The Likelihood of Gene Trees under Selective Models

Graham M. Coop

Lady Margaret Hall

A thesis submitted to the Faculty of Mathematical Sciences

for the degree of Doctor of Philosophy of the University of Oxford.

Department of Statistics

University of Oxford

Trinity Term, 2004

The Likelihood of Gene Trees under Selective Models

Graham M. Coop, Lady Margaret Hall

D.Phil Thesis, Trinity Term, 2004

Abstract

The extent to which natural selection shapes diversity within populations is a key question for population genetics. Thus, there is considerable interest in quantifying the strength of selection. In this thesis a full likelihood approach for inference about selection at a single site within an otherwise neutral fully-linked sequence of sites is developed. Integral to many of the ideas introduced in this thesis is the reversibility of the diffusion process, and some past approaches to this concept are reviewed. A coalescent model of evolution is used to model the ancestry of a sample of DNA sequences which have the selected site segregating. A novel method for simulating the coalescent with selection, acting at a single biallelic site, is described. Selection is incorporated through modelling the frequency of the selected and neutral allelic classes stochastically back in time. The ancestry is then simulated using a subdivided population model considering the population frequencies through time as variable population sizes. The approach is general and can be used for any selection scheme at a biallelic locus. The mutation model, for the selected and neutral sites, is the infinitely-many-sites model where there is no back or parallel mutation at sites. This allows a unique perfect phylogeny, a gene tree, to be constructed from the configuration of mutations on the sample sequences. An importance sampling algorithm is described to explore over coalescent tree space consistent with this gene tree. The method is used to assess the evidence for selection in a number of data sets. These are as follows: a partial selective sweep in the G6PD gene (Verrelli et al., 2002); a recent full sweep in the Factor IX gene (Harris and Hey, 2001); and balancing selection in the DCP1 gene (Rieder et al., 1999). Little evidence of the action of selection is found in the data set of Verrelli et al. (2002) and the data

set of Rieder et al. (1999) seems inconsistent with the model of balancing selection. The patterns of diversity in the data set of Harris and Hey (2001) offer support of the hypothesis of a full sweep.

Acknowledgements

I would like to thank my supervisor Prof. R. C. Griffiths for his support, patient editing, advice and encouragement throughout my D. Phil and the writing of this thesis. Everyone in the group and wider department has always been very friendly and helpful, particularly the IT and support staff. Dr. Simon Myers, Dr. Gil McVean and Chris Spencer have been of great help and I thank them for their insights, discussion and advice during this work. I would also like to thank my office mates Simon, Danny, Chris, Don and Niall and for their friendship, advice and constant welcome distraction over my time here. Yvonne Griffiths and Muireann Ocinneide have been of great help in proof reading various chapters. Thanks to the Engineering and Physical Sciences Research Council for their financial support through this project. Finally, I thank my Parents, my sister Jenny and Hannah for their continual support both academically and emotionally.

Graham Coop

Oxford, England

July, 2004.

Contents

1 Introduction **2**

2 Models Arising in Population Genetics **10**

2.0.1 Models of Selection 11

2.0.2 Mutational Models 14

2.1 Discrete State Space Models 15

2.1.1 Moran Model 15

2.1.2 Wright-Fisher Model 17

2.2 The Diffusion Approximation 19

2.2.1 The Scale Function and the Speed Measure 21

2.2.2 The Generator of a Diffusion 22

2.3 The Coalescent 25

2.3.1 The Coalescent and Variable Population Size 27

2.4 Gene Trees 28

2.5 Previous Approaches to Modelling the Coalescent with Selection 34

2.5.1 Ancestral Selection Graph (ASG) 35

2.5.2 Expectations of the Coalescent and Path Integrals 37

2.6 Explicit Trajectory Models 40

2.6.1 Deterministic Selection 40

3 The Wright-Fisher Diffusion Process and Reversibility 42

3.1 Properties of an Allele Segregating in the Population 42

3.2 Properties of an Allele Segregating in a Sample 57

3.3 The Moran Model and Reversibility 61

4 Simulation of Coalescent Genealogies with Selection 68

4.1 The Coalescent and the Moran Model 68

4.2 Simulating a Coalescent History 69

4.3 Extensions 72

5 Importance Sampling, Selection and the Coalescent 74

5.1 Monte Carlo Methods 74

5.2 Importance Sampling 75

5.2.1 The Likelihood of a Gene Tree 76

5.2.2 History States of a Neutral Gene Tree 79

5.2.3 History States of a Gene Tree Conditional on a Trajectory . . . 82

5.3 Importance Sampling Conditional on a Trajectory 84

5.3.1 Importance Sampling Within a Subsample 86

5.3.2 Removal of the Selected Mutation 86

5.3.3 The Trajectory Starting Frequency 88

5.3.4 Computational Features 91

5.4 Convergence 93

6 Data Analysis **101**

6.1 G6PD and a Partial Sweep 101

6.2 The Factor IX Region and a Model of a Full Sweep 104

6.3 DCP1 and Balancing Selection 113

7 Conclusion **117**

List of Figures

2.1 An illustration of the Moran model 16

2.2 Moran model trajectories for various strengths of selection 18

2.3 An example data set 31

2.4 The gene tree for the example data set 32

2.5 A possible coalescent tree for the example data 32

2.6 An example Ancestral Selection Graph 37

2.7 An example deterministic trajectory and stochastic trajectories 41

3.1 The population frequency spectrum for different selection coefficients . 45

3.2 An illustration of a trajectory forward and backward in time 48

3.3 The frequency spectrum of a derived mutation conditional on observing a sample of 10 containing 5 copies of the selected allele 62

3.4 The frequency spectrum of a derived mutation conditional on observing a sample of 10 containing 3 copies of the selected allele 62

3.5 The frequency spectrum of a derived mutation conditional on observing a sample of 10 containing 7 copies of the selected allele 63

