,j), \tag{4.10}$$

where $I_C(.)$ indicates whether a pair of sequences can coalesce under

the appropriate prior. Under the Coalescent model Equation 4.10 gives

$$\lambda_C = \sum_{i,j \leq k,\ i \neq j} 1 \ = \ k(k-1)/2 \tag{4.11}$$

where k is the number of extant sequences. Under the SMC it is necessary to calculate which pairs of sequences share ancestral sites, then λ_C is the total number of such pairs.

2. Recombination: In this sampler the sequences are modelled as discrete with gaps of different sizes between sites, the recombination events are forced always to occur precisely half way between each site. The recombination parameter above specifies the total recombination rate, per sequence, at the start of the simulation. This is then broken up between each of the gaps between segregating sites according to the distance between each pair of sites, the total length of each sequence is defined as 1 for this purpose. The total instantaneous recombination rate, λ_R, is then taken as the sum over all sequences:

$$\lambda_R = \sum_{i=1}^{k} \frac{\rho}{2} \times (d_{r,i} - d_{l,i}) \tag{4.12}$$

where $d_{r,i}$ and $d_{l,i}$ denote the right and leftmost boundaries of ancestral material on the i^{th} sequence respectively. Initially $d_{l,i} = 0$ and $d_{r,i} = 1$

for all i. The prior weight of a recombination event between sites j and $j+1$ on a given sequence is $\frac{\rho}{2} \times \delta_j$ where δ_j is the distance between sites j and $j+1$.

3. Mutation: A constant mutation rate is assumed and the mutation rate is divided evenly between each of the (finitely many) sites. To obtain the total instantaneous mutation rate, λ_M, requires a sum over every site on every sequence:

$$\lambda_M = \sum_{i=1}^{k} \sum_{j=0}^{L} \frac{\theta}{2L} \times I_M(j) \tag{4.13}$$

where each sequence contains L sites (segregating in the population) and $I_M(.)$ is an indicator function which takes value 1 when site j is ancestral and 0 otherwise. The prior weight for each individual possible mutation is $\frac{\theta}{L}$.

These formulae specify the prior rates of each event at a given time in the history of the sample. This, in turn, provides the distribution of the type of the next event. If the genealogy is also of interest then given the total rate of events, $\lambda = \lambda_C + \lambda_M + \lambda_R$, it is possible to simulate a *time*, $t \sim \text{Exp}(\lambda)$, to the next event. However, the likelihood of the data is independent of this time and so I consider only the event sequence here. Having calculated

the prior rates $P(E)$ in Equation 4.6 for each event it is then necessary to estimate the relative probabilities of the data before and after each potential event. First however it is possible to eliminate certain events by simple observation. As an example, consider the data set in Figure 4.1. If the first event backwards in time were a coalescence event between sequences 1 and 2 then their haplotypes would be the same. However, as this is not true such an event would give rise to a genealogy, G, which gave $P(D \mid G) = 0$. The importance sampler therefore never generates such events. Similarly, under the infinite sites assumption each site can undergo mutation at most once in the genealogy. Consequently mutation events can only occur at those sites which contain a singleton in the sample. In practice it is necessary to enumerate all of the possible events that could give rise to an infinite sites genealogy that is compatible with the data. The only possible events for the data in Figure 4.1 are:

1. A coalescence event between sequences C and D.

2. A mutation event at site one on sequence B.

3. A recombination event at any of the 8 gaps between the sites on the sequences.

To calculate the weights given to each of these events by the importance sampler it is necessary to go back to Equations 4.7 to 4.9. In order to

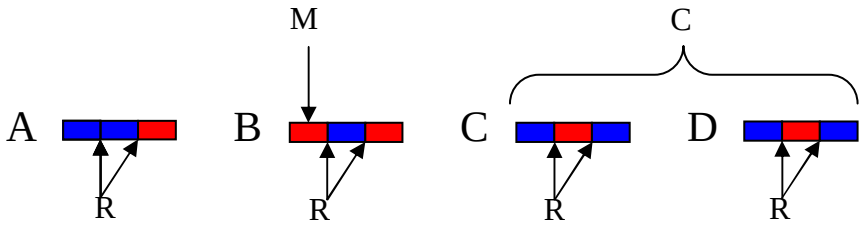

Figure 4.1: This figure shows all of the possible first events backwards in time under the infinite sites model for this toy data-set.

use these formulae an approximation to the quantity $P(h^*|h_1, \ldots h_k)$ for any haplotype h^* and set of existing haplotypes $\{h_1, \ldots, h_k\}$ is needed. In their paper Fearnhead and Donnelly propose a dynamic programming method that approximates these values. In Chapter 2 I investigate a range of methods which approximate these likelihoods. Using the results obtained in that work I decided to use the scheme $\pi_{L\&S}$ as this scheme was by far the fastest and the likelihoods are strongly related to those of the other approaches.

Note that in the following description I order the sequences in the equations such that h_1 (and sometimes h_2) are the sequences affected by the event being considered, merely for notational convenience.

1. In order to apply $\pi_{L\&S}$ to Equation 4.7 it is first necessary to construct the sequence h_C (from Equation 4.7) using the proposed coalescing sequences (which I name h_1 and h_2). This sequence is defined to be non-ancestral at all those sites where both h_1 and h_2 are non ancestral and due to the restrictions on coalescence events the types of h_1 and

h_2 are identical at all sites where both are ancestral, so h_C then takes the types of h_1 and/or h_2 wherever these are ancestral. Note that when recording these weights for the purposes of likelihood calculation the coalescence weights must be multiplied by two. This is because coalescing pairs are chosen in an ordered way, but there is no ordering of the two sequences in coalescence events.

2. Similarly, in order to calculate $\pi_{L\&S}(h_M \mid h2, \ldots, h_k)$ the haplotype h_M must be constructed. This is achieved simply by altering h_1 at the appropriate site - so that it has the same type as the rest of the sample.

3. In the case of recombination a further approximation is made. This is important as there are usually many possible recombination events and the quantity $\pi_{L\&S}(h^* \mid h_1, \ldots, h_k)$ is expensive to compute. Using the same notation as in Equation 4.9 I now make the approximation that

$$P(h_{1,<j} \mid h_2, \ldots, h_k)P(h_{1,>j} \mid h_{1,<j}, h_2, \ldots, h_k) \approx$$
$$P(h_{1,<j} \mid h_2, \ldots, h_k)P(h_{1,>j} \mid h_2, \ldots, h_k). \qquad (4.14)$$

This is close approximation as $h_{1,<j}$ shares no ancestral sites with $h_{1,>j}$ so only affects the distribution of $h_{1,<j}$ indirectly. However, this approximation allows a very significant computational saving. As discussed

later I make the approximation that, in the dynamic programming algorithm of $\pi_{L\&S}$, the probability of the type of a non ancestral site given the rest of the sample is always 1. Note that intermediate values (or partial probabilities), $P(h_1 \mid h_2, \ldots, h_n)_{<j}$, from calculating $P(h_1 \mid h_2, \ldots, h_n)$ in the dynamic programming algorithm correspond to the probability of observing the partially ancestral sequence: $h_{1,<j}$. By creating a sequence h'_1 constructed by reversing sequence h_1 the quantities $P(h_{1,>j} \mid h_2, \ldots, h_n)$ can be seen to be the partial probabilities $P(h'_1 \mid h_2, \ldots, h_n)_{<(L-j)}$. By storing these probabilities for each j it is then possible to calculate the probabilities in Equation 4.14 using only the partial probabilities used in calculating $\pi_{L\&S}(h_1)$ and $\pi_{L\&S}(h'_1)$.

To choose from the possible events it is first necessary to normalise the quotients from Equation 4.6 and then multiply them by their respective probabilities under the prior.

To calculate the likelihood for each event it is necessary to record the quantities $P(E)$, the probability of E under the prior, and $Q(E)$, the probability of E under the proposal distribution. Under both the prior and this proposal distribution the probability of generating each genealogy is the product of the probabilities of each event . Hence it is possible to calculate the

importance weight, $P(E)/Q(E)$, for each event and take the product of these as the importance weight for the whole graph. After each event it is then necessary to update the sampler, the new haplotypes need to be created, the new rates for each event must be calculated and any record of the genealogy may need to be updated. An example sequence of events is given in Figure 4.2. Note that this sequence of events is not equivalent to an ARG, and if the genealogy itself is of interest then that needs to be stored separately. Given the sequence of events the probability of the sample configuration is simply 1 as all mutations are specified and uniquely specify the data. The contribution towards the likelihood (in Equation 4.2) for this genealogy is then the importance weight $P(G)/Q(G)$.

I now consider some of the technicalities with the above approach. The most important consideration being how best to handle non ancestral material in the scheme $\pi_{L\&S}$.

4.2.1.1 The problems with Non Ancestral Material

When a recombination event occurs backwards in time the sequence on which it acts is split into two parts. One part carries the ancestral material from the left of the breakpoint and the other from the right. The rest of these two sequences contain material which is not ancestral to the sample. There is no need to simulate events that affect only non ancestral material. Non

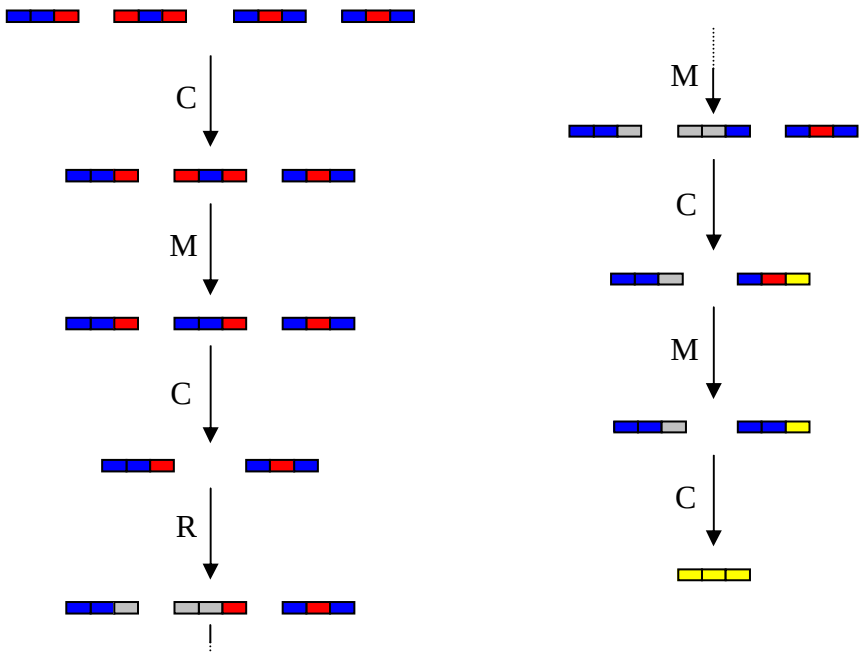

Figure 4.2: This diagram shows a full simulation using the importance sampler. The colour grey indicates non ancestral material and the colour yellow indicates material that has reached its most recent common ancestor. The letters C,M and R indicate Coalescence, Mutation and Recombination events respectively. The sequence continues on the right hand side. Note that events that occur in either non ancestral material or affecting sites which have reached their most recent common ancestor need only be simulated if they also affect ancestral sites that have not reached their MRCA. When calculating the likelihood simulation stops after the MRCA ancestor has been reached at each site. Note that there may still be more than one lineage in the ARG when this happens. Although, under the coalescent, it is possible to simulate genealogies back to the single most recent common ancestor of the whole sample this is not always possible under the SMC - there may remain pairs of lineages which can never coalesce.

ancestral types are unknown, and in any given epoch non ancestral material can be treated as missing data. An important question is then: how can we calculate the likelihood of data when some of the types are unknown? In particular, how can we calculate the quantities in Equations 4.7 to 4.9 in the presence of missing genotypes? There are two distinct problems here, and they are often concurrent.