3.6 Frequency spectrums of a neutral derived mutation conditional on observing a variety of samples 63

3.7 Moran model trajectories backwards in time for a variety of selection coefficients 67

5.1 A set of four trees showing the original gene tree and three permissable moves and their notation 80

5.2 A simple gene tree 95

5.3 A comparison of naive simulation estimates to those of the importance sampling algorithm 96

5.4 An example gene tree 97

5.5 Six likelihood surfaces for β for the example gene tree 98

6.1 Log likelihood surface for the G6PD data set 102

6.2 The G6PD gene tree with estimated times 106

6.3 The FIX gene tree 107

6.4 Caricature of the genealogy and the population trajectory of the selected allele 108

6.5 Joint log likelihood surfaces for the FIX gene tree for different θ 110

6.6 Joint log likelihood surface for $\theta = 3.5$ for the FIX gene tree 112

6.7 The DCP1 gene tree 114

6.8 Joint log likelihood surface for the DCP1 gene tree 116

List of Tables

2.1 The relative fitnesses of the different genotypes at a biallelic locus . . . 12

2.2 The binary incidence matrix for the example data set 31

5.1 The importance sampling scheme when both subtrees still have events to perform 87

5.2 The importance sampling scheme when one or both subtrees have no more events to perform. 89

5.3 The average variance of the estimate likelihood within a trajectory . . . 100

5.4 The mean and variance of the estimate of the likelihood of the trajectory 100

6.1 Summary of the age estimates for the African G6PD gene tree 105

Chapter 1

Introduction

The role that natural selection plays in evolution has been subject to intense debate since the publishing of the Origin of the Species (Darwin, 1859), and the question is still one of the most important in evolutionary biology. This question has been approached in a wide variety of ways, from 'in the field' experiments to find the reproductive advantage or increased probability of survival due to a particular trait to the analysis of the long term evolution of genes. A common approach is to examine the role that natural selection has played in the recent evolution of a population. This has the potential to elucidate the genome-wide pattern of selection within a species (Bustamante et al., 2002) and to identify regions where positive selection has acted recently, and hence locate genes that have been the focus of recent evolution, see for example Akey et al. (2002). The role of selection in human populations is of interest, e.g. Bamshad and Wooding (2003), for diverse reasons as understanding the role that selection has had in the prevalence of some genetic diseases (Diamond, 2003) and in human disease resistance (Tishkoff et al., 2001) to identifying regions important in the evolution of the facility for language (Enard et al., 2002). Patterns of genetic diversity within a population around a locus of known biological importance are often investigated to examine the possible role of natural selection in the evolution of the locus. In this thesis we

concentrate on developing a method to assess the strength of selection that has acted at a single site.

A wide variety of approaches to detect the action of natural selection have been developed over the years and a commonly used approach is that of summary statistics (see Nielsen (2001b) for a review). Methods for examining the role of selection can be divided into two broad groups. The first set of approaches are designed to detect non-neutral behaviour. This is achieved by taking a summary of the data and seeing how likely this summary statistic is under the standard neutral model using either analytical results or simulation. Commonly used are a family of frequency spectrum statistics, see for example Tajima (1989) and Fu and Li (1993). To detect adaptive or purifying selection the ratio of synonymous to nonsynonymous codon substitutions is often examined. Adaptive selection acting within one population and not another will increase the differentiation between the two populations and this has been the focus of a number of papers (Lewontin and Krakauer, 1973; Akey et al., 2002; Beaumont and Balding, 2004). The level of diversity between species and divergence within a population under neutrality should be correlated and a test of this, based on the comparison of multiple loci, was developed by Hudson et al. (1987). A positively selected allele will quickly propagate through the population allowing little time for recombination to act. The allele therefore should show a high degree of correlation with the haplotype on which it arose. This fact is exploited by methods that take as the test statistic a summary of the patterns of haplotype sharing away from a putative selected site (Hudson et al., 1994; Sabeti et al., 2002).

The second set of approaches to detecting selection explicitly model selection. This set of approaches can be further divided into moment estimators and likelihood or Bayesian approaches. Moment estimators fit the parameters of a nonneutral model by match-

ing the summary of the data to the expectation of that statistic under the nonneutral model, see for example Hudson and Kaplan (1988) and Wiehe (1998). In likelihood or Bayesian methods, the likelihood or posterior probability of the parameters of the nonneutral model is calculated. These include methods for quantifying the ratio of synonomous to nonsynonomous mutation rates at various sites (Yang et al., 2000) and for assessing the distribution of selective effects within a population (Bustamante et al., 2003). When examining the role of selection acting on a particular site the effect of selection on the genealogy and hence the neutral diversity is critical. Therefore a coalescent approach where selection acts to change the shape and the time in various parts of the genealogy is appropriate. Various methods to perform inference about selection in a coalescent setting have recently been developed. Examples are the summary statistic approaches of Slatkin (2001) and Przeworski (2003) and the pairwise composite approach of Kim and Stephan (2002).

A number of authors (see, for example, Rieder et al. (1999); Fullerton et al. (2000); Saunders et al. (2002); Verrelli et al. (2002)) have used the full likelihood gene tree approach of Griffiths and Tavaré on sequence data from a locus believed to have experienced natural selection. The estimated age of the candidate selected mutation is examined for inconsistency with the neutral model. However, the times are biased towards their values under neutrality as the coalescent model used does not include selection. This comparison of ages approach has been developed into a test for negative selection by comparing the ages of all nonsynonymous sites to those of synonymous sites (Nielsen and Weinreich, 1999; Nielsen, 2001a). Another method to assess the time of the most recent common ancestor in samples undergoing a selective sweep is the use of a star-like-tree. Under the assumption of a star-like-tree all the time in branches other than singletons is assumed to be negligible (see for example Hamblin and Di Rienzo (2000)). This approach can incur biases and underestimate the confidence in-

terval (Rosenberg and Hirsh, 2003).