1. What is $P(h_1 \mid h2, \ldots, h_k)$ when one or more of $h_2, \ldots, h_k$ contain missing genotypes?

2. What is $P(h_1 \mid h2, \ldots, h_k)$ when h_1 contains unknown types?

One way to tackle such questions is to *impute* the types at non-ancestral loci. This is the approach taken by Fearnhead and Donnelly [21] in their importance sampler. The imputed types are generated according to their frequency in the remaining sample in the epoch considered. It is then possible, due to the approximate model of sequence evolution used in their likelihood calculations, to sum over all possible imputations.

The same approach is possible using the conditionals $\pi_{L\&S}$ used here but the approach of explicit imputation was not taken for this importance sampler. In case 1, the probability of observing an unknown type at site j

given either a known type or an unknown type I used

$$Pr(H_1(j)) = \begin{cases} 1/2 & \text{if site } j \text{ is still segregating in the sample} \\ 1 & \text{otherwise.} \end{cases}$$

In case 2 I used

$$Pr(H_1(j)) = 1. \tag{4.15}$$

The question of how to calculate these likelihoods refers only to calculations in the *proposal* distribution. I attempt to find the best trade-off between accuracy and computational speed. The equations above provide two gains in speed, firstly the frequency of each type at each site need not be calculated after every coalescence or recombination event. More importantly, the extra computational time of summing over all possible imputations adds a further burden to the most computationally intensive calculations in the importance sampler. It is worth noting that the approach of imputing types according to the frequency of types where those are observed is itself an approximation. The types at the missing data are, sometimes strongly, dependent on the surrounding sequence. An example of this is given in Figure 4.3.

In order to understand the reasoning behind Equation 4.15 it is useful to consider what is meant by $P(h_1 \mid h_2, \ldots, h_k)$. This is the probability of observing h_1 given the already observed sequences in the sample. However,

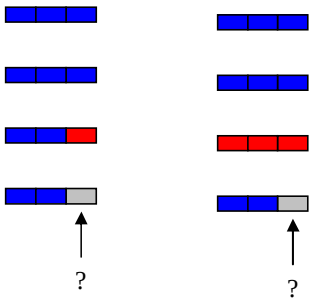

Figure 4.3: These two data sets illustrate how imputing types at non ancestral material based only on the frequency of that type in the sample might lead to inaccurate imputed frequencies of types. The grey colour indicates non ancestral material. On the left hand side the probability of observing the blue type might reasonably be assumed to be 2/3. However on the right hand side the rest of the haplotype structure indicates a higher probability of a blue type as there is some evidence for linkage disequilibrium between the first two sites and the third site.

we implicitly condition on having actually observed the types at h_1. If we consider the (unusual) question: What is the probability of observing the unknown genotype at position j on h_1 given that we were unable to type h_1 at this locus? This leads to the standard missing data approach - and we get a probability of 1.

A related problem, when considering the conditional probabilities from Equations 4.7 to 4.9, is how to treat ancestral material that has reached its MRCA. However in this case the MRCA has, by definition, been reached on *all* sequences and so all of the types are the same. These sites therefore do not contribute to the overall likelihood, and in the dynamic programming algorithm this is achieved by using a emission probability of 1.

4.2.2 Rephrasing the problem

Note that in the current formulation of the importance sampler the following approximation is used to calculate the probability of the data:

$$P(D) = \int Q(G)\frac{P(G)}{Q(G)}P(D \mid G)dG \approx \frac{1}{M}\sum_{i}^{M}\frac{P(G_i)}{Q(G_i)}P(D \mid G_i) \qquad (4.16)$$

where the genealogies in the summation on the right are simulated from $Q(.)$. However, when the genealogies are augmented with mutation the data is defined by the genealogy, so

$$P(D \mid G_i) =$$
$$= \begin{cases} 1 & \text{if } G_i \text{ gives rise to the data} \\ 0 & \text{otherwise.} \end{cases}$$

This means that many of the terms in equation 4.16 are zero and this is used to improve the efficiency of $Q(.)$. It is not possible for the Q used here to propose genealogies that do not give rise to the data. Technically this makes Q an invalid proposal distribution, as the support of P should be contained in the support of Q. However, the estimate is not affected as those genealogies not in the support of Q will always supply 0 weight in the approximation.

However, it is possible to reformulate this method such that this tech-

nicality does not arise, and it leads to a possible improvement in efficiency. Note that, by Equation 4.17

$$\int P(G)P(D \mid G)dG = \int_{G\in\Gamma} 1 \times P(G)dG \tag{4.17}$$

where Γ is the set of all genealogies that give rise to the data. In this formulation the total state space is Γ, the genealogies are distributed uniformly throughout Γ and the integral is that we wish to estimate is of the quantity $P(G)$, which represents the prior weight of that genealogy under the prior. Importance sampling can now be used to approximate this new integral, so that

$$\int_{G\in\Gamma} P(G) \times 1dG = \int_{G\in\Gamma} P(G) \times Q(G)\frac{1}{Q(G)}dG \approx \frac{1}{M}\sum_{i}^{M} P(G) \times \frac{1}{Q(G_i)} \tag{4.18}$$

where the genealogies are sampled from $Q(.)$, hence $G_i \in \Gamma$. Note that here the importance weights are proportional to $1/Q(G_i)$. This formulation leads to exactly the same estimate of the likelihood as in Equation 4.6.

One method for improving the efficiency of an importance sampler (see eg. [41]) is to replace $1/M$ in the sum on the Right hand side of Equation

4.18 by the sum of the importance weights. That is, proposing that

$$\frac{1}{M}\sum_{i}^{M}\frac{P(G_i)}{Q(G_i)} \approx \left(1/\sum_{i}^{M}\frac{1}{Q(G_i)}\right) \times \sum_{i}^{M}\frac{P(G_i)}{Q(G_i)} \tag{4.19}$$

Although it is stated in Liu [41] that this estimator is only slightly biased and that it is often a good approximation to the integral of interest this was not found here.

Consider the optimal importance sampler, that is, $Q^*(.)$ such that $Q(G) \propto P(G)$ (the prior probability of G_i), remember that the genealogies are restricted to those which give rise to the data. Note that this is equivalent to the optimal importance sampler in the first formulation because:

$$\begin{aligned} Q^*(G) &= P(G \mid D) = P(D \mid G)P(G)/P(D) \\ &= \begin{cases} P(G)/P(D) & \text{if } G_i \text{ gives rise to the data} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

It can now be seen that this does *not* lead to a good approximation to the likelihood as, substituting the optimal importance sampler into Equation

4.19 we get:

$$\left(1/\sum_{i}^{M}\frac{1}{Q(G_i)}\right)\times\sum_{i}^{M}\frac{P(G_i)}{Q(G_i)}=P(D)/P(D)\times\frac{\sum_{i}^{M}1/P(D\mid G)}{\sum_{i}^{M}1/P(G)P(D\mid G)}$$

and as $P(D \mid G) = 1$ for all G in this case this becomes

$$\frac{M}{\sum_{i}^{M}1/P(G_i)}. \tag{4.20}$$

Thus this approximations suggests that a good estimate of the likelihood is the reciprocal of the harmonic means of the prior densities of the genealogies compatible with the data. However, it is easy to see that this is not a good estimate, consider the approximation in 4.19, each of the terms $Q(G_i)/P(G_i)$ is an estimate of the probability of the data. For an optimal Q each of these terms is in fact precisely $P(D)$. However, in this case the expression on the right hand side of the approximation gives:

$$\left(1/\sum_{i}^{M}\frac{1}{P(G_i\mid D)}\right)\times\sum_{i}^{M}P(D)=\frac{M\times P(D)}{\sum_{i}^{M}\frac{1}{P(G_i|D)}}. \tag{4.21}$$

Given that $P(G_i \mid D) \leq 1$, this requires $P(G_i \mid D) \approx 1\ \forall i$. Hence this is only a good approximation to the $P(D)$ when the genealogies are well specified by the data. Unfortunately, in the presence of recombination, the data is highly uninformative about the genealogies and the quantities $Q(G_i)$ are generally

very small, this is true even under the optimal importance sampler.

4.2.2.1 An approach to improving efficiency

A natural approach to approximating the optimal importance sampler considers every possible event backwards in time and chooses according to the (approximate) distribution of events conditional on the data. However, calculating each of these approximate conditional probabilities is computationally very expensive so, in their paper [21], Fearnhead and Donnelly propose a simplification designed to reduce this computational burden. Instead of calculating the full weights for all possible events they first choose a haplotype according to the prior rate of events. More formally, denote the fraction of sites on sequence h_i that are ancestral by $p_{m,i}$ and let $d_{r,i}$ be the total recombination distance on sequence i between ancestral sites. Also note that the total number of sequences that h_i can coalesce with is $(k-1)$. Fearnhead and Donnelly then approximate the total rate of events involving h_i by the quantity

$$\frac{\theta}{2}p_{m,i} + \frac{\rho d_{r,i}}{2} + \frac{(k-1)}{2}. \tag{4.22}$$

To choose a sequence they then normalise these rates to create a probability of choosing each sequence. This method is not quite as accurate an approximation to the conditional probabilities as using the method so far described

as it does not condition on the data when choosing the sequence to take part in the next event. I shall call this the Fearnhead approximation.

A technicality that arises from choosing haplotypes in this way is that there are two potentially unequal routes to choosing each possible coalescence event between two sequences, denote the two relevant sequences as h_1 and h_2. In order to correct for this the coalescence rate in 4.22 is half of what might have been expected with a coalescence rate of 1 for each of the remaining $k-1$ sequences. However, in itself this is not sufficient. It is necessary to calculate the probability of generating this coalescence given that either h_1 or h_2 were chosen. That is, if a sequence, h_1, were chosen using the Fearnhead approximation, and the event chosen for h_1 were a coalescence event with h_2. It is then necessary to calculate all of the weights on h_2, normalise these and to calculate the probability of choosing h_2 and then a coalescence event with h_1 given that h_2 was chosen. Note that this correction *cannot* be ignored as it is required to calculate the appropriate importance weight, and is not a refinement to the proposal distribution.

I now investigate some of the properties of using the Fearnhead approximation. I first consider the case of very small ρ and try to demonstrate that the new approach may perform badly in here. First of all I consider a completely artificial data set which illustrates why choosing a sequence to act on without conditioning on the data has the potential to reduce ef-

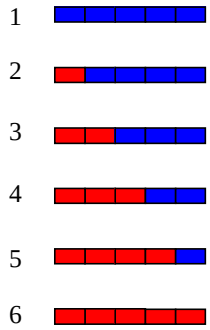

Figure 4.4: The data in this diagram is fully compatible with a tree. When the recombination rate is low the distribution of genealogies conditional on the data should contain very few recombination events.

ficiency. Figure 4.4 shows a data set with no incompatibilities. However, when considering the first event while simulating a genealogy backwards in time each sequence has an equal probability of being picked under the Fearnhead approximation. Assume that the ancestral and derived types are not known (the effect is more severe when they are, as only one sequence can mutate, not two). No coalescence events are possible with this data as no two sequences are identical. So, if any of the sequences from 2-5 are picked, mutation events cannot be simulated as the infinite sites assumption allows only singletons to mutate. Therefore recombination events must be simulated when sequences 2-5 are chosen. However, when $\rho \approx 0$ this is a very poor choice of event. To gauge the effect this has on the importance sampler I ran both the full conditional scheme and the Fearnhead approximation on this data set. I simulated 10,000 genealogies under $\rho = 1 \times 10^{-9}$ and both approximations calculated the log likelihood as -23.3. However, the average

number of recombination events under the Fearnhead approximation was 2.8 per simulation. These recombination events lead to considerable variation in the likelihoods estimated. Under the full method not a single recombination event was simulated and the estimates of the likelihood were very similar between runs. The first ten likelihoods under both schemes are displayed in Table 4.1. The full method took 77 seconds (for all of the runs) while the Fearnhead approximation took only 49 seconds to calculate. For data constructed in this way the Fearnhead approximation performs much worse as more sequences are added. These data sets are not typical of population genetic data or of data simulated under the coalescent. However, the Fearnhead approximation performs similarly poorly for most data sets analysed under small ρ, even those with incompatibilities.

The real purpose of these importance samplers is to analyse data in the presence of significant recombination and so this comparison does not reflect the true relative performance of the different approaches, it merely provides an intuition into the differences between the two schemes and when one might outperform the other. In fact, for higher values of the recombination rate the Fearnhead approximation seems to outperform the full method quite considerably. The time taken to simulate each genealogy is reduced, and the variation in the likelihoods is also often reduced. It is not clear why the variation in likelihoods should be reduced, however it is possible that

Run	Full Scheme	Fearnhead Approximation
1	-24.8	-67.2
2	-24.8	-113.2
3	-24.3	-43.0
4	-24.6	-90.3
5	-24.3	-111.9
6	-24.0	-113.1
7	-24.2	-89.9
8	-24.3	-88.3
9	-24.3	-88.7
10	-24.6	-109.5

Table 4.1: The first ten likelihoods calculated under both schemes, the full scheme and the Fearnhead approximation where a haplotype is first chosen according to the prior rates of events. Although there is considerable variation in the likelihoods in the Fearnhead scheme they do converge to the same likelihood as calculating the full scheme. However considerably more genealogies must be simulated in this case in order to achieve convergence.

there are certain situations where the approximate conditionals cause some events to be severely under-weighted causing certain genealogies to be very unlikely under the proposal distribution when the full method is used (see, for example, Figures 4.13 and 4.14 later on). Perhaps the method for choosing haplotypes according to the prior rates flattens the distribution of possible events and allows a more uniform exploration of certain parts of the state space.

4.2.3 Simulation Study

In order to assess the behaviour of the two models for performing inference I wrote a program to implement the Importance Sampling method described

above in C++, using both the coalescent and the SMC models.

To simulate the data I used the program 'make sample' by Hudson. To be confident that the importance sampling scheme would generate accurate likelihood estimates I used very small data sets. The values under which the data were simulated were $\theta = 3$ and $\rho = 5$, there were 10 sequences in each sample and 100 samples in total. This gave a range of numbers of segregating sites from 1 to 20, with the bulk being between 5 and 15 sites long.

The importance sampler, under both the Coalescent and the SMC, was run on each of these data sets. I calculated the likelihood on a grid of values,

$$\rho \in \{1 \times 10^{-9},\ 0.1,\ 0.3,\ 0.5,\ 1,\ 3,\ 5,\ 7,\ 10,\ 13,\ 16,\ 20\}$$

using 500,000 independent runs under both models, for each value of ρ and for each data set. These genealogies were also used to estimate the likelihoods at intermediate values of ρ, by correcting the importance weights from the genealogies for the different underlying recombination rates. However, under such a coarse grid it is questionable whether the likelihoods for intermediate values of ρ would be well estimated using this approach. This means that only approximate values for the MLE can be calculated. However, the signal for ρ is very weak in the data sets that the importance sampler can analyse effectively. It may not be possible to correctly estimate the underlying

recombination rate to within even a factor of ten in some of these samples. So there is little value in calculating ρ at a finer grid of points. The time to the most recent common ancestor as well as the total time in the tree at the left and right most edges were recorded, as well as the number of recombination events in each genealogy. The primary analysis consists of examining how well the likelihoods were estimated and comparing the two models, Coalescent and SMC.

4.3 Results

The SMC is only a useful model if inference under the SMC reflects biological reality. I use the ability to approximate the coalescent model as a measure of how successful the SMC is in this task. In particular I investigate whether likelihoods calculated under the SMC are similar to those calculated under the Coalescent. Encouragingly Figure 4.5 shows that the likelihoods calculated under the two models are very similar. However, in this implementation and using the parameter values here, the time taken to calculate the likelihoods was not greatly reduced under the SMC. This is because only relatively small values of ρ could be explored in this study. The Coalescent model and the SMC are very similar when ρ is small. Unfortunately the parameters of inference are currently severely restricted because the impor-

tance samplers converge very slowly for large numbers of segregating sites. It would be very interesting to understand the relationship between estimates of ρ under the two models. Preliminary analysis suggest that there is a slight decrease in estimated recombination rates under the SMC. However it is difficult to perform a proper analysis of this effect because very few of the data sets had significant maximum likelihood peaks that were not at the boundaries explored. This effect need not harm inference under the SMC a great deal, but it does suggest that recombination rate related results from the SMC cannot be implemented in full coalescent methods downstream, instead the SMC should be used throughout for consistency. Finally, it was also apparent in those data sets analysed that the correlations in tMRCA of the sample across sites separated by a certain recombination distance were reduced under the SMC, as expected (data not shown).

ESS

I wish to analyse the performance of the importance sampling scheme and how this performance varies with changes in the underlying rates and the dimensions of the data. With any Monte Carlo estimator it is necessary to assess whether or not the process has converged. Unfortunately it is often hard to assess how certain we can be of convergence given the results so far. As, in general, only a sample of the total space has been explored

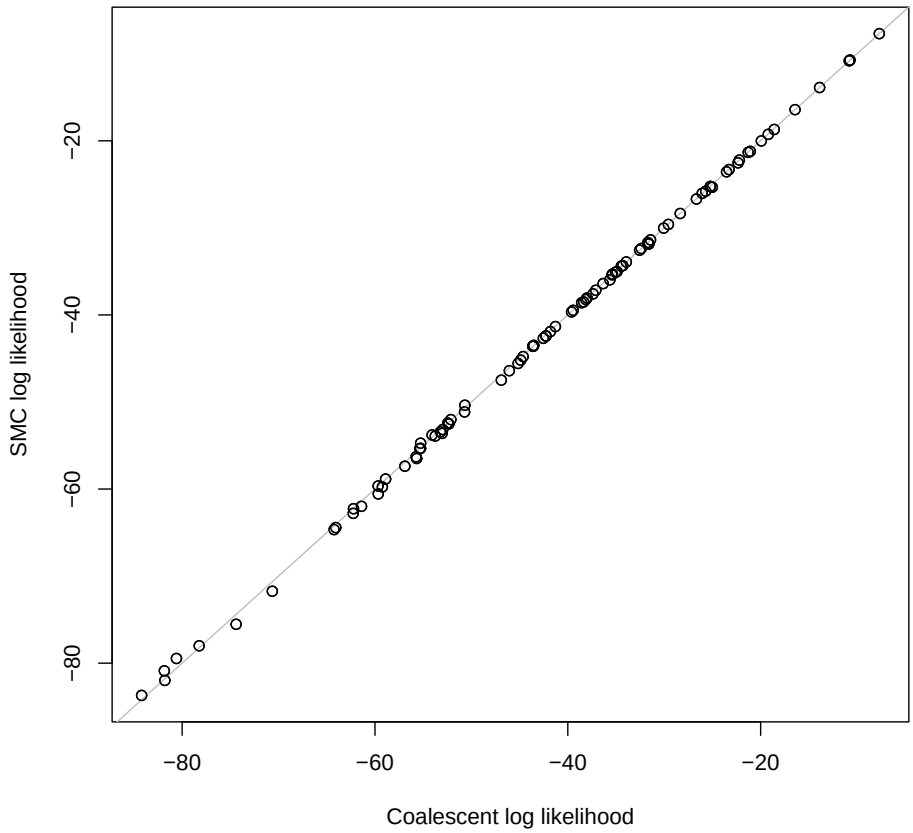

Figure 4.5: This graph shows the log likelihoods under the coalescent against those under the SMC for all of the datasets under $\rho = 20$. I chose this value as the differences between the SMC and coalescent likelihoods are expected to be greatest when ρ is large. It is worth noting that the estimates are least accurate when $\rho = 20$ and much of the deviation from perfect correlation for larger data sets (with lower likelihoods) is probably due to inaccuracy in the Monte Carlo estimates.

it is usually impossible to know how significant the contributions from the unexplored regions of the state space would be. One approach to estimating convergence is to assume that the variation in future samples will be similar to those already simulated. This is an imperfect approach, but I use it here to get an indicator of how well the importance sampler estimates different likelihoods.