In this thesis, the history of a sample of genes together with the population history of the frequency of the selected allele is considered. Our approach is similar to that of Slatkin (2001) in that it uses a fully stochastic trajectory generated backwards in time, and then generates a genealogy given this trajectory. By dealing with selection in a fully stochastic framework instead of a deterministic one, weaker selection can be dealt with. A weakly selected mutation might not allow neutrality to be rejected. However, if *a priori* it is known to have some biological effect, then its age and an indication of its selection coefficient are still of considerable interest. A full likelihood coalescent inference method for a single selected site in an otherwise neutral non-recombining sequence is described in this thesis, and in Coop and Griffiths (2004). The joint likelihood curve of the selection parameter and the distributions of the times at which the mutations occurred and the time to the most recent common ancestor can be found by this method, allowing a better understanding of the history underlying the data. The method is flexible and can be extended to any type of selection acting at a biallelic locus.

Throughout this thesis, a number of models are utilised to describe the evolution of a population and these are introduced in Chapter 2. Various simple models of different types of selection are given. The Moran and Wright-Fisher population models and their large population limit the Wright-Fisher diffusion process are introduced, and the coalescent process is briefly described. Also reviewed in this chapter are previous approaches to the simulation of the genealogy of a selected mutation and methods to evaluate expectations of quantities of interest relating to this genealogy.

A key idea in the approach to selection taken in this thesis is time reversibility of the

processes describing the frequency of the selected allele. A number of authors have considered this reversibility in a variety of ways. A review of some of these, with particular reference to the diffusion process, is given in Chapter 3. Applications of reversibility used in this thesis are also reviewed in detail in this chapter.

A new method for the simulation of a genealogy of a single biallelic site at which selection acts is given in Chapter 4. The frequency of the selected allele in the population (the trajectory of the selected allele) is explicitly modeled. This method utilises various reversibility results to simulate the frequency of the selected allele back through time. Previous approaches to trajectory simulation have either ignored the fluctuations in frequency, and used deterministic approximations, or used rejection or importance sampling to simulate the trajectory of the allele back from its current frequency. When simulating a history, for a site segregating in the population, the fact that the frequency of the allele in the current day is often only known in a sample must be accounted for. An approach introduced in this thesis is the simulation of a random current day frequency, the distribution of which is derived in Griffiths (2003). The trajectory may be simulated backwards in time from this current day frequency to when the selected allele is lost from the population. Given the trajectory, the genealogical process may be treated as a two deme subdivided population with varying population sizes determined by the trajectory (Kaplan et al., 1988). Extensions to the simulation method are also given in this chapter. The extension of the simulation method to allow the generation of genealogies for a recombining sequence featuring a selected site has been implemented by a coworker and myself (Spencer and Coop, 2004).

Under the assumption of an infinitely-many-sites mutation model and no recombination the data may be fully and uniquely represented by a perfect phylogeny, a gene tree (see for example Griffiths (2002)). This representation is useful for inference as it provides

a partial ordering to the events in the past genealogical history of the sample. The likelihood of the gene tree can be expressed as the integral over the missing genealogical history underlying the gene tree (see Stephens (2001) for an introduction). In the case of a gene tree featuring a selected allele two levels of missing data can be introduced, the first is the trajectory of the allele back through time and the second is the genealogical history given this trajectory. An importance sampling algorithm is used to integrate over possible trajectories and histories consistent with the data to find the likelihood with respect to the selection coefficient of the allele. The method belongs to a family of importance samplers developed by Griffiths and Tavaré (1994a) to deal with full information in both the finite and infinitely-many-sites models of mutation. The method has been extended to variable population size models (Griffiths and Tavaré, 1994b), subdivided populations (Bahlo and Griffiths, 2000; De Iorio and Griffiths, 2004b; De Iorio et al., 2004), and recombination (Griffiths and Majoram, 1996; Fearnhead and Donnelly, 2001). The efficiency of the importance sampling algorithm was improved by Stephens and Donnelly (2000) and a general importance sampling framework based on an approximation of the diffusion generator is developed by De Iorio and Griffiths (2004a). Felsenstein and colleagues have used Markov Chain Monte Carlo (MCMC) techniques to examine similar problems, see for example Felsenstein et al. (1999). Full likelihood MCMC methods have also been developed by Drummond et al. (2002) and Wilson et al. (2003) among others. In the importance sampling framework the history is a Markov chain on the possible history states embedded in the genealogical history of the sample. The concept of history states in a neutral gene tree is discussed in Chapter 5. Then the history states of a gene tree, given a trajectory, are described and the generator and recursion of the Markov chain for the history states is introduced in this thesis. An importance sampling algorithm for sampling trajectories and histories compatible with the gene tree is presented. The algorithm incorporates ideas developed in Chapter 4, and the infinitely-many-sites importance sampler of Stephens and Donnelly (2000).

Three main types of applications of the gene tree method described in this thesis are envisaged. These applications are given by the cases to consider in a gene tree with a single selected mutation:

(a) The selected site is segregating in the sample.

(b) The selected mutation has fixed in the population at some time in the past. Perlitz and Stephan (1997) have dealt with inference of this type based on the number of segregating sites in an infinitely-many-sites model with no recombination.

(c) The data have been collected in the subpopulation of sequences with a given mutation of known frequency, or the extension to case-control data. The mutation may be under selection. These data would be inappropriate for a coalescent analysis that assumes random sampling. The methods developed here can be used to correctly find the likelihood of a gene tree under this method of sampling. For example, consider the case where a sample is taken from a haplogroup defined by a particular neutral mutation found at a frequency of p in the population. To find the likelihood of the gene tree, say as a function of the population scaled mutation rate, trajectories of a neutral allele could be simulated back from the frequency p and then the importance sampling algorithm used to sample histories given these trajectories.

As in all population genetic models and inference the parameters of interest are scaled by the effective population size. For the models discussed here the main parameter of interest is the population scaled selection coefficient. If the effective population size is known then the unscaled selection coefficient can be calculated. Otherwise, the two cannot be determined separately. However the population scaled selection coefficient is itself of interest as it gives an indication of the relative effect of genetic drift compared to selection.

The method is finally applied to three published data sets in Chapter 6, where the authors have suggested that natural selection has acted. The first data set is from a paper by Verrelli et al. (2002) where a region in the G6PD gene was sequenced to examine evidence for a partial sweep of an allele known to be associated with resistance to malaria. The data set of Harris and Hey (2001) from the FIX gene is the second data set, where reduced levels of diversity led the authors to suggest that a full sweep had occurred, near the region, at some time in the recent past. The third data set is that of Rieder et al. (1999), where a region of the DCP1 gene is sequenced, and the authors postulated that balancing selection may have led to the patterns of diversity seen at the locus.

Chapter 2

Models Arising in Population Genetics

Genetics, Probability, and Statistics have long shared a fruitful relationship. Much of the early work in theoretical population genetics was devoted to showing the compatibility, and the resulting combination, of the ideas of Mendel and Darwin (Provine, 1971). In exploring this mathematical synthesis, probabilistic models to describe gene frequencies in a population were introduced (Fisher, 1930; Wright, 1968). In population genetics, as in all modelling applications, the models do not attempt to incorporate all of the complexity of real life but instead try to capture some of the salient features. In doing this a number of simplifying assumptions are often made, and these will now be discussed. The population is assumed to be randomly mating in that all individuals have an equal probability of mating with any other individual, i.e. there is no subdivision, isolation by distance or assortative mating. The second assumption is that a mutant allele, confers on the individuals it is present in, no beneficial or deleterious effect. This assumption, termed neutrality, arises from the neutral theory advocated by Motoo Kimura and others (see Ohta and Gillespie (1996) for a review) that states that the vast majority of polymorphisms present in a population have little or no effect on the fitness of organisms in that population. A third often made assumption is that the population is of a constant size through time. This set of assumptions, called the

standard neutral model, is not particularly valid for real populations and a number of modifications are possible to better reflect reality. However the standard neutral model is useful in that it presents a null model that can be tested and rejected if the data are better explained by a more realistic model.

In this chapter two discrete state space models, the Moran and the Wright-Fisher models, are discussed. Both are models of general population evolution and can incorporate complex mutation and selection schemes. For the purpose of this thesis we shall only discuss them as models of a biallelic locus. While the Moran and Wright-Fisher models are superficially quite different, in the limit of a large population size they behave identically. This is true of a large number of models of population evolution, and such models are said to lie in the same domain of attraction as the Wright-Fisher diffusion process. The Wright-Fisher diffusion process, a continuous state space and time model is described. Information about the whole population is rarely if ever known thus, particularly for statistical inference, properties of a sample from the population are of interest. In describing a sample of individuals it is helpful to consider the genealogical history relating the individuals in the sample. The Coalescent, the genealogical process embedded in the diffusion process, shall be discussed.

2.0.1 Models of Selection

Natural selection acts on heritable variation, that influences the number of offspring of an individual (fecundity selection) or the probability of those offspring surviving to have children (viability selection). In population genetics three types of selection are often studied, these are positive directional selection, diversity maintaining selection and negative selection. To illustrate the different types of selection consider a single

	A_1	A_2
A_1	1	w_{12}
A_2	w_{21}	w_{22}

Table 2.1: The relative fitnesses of the different genotypes at a biallelic locus. Note that $w_{12} = w_{21}$.

nucleotide where two types A_1 and A_2 are present. When quantifying selection, a model of fertility selection is focussed on. The genotypic reproductive fitness has the form $w_{..} = 1 + s_{..}$, see Table 2.1. Without loss of generality we can set $s_{11} = 0$, as the fitness of interest is the relative fitness (see for example Ewens (1979)). The birthrate can be either higher ($s_{..} > 0$) or lower than A_{11} ($s_{..} < 0$).

1. Positive Directional Selection

   Positive directional selection, sometimes called classical Darwinian selection, acts when the change to the phenotype caused by the mutant allele increases the carrier's relative contribution of offspring to the next generation. The mutant allele if fortunate will increase in number in the population until every member of the population carries it. The exact dynamics of the frequency of the allele depend on the population size, the size of the relative contribution (i.e. the relative fitness and frequency of the selected allele) and the relationship between genotype and phenotype. The relationship between genotype and phenotype in a simple Mendelian model is given by letting $w_{22} = 1 + s$ and $w_{12} = 1 + sh$, where h expresses the interaction of the A_1 and A_2 alleles. For the selection to be classified as positive directional selection $h \leq 1$ is necessary. If $h = 1$ then there is complete dominance and the heterozygote A_{12} has the same fitness as A_{11}. A recessive A_2 allele is described by $h = 0$ as only in the homozygote is any effect seen. If $h = \frac{1}{2}$ then each allele acts independently. This is sometimes called genic

or haploid selection as it acts in such a way as to ignore genotypic configuration, thus the allele acts as though it was in a haploid population. Directional selection is often described by a model of genic selection as from a theoretical view it is appealing as the independence from the genotype state simplifies calculations.

2. Diversity Maintaining Selection

   This is a wide class of models where selection acts to maintain, or balance, polymorphism within the population. This encompasses models where new alleles to the population are initially favoured (Nielsen, 1999), and where any allele of low frequency can be favoured, for example the MHC locus (Takahata et al., 1992). Also included are models where the effect of selection is spatial-temporal, i.e. different alleles have the advantage at different times or spacial locations, for example the fast/slow alleles at the Adh locus in Drosophila (Begun et al., 1999). Balancing selection at a biallelic site can also occur when the heterozygote for the allele has an advantage over both of the homozygotes, a classic example is the sickle cell allele in humans. A simple model of heterozygote advantage is given by $w_{11} = w_{22} = 1$ and $w_{12} = 1 + s_h$. This is equivalent to the directional selection model given above as $s \to 0$, $h \to \infty$ and $sh \to s_h$. Balancing selection can lead to an allele persisting for a long time in the population once it has become established. In some cases current day polymorphisms are believed to have survived since distant species splits (Takahata et al., 1992).