Define the effective sample size, or ESS, of a set of simulations of size N to be

$$S_E = N \times \frac{1}{1 + V_C^2} \tag{4.23}$$

where V_C denotes the coefficient of variation of the N sampled likelihoods. The ESS (see eg. [43] pp283-284) is a linear approximation to the efficiency of an importance sampler relative to the optimal importance sampler. That is, how well the approximate sampler explores the posterior distribution of interest (in this case ARGs relating to the sample) and in particular how the variance in statistics of the genealogies decreases as extra samples are introduced. This formula is not guaranteed to converge to the true relative efficiency, but it is a simple measure of performance that is easy to calculate. Larger values of S_E indicate greater efficiency and $1 \leq S_E \leq N$. It is normally not possible to calculate the true value of even S_E as the mean and variance of the target distribution are, in general, unknown. However we can estimate

the mean and variance of the sample and use these to estimate S_E:

$$\widehat{S_E} = N \times \frac{1}{1 + \widehat{V_C^2}} \tag{4.24}$$

Unfortunately the performance of genealogical importance samplers is poor, that is: even when large numbers of genealogies have been simulated there may be genealogies which would contribute very large amounts to the estimated likelihood that have not been sampled. If these genealogies were sampled they would also significantly increase the sample variance and hence greatly reduce the effective sample size. This leads to the situation that sometimes estimates of S_E are strongly upwardly biased. Low values of S_E should therefore be treated with suspicion (as they may not be *low enough*) and in some situations even very high values may exaggerate the evidence for convergence. Another effect is that inflated ESS values tend to be underestimates of the likelihood while overestimates of the likelihood tend to have very low estimates of the ESS, this effect is shown in Figure 4.6

Despite the large degree of uncertainty in estimates of the ESS there are practical reasons to believe that there is information in this measure. When insufficient runs are used to effectively sample the rare highly contributing ARGs then estimates in the likelihood should be poor. However, by examining the similarities of the likelihoods between independent estimates of the

likelihood it can be seen that there is very little variation in the estimates produced for the number of runs used here. This encourages the belief that the ESS is approximately correct and can be used as a guide to comparing the performance of different methods when the estimates of ESS are high.

4.3.1 Performance differences between the SMC and the Coalescent

Although full inference under the SMC is not yet possible for large data sets, there is still a notable improvement in the efficiency of generating likelihood values under the SMC at higher values of ρ. This is due to the simpler space of ARGs under the SMC. In each epoch there are fewer possible events (due to restrictions on coalescence events). Also, as there is no trapped non ancestral material, there are (on average) fewer recombinations (see Figure 4.7) in the ARGs. This reduces the complexity of the space of ARGs that contribute significantly to the likelihood. Although it was not possible to calculate accurate likelihoods for all the data sets for very high ρ where this has been tried the SMC improvement is much more extreme for higher ρ (data not shown). Under $\rho > 100$ coalescent genealogies often contain thousands of recombination events created by coalescing lineages with very small amounts of ancestral material and then recombining on the resulting

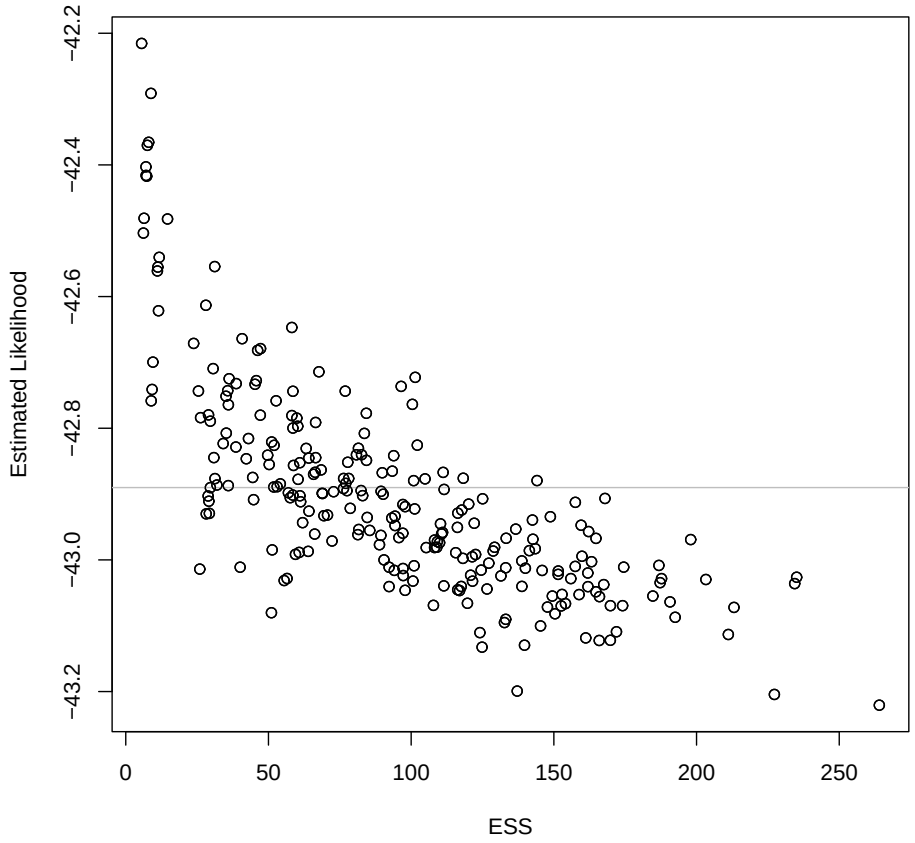

Figure 4.6: This graph shows estimates of the log coalescent likelihood plotted against the estimated ESS for 250 independent subsamples of 25000 likelihood estimates for data set 52 (9 segregating sites). It can be seen that there is significant difficulty in estimating the effective sample size. It can also be seen that high ESS values can be misleading and cause bias. The likelihood is strongly correlated with the reported ESS. This effect is caused by rare simulations that produced a genealogy with very high weight. Such runs increase the variance significantly while also increasing the estimated likelihood. The minimum estimated ESS was 5.56 while the maximum was 264, and this is very typical for such a sample. Note that the coefficient of variation in estimated ESS values was 0.6 while the coefficient of variation for the mean likelihood estimates was less than 0.2. In fact overall estimates of the ESS were significantly more variable than estimates of the likelihood itself. The estimate of the likelihood from 500,000 independent genealogies is given by the horizontal grey line.

lineage to give rise to the original pair of sequences. This pattern can repeat many times in one genealogy. Under the SMC far fewer recombinations occur as 'trapped non-ancestral material' (see Chapter 3) cannot be created under the SMC. Figures 4.8 and 4.9 show that this reduction in the state space does lead to improved efficiency, however the improvements due the to SMC are overshadowed by the fact that both models are severely restricted in the quantity of data that can be practically analysed.

4.3.2 The Performance Drops considerably as Data size increases

The performance of the importance sampler varied dramatically between data sets. For data sets with fewer than 10 segregating sites about 5 hours computing time on a standard Pentium 4 desktop computer was enough to calculate an very accurate likelihood curve from 500,000 samples for ρ up to 20. However for 20 segregating sites 100 hours of computing time produced likelihoods with evidence of significant variation, in one case by almost an order of magnitude. The rate of convergence of the likelihood also has a strong impact on the convergence to the true distribution of other statistics observed, such as the estimated average time to the most recent common ancestor or the estimated recombination rate. As the likelihoods differed

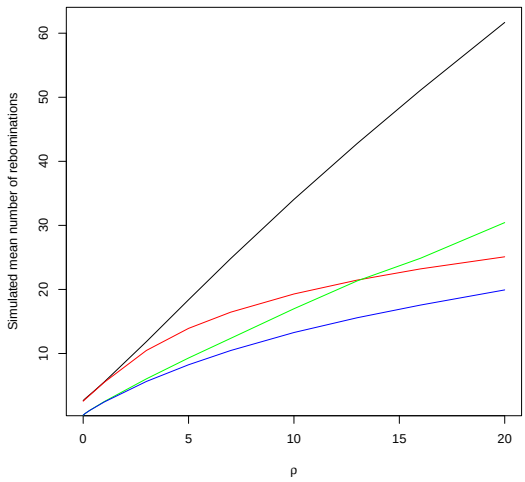

Figure 4.7: This graph shows the number of recombination events in each genealogy averaged over all data sets and all 500,000 runs. In black is the average number of simulated recombinations under the coalescent. In red is the average simulated under the SMC. The green and blue lines show the estimated posterior mean number of recombinations for the coalescent and SMC respectively. These were calculated by weighting the number of recombination events in each genealogy by its importance weight. The proposal schemes lead to, on average, more recombination events than the posterior distribution because, when incompatibilities are present, the proposal cannot always identify where recombination must occur (see Figure 4.13). This leads to recombinations in locations that do not remove any incompatibility (and coalescence events that may even introduce incompatibilities); further recombination events are then later required to remove these incompatibilities and reach the common ancestor.