3. Negative and Background Selection

   A change in the code for a protein can lead to that protein being defective, negative selection will act on the allele that encodes for this defect. This defect might be so fundamental that it leads to the death of the offspring before it is born, this leads to a distortion in the alleles passed on from parent to offspring (see for example Zollner et al. (2004)). However many deleterious mutations might not

be so severe and will leave the offspring functional but reduce its fitness slightly. The constant influx of deleterious mutations can to first approximation be taken as a drop in effective population size from that expected for a neutral population (Charlesworth et al., 1995). The reduction in the level of diversity and other effects caused at neutral sites by the constant influx of deleterious mutations is called background selection.

2.0.2 Mutational Models

A wide variety of mutation models are regularly used in population genetics, and a selection of those useful to this thesis are briefly discussed here.

1. Finitely-many-alleles model

   The allelic type of the individual, E, is one of a finite number of alleles, k, at the locus, $E \in \{E_1, \cdots, E_k\}$. The mutation process between alleles is governed by a rate matrix μ, $\mu_{ij} \geq 0$. In particular, we shall be interested in the case of a locus with just two alleles.

2. Infinitely-many-alleles model

   The allelic type of the individual, E, is one of an infinite number, assigned to the interval $E \in [0, 1]$. This model can be considered as the limit of the finite allele model as $k \to \infty$.

3. Infinitely-many-sites model

   If the mutation rate at an individual site is low, effectively zero, then at most one mutation is expected at a site. The infinitely-many-sites model arises as the limit of a model of DNA sequences made up of a large, effectively infinite, number of biallelic loci with very low mutation rates such that the overall sequence mutation rate is constant. This can be thought of as a finer modelling of the

infinitely-many-alleles model where all the previous allelic states of a gene in the population are recorded. A perfect record of the mutation history is kept. The data from the infinitely-many-sites model offers the full amount of information for a given overall mutation rate as the relationship between alleles in the population, including the number of mutations separating alleles, is preserved.

2.1 Discrete State Space Models

Let $Z(t)$ be the number of copies of the selected allele A_2 in a population of size N at time t. The Moran and Wright Fisher models describe the evolution of the number of copies of the allele A_2 through time, denoted by $\{Z(t), t \geq 0\}$, termed the trajectory.

2.1.1 Moran Model

The Moran model, introduced by Moran (1958), is a continuous time birth and death model in which a fixed population size is maintained by a birth always being accompanied by a death. The initial state of the process is given by $Z(0) = 1$, representing the allele arising in one gene in the population of genes by mutation. An individual of type A_1, has a birth rate $\lambda_1 = N/2$. While an individual of type A_2 has birth rate $\lambda_2 = \lambda_1(1 + s_N(Z))$, where $s_N(Z)$ is the birth rate advantage of the selected allele when there are Z copies of the allele in the population (for a model of genic selection $s_N(Z) = s_N$ i.e. a constant). The neutral birth rate of $N/2$ is chosen to place the Moran model on the same time scale as the Wright Fisher model. We will assume for simplicity that there is no recurrent mutation. An exponential distribution with rate parameter λ will be denoted by $\exp(\lambda)$. The waiting time till the next birth when there are z of the selected type is

$$\sim \exp(\lambda_1(1 + s_N(z))z + \lambda_1(N - z)). \tag{2.1}$$

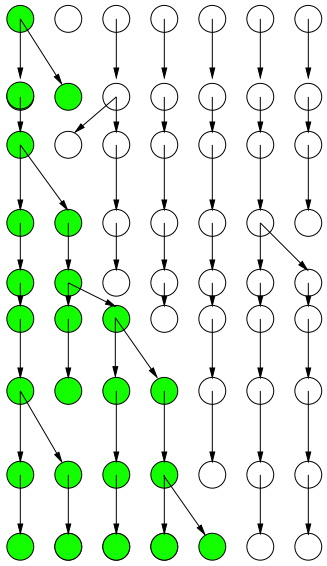

Figure 2.1: An illustration of the Moran model in a population of seven 7 individuals with time running down the vertical axis. The diagonal arrows show a birth event, vertical arrows a continuing individual. The selected allelic type is shown in green.

The allelic type that gives birth, increasing its number by one, is chosen with probability proportional to the population birth rates, and an individual is chosen to die uniformly. As there is no recurrent mutation the states 0 and N are absorbing. An example of this process is shown in Figure 2.1. Events which involve births and deaths within a type do not need to be considered, as we shall be only interested in the frequency of the allele. Times are simulated and events chosen conditional on a change in frequency. The birth and death rate (λ_z and μ_z) of allele A_2 when there are z in the population are now

$$\lambda_z = \lambda_1(1 + s_N(z))\frac{z(N-z)}{N}, \qquad \mu_z = \lambda_1\frac{z(N-z)}{N}, \tag{2.2}$$

and the time to the next event is

$$\sim \exp(\lambda_z + \mu_z). \tag{2.3}$$

A range of trajectories starting from a mutation arising in one individual for different selection values can be seen in Figure 2.2. As the strength of selection increases the trajectories become more deterministic, see Section 2.6.1. For a value of $s_N > 0.025$

a deterministic path through intermediate frequencies would be an appropriate approximation. However it is worth noting that even at the highest selection value, $s_N = 0.1$, only 4 of the 20 trajectories fix in the population, the other trajectories hit zero very quickly and thus are not visible.

2.1.2 Wright-Fisher Model

In the Wright-Fisher model generations are discrete, and every individual gives birth and is replaced every generation. The generation $t+1$ is constructed by sampling N individuals with replacement (i.e. binomially) from an infinite gamete pool constructed from the previous generation of $Z(t)$ A_2 genes. We assume a mutation rate u from type $A_1 \rightarrow A_2$ and v from type $A_2 \rightarrow A_1$. While later we will set $u = v = 0$, the general model is introduced here. The frequencies of alleles in the infinite gamete pool are constructed by allowing every individual to mate randomly. The gamete pool contribution to the A_2^{th} type by A_1 is denoted by $f_{A_1}(A_2, t)$. The relative frequency of A_2 in the generation t is $X_t = \frac{Z(t)}{N}$. The various contributions to the infinite gamete pool for the generation t are given below.

$$f_{A_2}(A_2,t) = \frac{1-v}{\overline{w}_t}\left(X_t^2 w_{22} + X_t(1-X_t)w_{12}\right) \tag{2.4}$$

$$f_{A_1}(A_2,t) = \frac{u}{\overline{w}_t}\left(X_t(1-X_t)w_{12} + (1-X_t)^2 w_{11}\right), \tag{2.5}$$

$$f(A_2,t) = f_{A_2}(A_2,t) + f_{A_1}(A_2,t), \tag{2.6}$$

where $\overline{w}_t$ is the population fitness given by

$$X_t^2 w_{11} + 2X_t(1-X_t)w_{12} + (1-X_t)^2 w_{22}. \tag{2.7}$$

There are similar equations for $f_{A_2}(A_1,t)$ and $f_{A_1}(A_1,t)$. The number of the allele A_2 in the next generation is sampled binomially from these frequencies.

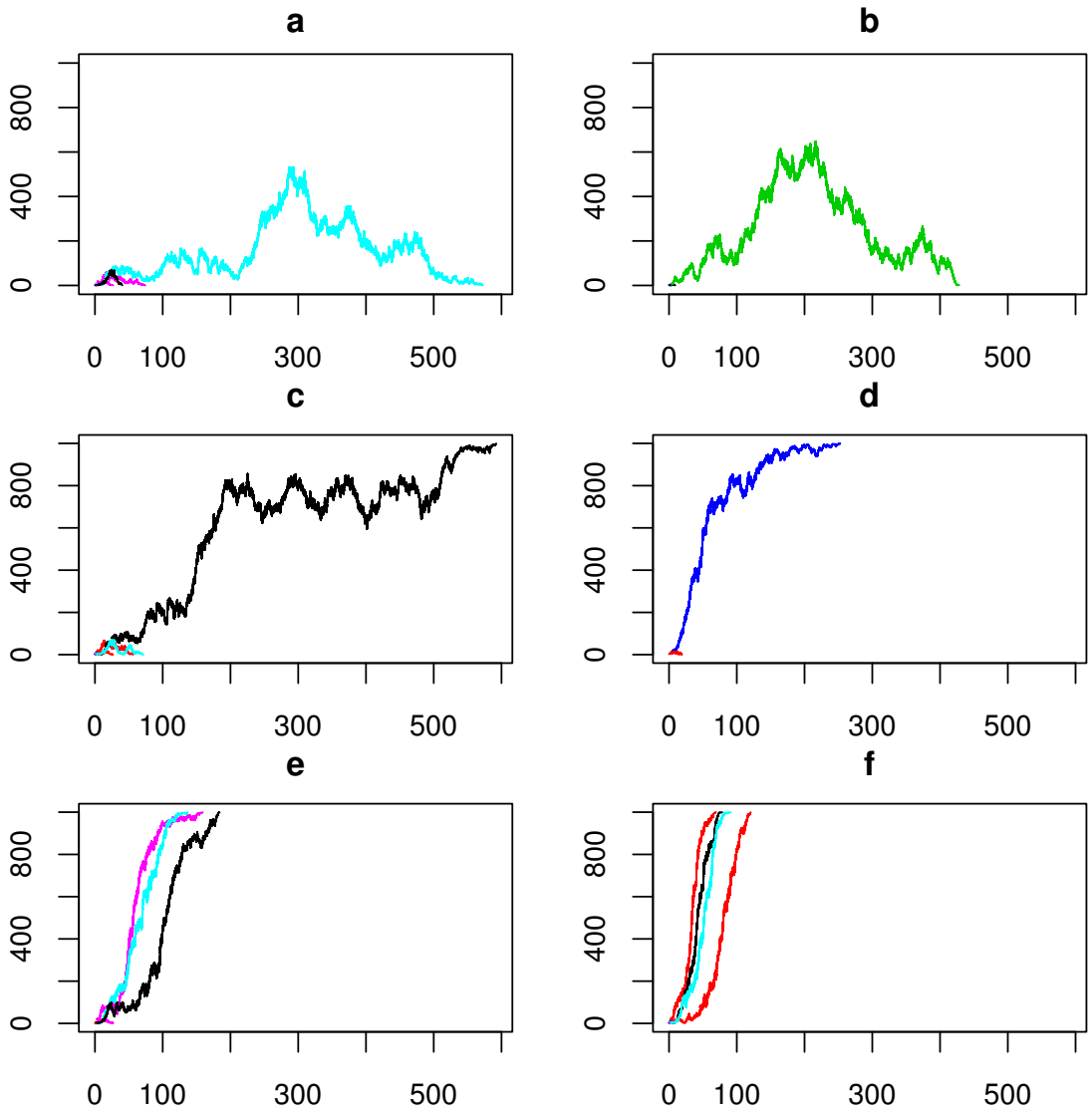

Figure 2.2: A sample of twenty trajectories generated by a forward in time Moran model ($N = 1000$) from an initial occurrence of the mutation in one individual for a variety of genic selection coefficients a) neutral $s_N = 0$, b) $s_N = 0.0001$, c) $s_N = 0.001$ d) $s_N = 0.025$ e) $s_N = 0.05$, f) $s_N = 0.1$. Time is given along the x-axis and number of the selected mutation in the population on the y axis

$$P(z(t+1) = Z \mid X_t) = \frac{N!}{Z!(N-Z)!} f(A_2,t)^Z (1 - f(A_2,t))^{N-Z} \qquad (2.8)$$

2.2 The Diffusion Approximation

Biologists and mathematicians, given the approximations made in constructing a model, are often not interested in the different detailed behaviour of discrete state space models and only care about behaviour that is general to many models. The shared behaviour is often observed when a large population size is considered. The large population limit of the Moran and Wright-Fisher models, with appropriate time scaling, is a diffusion process. The diffusion process is an infinite population limit and thus should only be applied to modelling large populations, however in even models of relatively small populations the diffusion process approximation can be useful. Many results are more tractable in the large population limit and so diffusion processes are often of great use, as long as their approximate nature is borne in mind. Diffusion processes were used extensively by Motoo Kimura (see Watterson (1996) for a review) and they underlie many results commonly used.