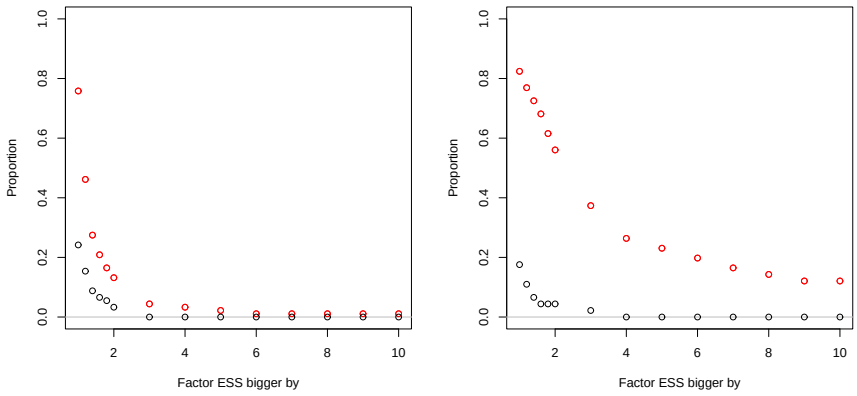

Figure 4.8: These graphs show the proportion of time that, for the 100 data sets analysed in the study, the coalescent ESS was bigger than the SMC ESS (black) and vice versa (red) (500,000 independent runs were used to calculate each likelihood). On the left hand side the genealogies were simulated with $\rho = 1$ and on the right with $\rho = 20$. The horizontal axis denotes the factor by which the the ESS must have increased for inclusion at that point. For example, the right hand graph shows that roughly 1 in 5 likelihoods calculated under the SMC had an associated ESS of more than 5 times that of those generated under the coalescent. It is clear that the advantages of using the SMC only start to become significant for larger values of ρ. The time taken per genealogy simulated is not taken into consideration here, that can be seen in Figure 4.9.

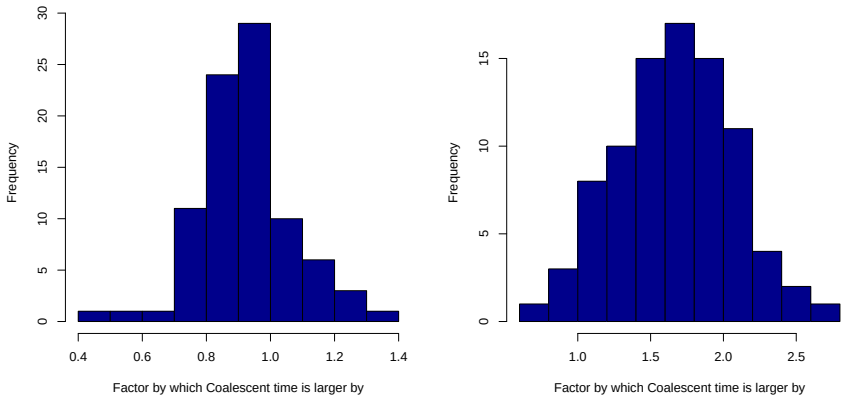

Figure 4.9: These histograms show the distribution of the CPU time under the coalescent (to estimate the likelihood using a 500,000 runs) divided by the CPU time under the SMC for the same likelihood also using 500,000 runs. On the left is the histogram for $\rho = 1$ and on the right for $\rho = 20$. Note that the SMC performs slightly worse for very small values of ρ, as there is slightly more to check under the SMC. However, when ρ increases there is a significant saving in the time required to simulate each genealogy.

by significantly less than an order of magnitude between different values of ρ the lack of convergence for large numbers of sites severely reduces our ability to estimate a likelihood curve or maximum likelihood estimate, unless methods using driving values were used. However, these approaches were not explored extensively here after initial results showed that highly misleading results may be obtained with little diagnostic power. Figures 4.10 and 4.11 give an indication of how the performance changes with size of the data set, measured by the number of segregating sites.

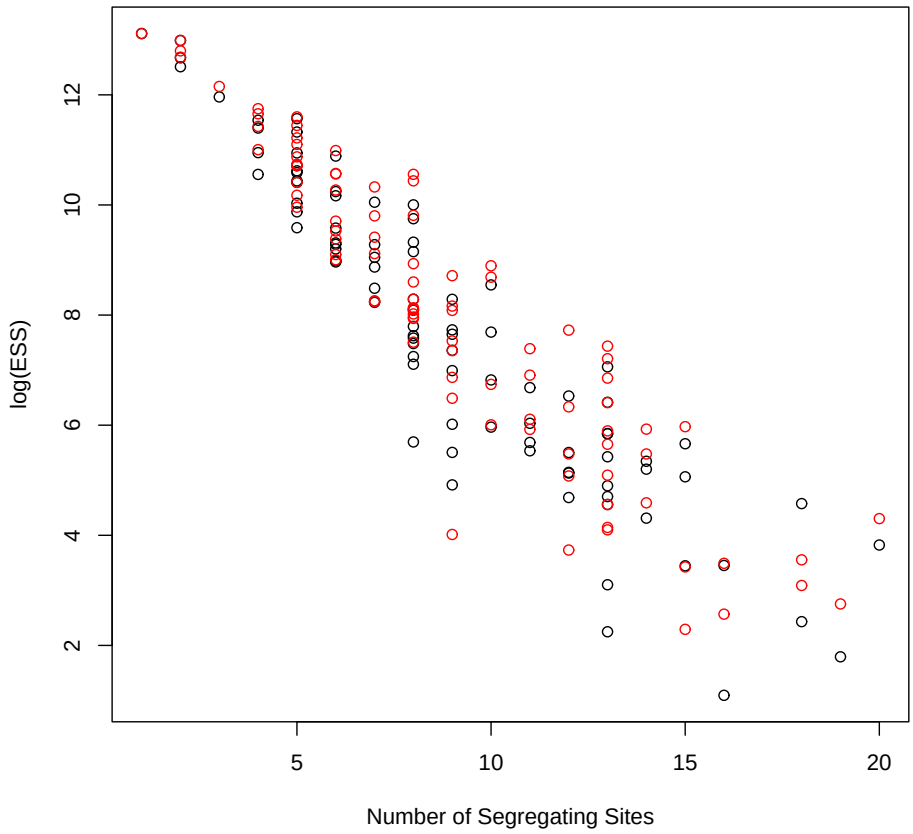

Figure 4.10: This scatterplot shows the log of the ESS estimates against the number of segregating sites for 500,000 runs, calculated for each data set at $\rho = 5$. The Coalescent values are in black and the SMC values are in red. The models suffer very strongly from increasing the length of the sequences. Note also that as the estimated ESS gets lower the estimates will also tend to be an overestimate of the likelihood. This means that the true relationship is probably even more severe.

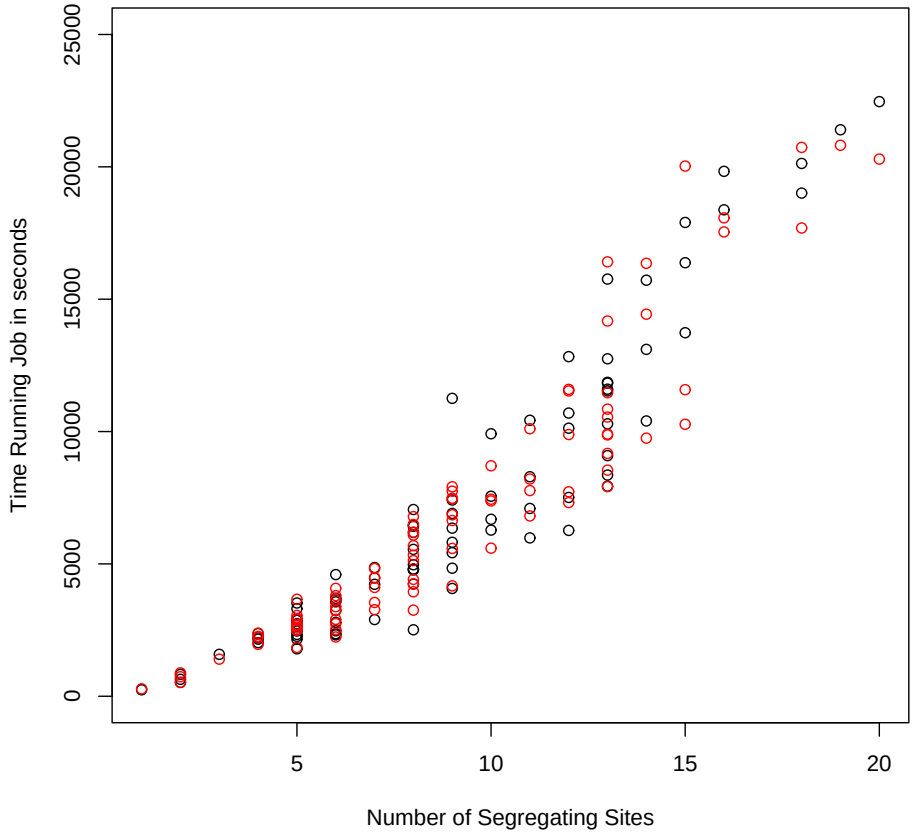

Figure 4.11: This scatterplot gives an indication of the computing time taken to generate 500,000 genealogies with 10 sequences under $\rho = 5$ against the number of segregating sites in each sample. The values for the SMC are plotted in red and the coalescent times in black. It is worth noting that this graph can only be used as an indication of the times that various data sets required to be completed as this is a naïve measure, simply taken by subtracting the time that the particular analysis was started from its completion time. Many factors may have influenced this, such as jobs being temporarily halted for other purposes. However, the trend seems clear and it's apparent that the time per simulation increases significantly with the number of sites to be analysed.

How Performance Changes with Different Sample Sizes

As well as changes in the number of segregating sites I also investigated changes in performance when different numbers of sequences were analysed. A full scale analysis with hundreds of data sets with large numbers of sequences is impractical. I analysed a small number of data sets with many sequences, but repeated the analyses using different proportions of the total set of sequences. From this it seems that increasing the number of sequences has a much less critical effect on the ability of the importance sampler to estimate the likelihood (than increasing the number of sites does). A summary of these results can be found in table 4.2.