A continuous time Markov stochastic process whose sample paths are almost surely continuous functions of time, is called a diffusion process, see Karlin and Taylor (1981) for a general introduction to all topics relating to the diffusion processes discussed in this thesis. Let $\Delta_h X(t) = X(t+h) - X(t)$. As $h \to 0$ this is the infinitesimal increment. In practical applications diffusion processes are characterized by the following three features:

$$\lim_{h \to 0} \frac{1}{h} E[\Delta_h X(t) \mid X(t) = x] = \mu(x,t)dt, \qquad (2.9)$$

$$\lim_{h \to 0} \frac{1}{h} E[(\Delta_h X(t))^2 \mid X(t) = x] = \sigma^2(x,t)dt, \qquad (2.10)$$

$\mu(x,t)$ is the infinitesimal mean or drift parameter, and $\sigma^2(x,t)$ is the infinitesimal variance. All higher infinitesimal moments are zero

$$\lim_{h\to 0}\frac{1}{h}E[(\Delta_h X(t))^r \mid X(t)=x]=0, \qquad r=3,4\cdots \quad . \tag{2.11}$$

In general a Markov chain can be approximated by a diffusion process when (2.9), (2.10) and (2.11) hold. This will be informally shown for the Moran and Wright-Fisher models.

Moran Model to Diffusion Process

In the Moran model described above if at time t there are z genes of type A_2 with a selective advantage of $1+s_N(z)$ over A_1 then the expected change in frequency is

$$\begin{aligned}
\mathbb{E}[\Delta_h X(t) \mid X(t)=x] &= \frac{1}{N}P(Z(t+h)=z+1 \mid Z(t)=z) \\
&\quad -\frac{1}{N}P(Z(t+h)=z-1 \mid Z(t)=z) \\
&= \lambda_1 z\frac{N-z}{N^2}((1+s_N(z))-1)h+O(h^2) \\
&\to \beta x(1-x)h, \\
\text{as } N\to\infty, \qquad \frac{z}{N} &\to x \qquad \text{and } \frac{Ns_N(z)}{2}\to\beta(x).
\end{aligned}$$

A similar limit can be found for for the infinitesimal variance, giving $\sigma^2(x)=x(1-x)$, and all higher infinitesimal moments are zero.

Wright-Fisher Model to Diffusion Process

In the Wright-Fisher model in the current generation suppose there are i type A_2 genes in a total population of N genes. The transition probabilities, from a state i to j, are given by binomial sampling of the next generation from the infinite gamete pool con-

structed from the i type A_2 individuals as

$$P_{ij} = \binom{N}{j} f(A_2,t)^{\,j}(1-f(A_2,t))^{N-j}, \;\; 0 \le j \le N$$

where $f(A_2,t)$ is the frequency of type A_2 in the infinite gamete pool. Considering for simplicity the case of no recurrent mutation with genic selection acting on the allele A_2, then

$$f(A_2,t) = \frac{(1+s_N)i}{(1+s_N)i+(N-i)}.$$

The infinitesimal drift, letting $h = \frac{1}{N}$, is

$$\begin{aligned} \frac{1}{h}\mathbb{E}[\Delta_h X(t) \mid X(t)=x] &= N\left(\frac{(1+s_N)i}{((1+s_N)i+(N-i))}\right) - i \\ &\to \beta x(1-x), \text{ as } x = \frac{i}{N}, \;\; \beta = Ns_N \end{aligned} \tag{2.12}$$

By a similar argument

$$\sigma^2(x,t) = x(1-x), \tag{2.13}$$

and all higher infinitesimal moments are zero

2.2.1 The Scale Function and the Speed Measure

Results for the one dimensional diffusion process are often expressed in terms of the scale function and speed measure of the diffusion process. A diffusion is said to be on the natural scale if the probability of going to one boundary before the other is proportional to distance to the boundary. The scale function, $S(x)$, is the function that maps a diffusion process onto its natural scale. An explicit expression for $S(x)$ is

$$S(x) = \int^x s(y)dy \tag{2.14}$$

$$s(y) = \exp\left\{-\int^y [2\mu(\zeta)/\sigma^2(\zeta)]d\zeta\right\}, \qquad 0<y<1, \tag{2.15}$$

so the probability of going to loss before fixation from x is

$$u_0(x) = \frac{S(1) - S(x)}{S(1) - S(0)}, \tag{2.16}$$

$$= \frac{\int_x^1 s(y)dy}{\int_0^1 s(y)dy}. \tag{2.17}$$

The neutral model is on the natural scale as $u_0(x) = 1 - x$. For a model of genic selection

$$s(y) = e^{-\beta y}, \tag{2.18}$$

$$u_0(x) = \frac{e^{-\beta x} - e^{-\beta}}{1 - e^{-\beta}}. \tag{2.19}$$

The speed measure, $m(x)$, has the interpretation of being the speed that the time runs at locally when the diffusion process is on the natural scale

$$m(x) = [\sigma^2(x)s(x)]^{-1}. \tag{2.20}$$

$m(x)dx$ is the expected time to leave an infinitesimal interval centered on $[x - dx/2, x + dx/2]$. Both the Scale function and Speed measure are found by considering solutions to the Kolmogorov backward equation, which arises naturally as the solution to the expectation of certain quantities considered along the diffusion path.