While analysing these data it became apparent that a further computational saving could be made for data with large numbers of sequences. In these cases coalescence events happen at a much greater rate than when the number of sequences is small. When the event types were recorded it was common to see many coalescence events before any other event types were observed. Furthermore most coalescence events have very similar, or identical weights in the importance sampler, when multiple mutation events are possible in an epoch these also tend to have identical weights. Note, however, that different recombination events can have very different weights, when ρ is small some recombinations are vastly more likely than others. The obser-

n	ρ	ESS	Time	ESS	Time
		Coalescent		SMC	
2	10^{-9}	50000	59	50000	58
2	1	36797	93	36515	86
2	20	6769	680	5930	463
3	10^{-9}	25128	131	25175	135
3	1	18586	145	18507	205
3	20	2594	1428	3233	896
5	10^{-9}	30733	195	30730	188
5	1	22014	276	22878	272
5	20	821	1997	353	1225
10	10^{-9}	5902	583	5984	574
10	1	9083	816	9621	835
10	20	15	3489	108	2562
20	10^{-9}	595	2365	611	2434
20	1	1567	3439	2654	3494
20	20	31	12730	5.2	9291

Table 4.2: The performance of the importance sampler when subsets of various sizes are taken from a data set with 100 sequences. Similar results were found in all data sets observed (data not shown). The data was simulated with $\theta = 3$ and $\rho = 5$, there were 13 segregating sites and 50,000 independent runs were used to calculate the likelihood for each configuration of parameters. The time is given in seconds (real time) taken to complete the 50,000 runs for that number of sequences and value of ρ. The runs for 100 sequences would take about 9 days of computing time. Note that the ESS values for $n = 20$ show signs of unreliability (ESS is normally seen to, and is expected to, drop as ρ increases). Given these probably unreliable ESS values for 20 sequences it is likely that many months of computing time would be required to properly estimate the ESS values for $n = 100$.

vation that coalescence events comprise nearly all of the initial events when large numbers of sequences are analysed allows the construction of simplified approaches. These can produce significant reduction in computational burden of the likelihood calculation. There are many possible implementations that could be used, so I describe a general approach here.

1. Choose an event *type* according to the prior rates of events.

2. If this event is a recombination event then use the usual dynamic programming methods to approximate the likelihood of the data given each recombination event. Choose from each possible recombination event using the resulting weights (as in the earlier parts of this chapter, but with no coalescence or mutation events possible).

3. If the event type is a coalescence (or mutation) then choose uniformly and at random from the possible coalescence (or mutation) events.

Note that under this scheme certain significant errors could potentially be made. For example, if the recombination rate is low, but the data contain an incompatibility then recombination events that remove this incompatibility will be chosen too rarely, giving rise to high variance in likelihood estimates. Also, when sequences contain non-ancestral material some coalescence events could create new incompatibilities - greatly reducing the contribution for that

genealogy. In order to make this method effective it is necessary to exercise care in its use. It will take considerable thought and testing to find the best compromise between fast simulation of genealogies and accurate proposal distributions.

4.3.2.1 How does approximating the optimal importance sampler compare to previous approaches?

Although it is discussed extensively in their paper [21] it is worth noting again that the approach developed by Stephens and Donnelly [20], later extended by Fearnhead and Donnelly, is a considerable improvement over pre-existing methods. The previous approach used by Griffiths and Marjoram [17] did not consider the form of optimal importance sampler. In this approach although certain events could be eliminated through the infinite sites assumption and compatible genealogies could then be produced, the rates of the events at each stage were not adjusted for the likelihood of the resulting sample configuration.

Although this should lead to much faster simulation of events, and perhaps genealogies, the resulting importance sampler is in fact far less efficient than that of Fearnhead and Donnelly. Intuitively this is because recombination events are not placed in regions of incompatibility, but uniformly throughout the data. This leads to a situation where, although more recom-

bination events are simulated when the recombination rate is higher, when the recombination rate is low many events might still need to be simulated in order to remove incompatibilities from the data. Indeed even for higher recombination rates the average number of recombination events simulated under this scheme was considerably higher than under the scheme used here.

Tables 4.3 and 4.4 show the results from the Griffiths and Marjoram scheme and the scheme used here when run for 30,000 iterations on data set 1. Figure 4.12 shows the likelihood curves for ρ under the two schemes to give a more visual indication of convergence. Note that the increased number of recombination events simulated under the conditionals which did not use the information in the data led to more computing time being required than under the more expensive conditional schemes in some cases. Given that the optimal importance sampler requires information about the likelihood of the resulting sample configuration it is not surprising that methods which ignore this factor perform badly. This important observation applies to all importance sampling - in any scenario where chosen events will affect the likelihood of the resulting sample configuration to significantly differing degrees these likelihoods must be considered to achieve maximum efficiency.

ρ	Likelihood	ESS	$\hat{R}$	$\widehat{R_W}$	Time Taken
1×10^{-9}	-33.9	9445.9	0.5	0.0	594
0.1	-34.0	8256.9	0.7	0.2	548
0.3	-34.0	6835.1	1.1	0.5	583
1	-34.0	6336.8	1.6	0.8	607
1	-34.1	2836.1	2.9	1.7	646
3	-34.3	975.4	8.8	5.0	975
5	-34.5	769.4	15.4	8.4	1321
7	-34.6	148.5	21.8	11.7	1621
10	-34.9	222.1	30.9	16.4	2085
13	-35.0	116.8	39.9	20.8	2607
16	-34.9	23.8	49.3	25.9	4176
20	-35.5	82.1	59.9	33.4	5087

Table 4.3: This table shows the results for the first dataset obtained from 30,000 genealogies simulated under the scheme described in this Chapter. I used the approximation of haplotypes being chosen according to their prior rate of events to improve efficiency. The symbols $\hat{R}$ and $\hat{R_W}$ denote the average, and weighted average numbers of recombination events in the simulations. Under this scheme the likelihoods are well estimated. This can be seen from the consistent trend in ρ and the high values of ESS reported. Repeating the analysis also gave very similar values for the likelihoods (although some variation does occur at the third significant figure for higher ρ.) It is also worth noting that the average number of recombination events simulated under the proposal distribution has a moderately good relationship with the expect number of recombination events under the target (weighted) distribution.

ρ	Likelihood	ESS	$\hat{R}$	$\widehat{R_W}$	Time Taken
1×10^{-9}	-34.1	13.8	0.0	0.0	223
0.1	-34.3	32.4	2.0	0.2	255
0.3	-34.4	8.4	5.5	0.5	385
1	-34.8	7.9	8.7	1.0	658
1	-34.8	3.3	15.4	1.7	872
3	-35.5	4.0	35.1	9.7	1776
5	-38.2	1.9	49.8	17.5	2547
7	-38.0	2.0	62.2	19.3	3329
10	-41.2	3.6	78.4	30.5	4501
13	-41.2	6.3	92.9	43.8	7083
16	-41.3	3.1	105.9	51.8	4828
20	-38.3	1.0	121.5	55.0	5869

Table 4.4: This table shows the results for data set 1 obtained using 30,000 genealogies simulated under a scheme where the likelihood of the sample configuration after each event is not used. The likelihoods do not appear to be well estimated, the pattern observed between different values of ρ describes an implausible likelihood surface and the estimated values of the ESS are very low (and likely to be significantly over estimated). The reason that this scheme performs so badly is related to the average number of simulated recombination events in each genealogy. The scheme proposes very large numbers of recombination events, even when the underlying recombination rate is very low. This shows that the genealogies simulated are not a good approximation to the true conditional distribution of ARGs for this data.

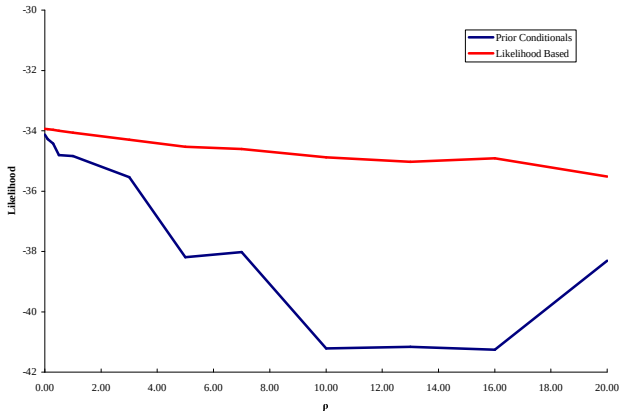

Figure 4.12: This figure shows the likelihood curves generated by using the prior rates of events when importance sampling (in blue) and using the likelihood of the resulting sample configuration (in red). The blue curve seems not to have converged as the shape seems unlikely. The red curve shows roughly consistent estimates on all 20 independent values of ρ calculated. It is possible to tell from the red curve that this data provides no evidence for recombination. The total time to produce these results was 3.5 hours for the red curve and 5.5 hours for the blue curve. However the program was not optimised for the approach which only used the prior rates of events and some further gains could probably be made if dedicated software were used.

4.3.3 Discussion

The data sets analysed here used only 10 sequences and all of the methods struggled when the number of segregating sites approached 20, even for days of computing time (using one processor). Although such data could be comfortably analysed using a collection of computers or with a more powerful processor it is close to the limit of this approach. Unfortunately modern data sets are usually far bigger than those considered here and an importance sampling approach would be completely impractical for most applications.

The methods I believe would be most effective at increasing the efficiency of importance sampling methods would focus on improving the conditional likelihoods used to calculate the weight of each of the possible next events in each epoch. The current approaches ignore much of the genealogical information in the data, these ideas are discussed more fully in Chapter 2. This lack of genealogical interpretation leads to underestimating the full importance of a recombination event that removes an incompatibility in the data - when the recombination rate is low these events will not be proposed often enough. When the recombination rate is low incompatibilities are confused with mutation events. This leads to the schemes underestimating the impact that incompatibilities have on the likelihood of a sample configuration (see Figure 4.13). Also even when the schemes recognise that it is essential to

simulate recombination events there can still be considerable confusion about the best position for this event and the consequences of each possible recombination, some intuitive examples are given in Figure 4.14. Although these examples assume a low recombination rate to clearly expose the inaccuracies of the likelihood approximations, these sorts of problem can be observed to cause serious inaccuracies in the distribution of genealogies given the data. Tackling these issues and providing a closer approximation to the coalescent likelihoods may well be the key to creating powerful importance samplers.

One approach to including this information would be to use summaries of incompatibility in the data, as this is perhaps the strongest evidence for recombination in the data. When an event under consideration reduces the estimated minimum number of recombination events required to explain the data this event could be up-weighted; this would reflect the increase in likelihood of the resulting data missed by the conditional approximations.

Another approach might be to try to characterise the simulations that led to very high importance weights. The quotient $P(G)/Q(G)$ is high for these genealogies (that is, the genealogy was relatively likely under the prior, compared to the proposal distribution). It could be useful to understand what aspects of these genealogies are unlikely under $Q(.)$ when in fact they are genealogies with high weight under the prior that are compatible with the data. Understanding these anomalies and appropriately altering the pro-

Figure 4.13: This diagram represents a toy data set which illustrates the inability of the $\pi_{L\&S}$ scheme to recognise the importance of incompatibilities in data. The data here contains an incompatibility and so at some point in the history must undergo a recombination event. Under a small value of ρ genealogies for this data will undergo the two specific events before any others are likely occur. One is a coalescence event between the two identical sequences, another is a recombination event that on any unique type results in data with no incompatibility. Under the schemes described here the coalescence event gets a much higher weight and is chosen much more often. However under the true posterior distribution recombinations and coalescence events are roughly equally likely. The reason for the disparity is that the approximation $\pi_{L\&S}$ (or any of the approximations in Chapter 2) do not give sufficiently low likelihoods for the data which is incompatible. As explained in Chapter 2 the schemes confuse these incompatibilities with repeat mutations and when the recombination rate is significantly different to the mutation rate this causes very inaccurate likelihood estimates. An interesting side effect of using the Fearnhead approximation mentioned earlier in this chapter is that the ESS is much greater. This is because the recombinant sequences are chosen at random and, once such a sequence is chosen, a recombination event must occur. This may be an example of why this approximation manages to perform well.

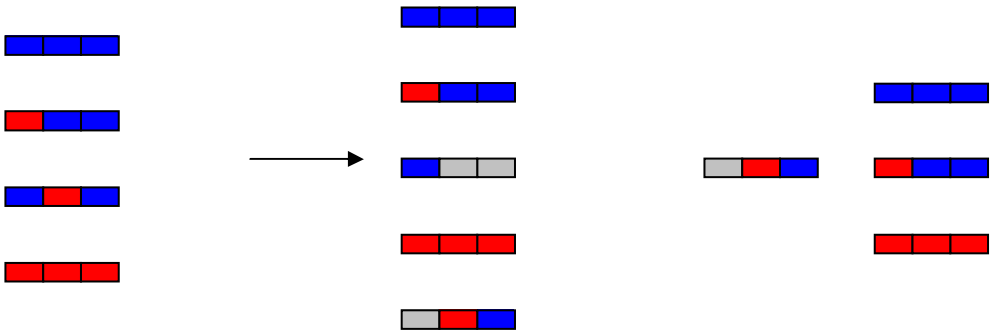

Figure 4.14: The diagram on the left represents a simple data set which can be used to illustrate some subtle problems with the $\pi_{L\&S}$ approximation. The first inaccuracy created in using these approximations is that, when the recombination is very low, recombinations are simulated similarly often at the first gap between sites as the second. However, with a low recombination event only genealogies with one recombination event should commonly be sampled. A recombination event *must* occur in the first gap between sites but none is needed in the second. The inaccuracy arises due to the approximate schemes allowing repeat mutation events which generate incompatibilities without having to invoke recombination events. When the recombination rate is much lower than the mutation rate this can strongly affect likelihood approximations. The second interesting feature in this data set is that even with low mutation (and recombination) rates the frequency of recombination events on the *third* sequence at the first site is very low. While the frequency of recombination events on the other sequences at the first gap (only) is high. This arises due to the approximation of the probability of the data given a recombination event at the first site on the third sequence (the resulting configuration is the diagram in the middle, where grey denotes non ancestral material). The key term in this calculation is represented pictorially on the right hand side. The probability of the single sequence given the three that remain. This is very low using $\pi_{L\&S}$ as every path in the dynamic programming algorithm requires either a recombination of repeat mutation to generate the single sequence given the others. Less formally $\pi_{L\&S}$ perceives the need for an extra mutation or recombination in order to explain the presence of this sequence given that those three exist in the population. However, under the full model this sequence can be explained without the need for further recombination or mutation.

posal distribution could be used to reduce the variance in the distribution in the importance weights which would improve the convergence times of the importance sampler.

Other possible approaches to increasing the scale of inference include using the genealogies to construct approximate methods of inference. For example the data could be split into smaller regions, such as in Fearnhead and Donnelly [38]. The time it would take to properly analyse 50 regions each with 10 segregating sites in is extremely small compared to the time it would take to analyse a single data set 500 segregating sites. Alternatively the weights of the genealogies could be discarded and quantities of interest could be averaged over the distribution of un-weighted genealogies, which approximate the true distribution. There are many other possibilities and a careful analysis into the properties of such methods for specific applications of interest would be required to assess performance.

In summary: current full likelihood methods are severely restricted in the size of data sets that they can be used to analyse. However, understanding the problems with the current methodology could create significant improvements in performance. Also, using these methods to perform approximate inference could potentially save considerable computing time. Unfortunately even such improvements and approximations currently seem highly unlikely to allow full genealogical inference to be applied to modern large scale vari-

ation data sets. These fully genealogical approaches are most likely to be useful when only small regions of the genome are under consideration. In these cases it will normally be necessary to model the recombination process and genealogical approaches are likely to provide the most accurate inference.

Chapter 5

Discussion

5.1 Introduction

In this thesis I have discussed various statistical models for genetic data in the presence of recombination, how these methods can be used for inference about population parameters and the calculation of likelihoods for a sample of population genetic data. The focus of this thesis is restricted to two types of model for genetic variation within a population subject to recombination. The first type of model is based on the notion of 'copying' where new sequences are constructed as imperfect mosaics of preexisting sequences. The second type of model uses the notion of a genealogy which describes the ancestry and relationships of the sampled sequences.

5.2 Using a Product of Approximate Conditionals

Building on the work of Stephens and Donnelly [20] Fearnhead and Donnelly [21] constructed a copying scheme that was able to mimic the effects of mutation and recombination. Li and Stephens [22] took this model and created a faster, simplified version as a direct approach to calculating the likelihood through the product of these conditional likelihoods for all of the sequences in the sample.

This approach has been extremely successful due to the small computing time required to calculate the likelihood under this model. Although the model was introduced for estimating recombination rates it is much more flexible than previous methods for approximating the likelihood. Unlike the use of summary statistics or composite likelihood methods the PAC model can be directly used for a wide range of applications such as phasing genotype data [29], simulating population data, elucidating population structure and imputing missing data. The success of this approach has also led to the development of even faster algorithms, such as that designed by Stephens and Scheet [23], although this approach does not use an explicit recombination parameter, so it is not directly comparable with these approaches.

In this thesis I do not focus on the speed of various PAC approaches but

on their properties, in particular, their accuracy when estimating a constant recombination rate. I have compared four alternative models, including two novel models (π_R and π_{L^2}) and reported their individual performance as well as the relative success of each model at estimating the recombination rate. There is some evidence that improvements can be made with improved conditional schemes, and perhaps some improvement has been found in the scheme π_{L^2}. However in the course of the investigation it became apparent that there existed some fundamental problems with all of the approaches tried here.

The PAC approach, whilst able to distinguish between various degrees of recombination, does not provide unbiased recombination rate estimates. The schemes investigated here all estimate a non-zero recombination rate in a high percentage of cases where the data were simulated with no recombination. Also recombination rate estimates are biased downwards when the recombination rate is very high. These biases are complex in nature and vary with the number of segregating sites in a region of fixed length, or equivalently, with the average distance between segregating sites. Li and Stephens provide an empirical bias correction for data with a constant recombination rate. However, when recombination rates are allowed to vary this correction may not lead to unbiased estimates.

By examining the schemes in more detail, and by using genealogical ap-

proaches as a gold standard, it is possible to identify certain features of the data that are incorrectly interpreted by PAC schemes. As a result a greater number of mutation or recombination events are sometimes required to explain the data than in the genealogical case. Similarly PAC models confuse the signal for recombination with mutation which can result in an underestimate of the recombination rate. These effects also make a broad bias correction term inadequate to provide completely accurate inference - this only affects the *average* estimate of the recombination rate.

There have been attempts to improve the choice of orderings used by the PAC schemes to reduce the effect of spurious recombination signals being inferred, although no such work has yet been published. Altering the order in which sequences are considered, or other more severe changes to the schemes designed to deal with various signals for recombination are not sufficient to overcome these difficulties. In most cases there will exist no ordering or sequence sampling scheme which will does not infer unnecessary events when generating sequences given the others in the sample.

Many of the problems suffered by the copying models of sequence evolution stem from the fact that the true evolutionary process has a complex structure in which sequences change through time. This induces an ordering on events that is not reflected in copying processes, as discussed in Chapter 2. The time ordering of events can lead to sequences which, although differ-

ent to others in the sample, can be explained (on examination of these other sequences) without inferring further mutation or recombination events. If a copying approach could be designed which was able to incorporate such aspects of the ancestry while retaining computational efficiency it might provide a means of performing much more accurate inference on recombinant data.

One possible approach to this is to make the following distinctions. Consider a sample of k haplotypes and the probability of observing a $k + 1^{\text{th}}$ haplotype, h_{k+1}, that is distinct from all of the first k at some locus. Then h_{k+1} can be viewed as being derived from these k haplotypes with

1. A novel mutation

2. A repeat mutation or recombination

3. No mutations or recombinations.

For examples of these three situations see Figure 5.1. Under the PAC approach it is impossible to distinguish between situations 2 and 3 in the list above and this can lead to false signals for recombination. One approach to distinguishing these situation would be to use estimates of $R_{\min}$ which try to count the number of events of type 2 in the sample. The most accurate method that does not require the simulation of genealogies is the bound R_h introduced by Myers and Griffiths [34], although more accurate, although

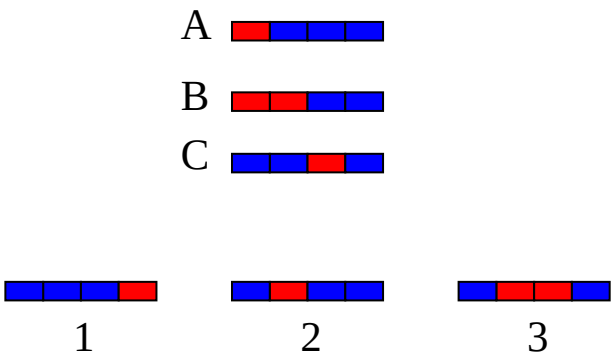

Figure 5.1: This diagram shows 3 different possibilities that might arise when calculating the probability of new haplotype, h, given a pre existing set of sequences (with none the same as H). Haplotype 1 contains a novel mutation, these are trivial to observe. Haplotype 2 can be explained through a time ordering effect (that requires no extra mutation) (eg. see Figure 2.10 from Chapter 2). Haplotype 3 creates an incompatibility in the data that was not previously present.

computationally intensive, bounds have been proposed by Song and Hein [44]. This method can also be used to (partially) localise incompatibilities. When a new sequence is added to the sample, if R_h is increased then this is evidence for recombination and so recombination should be simulated within the algorithm. When R_h does not increase this is evidence (but as R_h is imperfect there is no certainty) that a time ordering effect can explain some of the differences in the new sequence: so these differences should be treated in a different way.

Another possible approach would be to construct hybrid haplotypes from those in the sample representing ancestral individuals that could be 'copied' from. This gives a more natural, and therefore potentially more biologically

meaningful, interpretation of the copying process. Combinatorial approaches to manufacturing such ancestral sequences have been proposed [45], but much development is required. Modelling of underlying biological processes and a statistical approach including the notion of likelihood would be required to make use of such ideas in this setting. Alternatively, a graph-like approach with the extant lineages as internal nodes, such as that developed by Fitch in 1977 [46], could be extended with directional edges denoting the direction of copying and intermediate sequences from which other sequences could copy.

5.3 Genealogical Models

The second major theme of this thesis was to reduce the computational burden of performing genealogical inference in the presence of recombination. I describe a new model, the SMC, of ancestry which is based on the coalescent and which is identical to the coalescent in the absence of recombination [36]. The SMC produces provides a Markovian structure when simulating genealogies along a sequence and reduces the overall state space of ARGs. In this thesis I compare the coalescent and the SMC, both in terms of the properties of the models, as well as the impact that the simplified structure of the SMC has on performing inference using importance sampling.

When importance sampling is used to calculate the likelihood of small

data sets the SMC and the coalescent perform similarly for small amounts of recombination. When the recombination rate increases there is a systematic decrease in the average number of recombination events under the SMC which leads to genealogies being simulated faster and to lower variation in likelihood estimates. Unfortunately this effect starts to become important only when the recombination rate is high, usually this arises in data so large that inference under both models is usually very challenging or impossible. Only in organisms, or genomic regions, where the recombination rate is very high compared to the mutation rate will this simple use of the SMC provide truly significant gains in efficiency. Another possibility for improving inference is that the sequential form of the SMC could be used, perhaps to construct an MCMC scheme where local genealogies were updated conditional on the two immediately adjoining trees. This would be a valid approach within the SMC because of the Markov property when genealogies are viewed along a sequence. However, it was found to be challenging to correctly condition on neighbouring genealogies when performing an update.

The importance sampling schemes used here were unable to perform accurate inference on data sets with a large number of segregating sites. Due to the reduced cost of genotyping it is now common for data to contain many segregating sites. This limitation makes the impact of improvements in full likelihood methods questionable. However, there are both theoretical

and practical reasons why such schemes may be further pursued. Firstly, note that improving the methods for calculation of approximate likelihoods has the potential to revolutionise importance sampling. Under the optimal importance sampler only one genealogy need be simulated to calculate the likelihood. Using the ideas in Chapter 2 and above in this discussion it may be possible to provide closer approximations to these optimal conditionals. However, how the performance changes as the importance sampler *approaches* the optimal sampler has not been explored here, and is perhaps an important question.

On a more practical level, approximate methods may be constructed using full likelihood methods. Composite likelihood approaches (see eg. McVean et. al. [14] or Fearnhead and Donnelly [38]) use likelihoods constructed from multiple subsets of segregating sites and combine the results to provide inference for large regions. These approaches are designed to infer recombination rates however it may be possible to apply the same approaches to a range of population genetic questions when recombination is present.

5.4 Summary

Modern genetic data can help to provide account of the history of populations. Using this we can learn about the biological processes that change

individual organisms through time. This information can be helpful in understanding underlying biological mechanisms that give rise to different phenotypes, and even disease in human populations. Unfortunately, the processes that give rise to this data are highly random and the data that we observe does not readily yield the information that we desire.

Statistical models have the potential to account for random elements, and to pinpoint quantities of interest within the data. However, designing accurate models which allow the construction of computationally efficient algorithms to perform these inferences remains a challenging task. The coalescent with recombination directly models the ancestry of a sample, and so provides the opportunity to distinguish between different forces and processes affecting the data. It is also theoretically possible to estimate parameters of interest, such as the mutation and recombination rates. However performing efficient inference under the coalescent has proved to be extremely challenging.

Approximations to the coalescent have provided a compromise between accuracy and efficiency. While there are currently no models that can truly claim to have found the perfect balance between simplicity and effectiveness, there is much evidence for progress. By understanding the shortcomings of the methods of today it may be possible to design faster and better methods for understanding the data of tomorrow.

Bibliography

[1] J. F. C. Kingman. The coalescent. *Stochastic Processes Appl.*, 13:235–248, 1982.

[2] R R Hudson. Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*, 23(2):183–201, Apr 1983.

[3] G. Grimmet and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, third edition, 2002.

[4] J K Pritchard, M Stephens, and P Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, Jun 2000.

[5] Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, Aug 2003.

[6] Nick Patterson, Neil Hattangadi, Barton Lane, Kirk E Lohmueller,

David A Hafler, Jorge R Oksenberg, Stephen L Hauser, Michael W Smith, Stephen J O'Brien, David Altshuler, Mark J Daly, and David Reich. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet*, 74(5):979–1000, May 2004.

[7] Michael W Smith, Nick Patterson, James A Lautenberger, Ann L Truelove, Gavin J McDonald, Alicja Waliszewska, Bailey D Kessing, Michael J Malasky, Charles Scafe, Ernest Le, Philip L De Jager, Andre A Mignault, Zeng Yi, Guy De The, Myron Essex, Jean-Louis Sankale, Jason H Moore, Kwabena Poku, John P Phair, James J Goedert, David Vlahov, Scott M Williams, Sarah A Tishkoff, Cheryl A Winkler, Francisco M De La Vega, Trevor Woodage, John J Sninsky, David A Hafler, David Altshuler, Dennis A Gilbert, Stephen J O'Brien, and David Reich. A high-density admixture map for disease gene discovery in african americans. *Am J Hum Genet*, 74(5):1001–1013, May 2004.

[8] E S Lander and P Green. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A*, 84(8):2363–2367, Apr 1987.

[9] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, 2003.

[10] R.A. Fisher. The correlation between relatives on the supposition of

mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[11] P.A.P. Moran. Random Processes in Genetics. *Proc. Camb. Phil. Soc.*, 54:60–72, 1958.

[12] S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.

[13] S. Wright. Size of population and breeding structure in relation to evolution. *Science*, 87:430–431, 1938.

[14] Gil McVean, Philip Awadalla, and Paul Fearnhead. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160(3):1231–1241, Mar 2002.

[15] Gilean A T McVean, Simon R Myers, Sarah Hunt, Panos Deloukas, David R Bentley, and Peter Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–584, Apr 2004.

[16] A. J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, 29(2):217–222, 2001.

[17] R C Griffiths and P Marjoram. Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol*, 3(4):479–502, Winter 1996.

[18] M K Kuhner, J Yamato, and J Felsenstein. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, 140(4):1421–1430, 1995.

[19] M K Kuhner, J Yamato, and J Felsenstein. Maximum likelihood estimation of recombination rates from population data. *Genetics*, 156(3):1393–1401, Nov 2000.

[20] M. Stephens and P. Donnelly. Inference in molecular population genetics. *Journal of the Royal Statistical Society B*, 62:605–655, 2000.

[21] P. Fearnhead and P. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159:1299–1318, 2001.

[22] N. Li and M. Stephens. Modelling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 2003.

[23] Matthew Stephens and Paul Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet*, 76(3):449–462, Mar 2005.

[24] Michael P H Stumpf and Gilean A T McVean. Estimating recombination rates from population-genetic data. *Nat Rev Genet*, 4(12):959–968, 2003.

[25] Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78(4):629–644, Apr 2006.

[26] E. Kreyszig. *Advanced Engineering Mathematics*. John Wiley and Sons, 1998.

[27] R R Hudson. Two-locus sampling distributions and their application. *Genetics*, 159(4):1805–1817, 2001.

[28] J D Wall. A comparison of estimators of the population recombination rate. *Mol Biol Evol*, 17(1):156–163, Jan 2000.

[29] Matthew Stephens and Peter Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, 73(5):1162–1169, Nov 2003.

[30] Susan E Ptak, Amy D Roeder, Matthew Stephens, Yoav Gilad, Svante Paabo, and Molly Przeworski. Absence of the TAP2 human recombination hotspot in chimpanzees. *PLoS Biol*, 2(6):e155, Jun 2004.

[31] Pierre Nicolas, Fengzhu Sun, and Lei M Li. A model-based approach to selection of tag SNPs. *BMC Bioinformatics*, 7:303, 2006.

[32] Jonathan Marchini, David Cutler, Nick Patterson, Matthew Stephens, Eleazar Eskin, Eran Halperin, Shin Lin, Zhaohui S Qin, Heather M Munro, Goncalo R Abecasis, and Peter Donnelly. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet*, 78(3):437–450, Mar 2006.

[33] R R Hudson, M Slatkin, and W P Maddison. Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132(2):583–589, Oct 1992.

[34] Simon R Myers and Robert C Griffiths. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163(1):375–394, Jan 2003.

[35] C. Wiuf and J. Hein. Recombination as a point process along sequences. *Theoretical Population Biology*, 1999.

[36] Gilean A T McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*, 360(1459):1387–1393, Jul 2005.

[37] W Hill and A R Robertson. Linkage disequilibrium in finite populations. *Theoretical Population Biology*, 38:226–231, 1968.

[38] P. Fearnhead and P. Donnelly. Approximate likelihood methods for estimating local recombination rates. *JRSS (B)*, 64:657–680, 2002.

[39] Paul Fearnhead, Rosalind M Harding, Julie A Schneider, Simon Myers, and Peter Donnelly. Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics*, 167(4):2067–2081, Aug 2004.

[40] Jeffrey D Wall. Estimating recombination rates using three-site likelihoods. *Genetics*, 167(3):1461–1473, Jul 2004.

[41] J.S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.

[42] G. Watterson. On the number of segregating sites. *Theoretical Population Biology*, 7:256–276, 1975.

[43] Liu Kong and Wong. *Topics in Monte Carlo Inference*. Unknown as yet, 1999.

[44] Yun S Song and Jotun Hein. Constructing minimal ancestral recombination graphs. *J Comput Biol*, 12(2):147–169, Mar 2005.

[45] E. Ukkonen. Finding founder sequences from a set of recombinants. *Algorithms in Bioinformatics (WABI-2002), Lect. Notes in Computer Science*, 2452:277–286, 2002.

[46] W. M. Fitch. On the Problem of Discovering the Most Parsimonious Tree. *The American Naturalist*, 111:223–257, 1977.