2.2.2 The Generator of a Diffusion

The transition function $f(x, y, t)$ is the probability density function of $X(t)$ given $X(0) = x$. The Kolmogorov backwards equation describes how the transition function behaves as a function of its initial condition x. The generator $\mathcal{L}$ of a diffusion process specifies the generalised Kolmogorov backwards equation of the process. By considering the operator for the expectation $h(x) = \mathbb{E}_x\{g(X(t))\}$, and examining how the diffusion $g(x)$ changes in a small time interval dt, the generator

$$\frac{dh(x)}{dt} = \mathcal{L}h(x) = \mu(x)\frac{dh(x)}{dx} + \sigma^2(x)\frac{d^2h(x)}{dx^2} \tag{2.21}$$

is derived. The function $h(x)$ has to be twice differentiable with respect to x. If there is recurrent mutation between the two alleles in the model, see for example the Wright-Fisher model described above, then the infinitesimal mean, $\mu_m(x)$, is

$$\mu_m(x) = \mu(x) + \mu_{12}(1-x) - \mu_{21}x \tag{2.22}$$

where $2Nu \to \mu_{12}$ and $2Nv \to \mu_{21}$.

Path Integrals

The integral of a functional, $f(x)$, along the diffusion path $\{x(t), t > 0\}$ is of interest. The expectation of such an integral over all sample paths that the diffusion process takes is written as

$$\mathbb{E}\left(\int_0^T f(x(t))dt \mid x(0) = p \right) = w_p, \tag{2.23}$$

where T is the time of fixation or loss. The solution to (2.23) may be derived by considering a small time interval and the accumulation of the integral in this time. The following differential equation must be solved to find a solution to (2.23)

$$-f(x) = \left\{ \mu(x)\frac{d}{dx} + \sigma^2(x)\frac{d^2}{dx^2} \right\} w_x, \tag{2.24}$$

with boundary conditions $w_0 = 0, \;\; w_1 = 0$. The expected time to obtain a boundary from a frequency p, $\mathbb{E}_p(T)$, is the solution of (2.24) when $f(x) = 1$. The expected total time spent in the frequency interval $[y - dy/2, y + dy/2]$, for a process initially at p, before hitting either boundary is of interest. This can be found by considering the path integral

$$\mathbb{E}\left\{ \int_0^T \delta(y - x(t))dt \mid x(0) = p \right\}. \tag{2.25}$$

where $\delta(v)$ denotes the dirac delta ($\delta(x) = 1$ for $x = 0$ and $\delta(x) = 0$ otherwise). The solution of which is the Green function, $G(p, y)$,

$$G(p,y) = \begin{cases} 2u_0(p)m(y)\int_0^y s(v)dv & 0 \le y \le p \\ 2u_1(p)m(y)\int_y^1 s(v)dv & p < y \le 1. \end{cases}$$

The Green Function is a useful quantity as it presents a general solution for the expectation (2.23) as it is the expected total time spent in a small frequency interval $[y - dy/2, y + dy/2]$, for a process initially at p. This allows (2.23) to be expressed as

$$w_p = \int_0^1 G(p, y) f(y) dy. \tag{2.26}$$

Higher moments of path integrals, some functionals of path integrals and even their distributions can be found, see for example Maruyama (1977).

Path Integrals with Killing

If the diffusion process is modified to one where a sample path at time t in the frequency interval $[x, x + dx]$ is killed with rate $k(x, t)$, then the resulting process is a diffusion process with exponential killing. These processes provide a natural way to describe path integral diffusion problems where the stopping conditions are not merely the reaching of a boundary. The expectation of the integral of a function, $f(x)$, along sample paths in a process with killing can be written as

$$w_p = \mathbb{E}\left\{ \int_0^T f(x(t)) \exp\{-\int_0^t k(x(u), u) du\} dt \mid x(0) = p \right\}, \tag{2.27}$$

where T is the time at which the sample path hits either boundary. For example the expectation of the total time before a homozygote for a particular allele is present in the population can be found by considering the integral of $f(x) = 1$ along sample paths which are killed at rate $k(x, t) = x^2$, the probability of homozygote appearing. w_x, if $k(x, t)$ is independent of t, satisfies the Feynman-Kac equation

$$-f(x) = \left\{ \mu(x) \frac{d}{dx} + \sigma^2(x) \frac{d^2}{dx^2} - k(x) \right\} u_x, \tag{2.28}$$

with boundary conditions $w_0 = 0, w_1 = 0$.

2.3 The Coalescent

When studying the properties of a sample from a neutral population it is sufficient to consider the genealogical history of the sample rather than that of the whole population. The Coalescent, derived by Kingman (1982) (see Nordborg (2001) for a review), is a genealogical model for the history of a sample and is the genealogical process embedded in the neutral diffusion process. In a neutral population the genealogy relating individuals is independent of the mutation process. Thus the genealogy of the sample may be considered and then mutations placed on the genealogy. Mutations are placed onto each edge of the genealogy as a Poisson process with parameter $\frac{\theta}{2}T$ where θ is the population mutation rate, $\theta = 2NL\mu$, where L is the sequence length and μ the per site mutation rate, and T is the time along an edge. Due to the sufficiency of the genealogy of the sample and the independence of the genealogical process from the mutation process many results are easily derived and simulation is computationally efficient in a coalescent setting.

The Coalescent may be derived by finding the probability that a pair of lineages in the sample of n find a common ancestor in the previous generation in a Wright Fisher population of size N. A particular pair find a common ancestor one generation back with probability $1/N$, thus the probability that any pair finds a common ancestor is $\binom{n}{2}\frac{1}{N}$. The probability that more than 2 lineages find a common ancestor in the same generation is $O(\frac{1}{N^2})$, thus in the large population limit a coalescent genealogy is bifurcating. The probability of no coalescences for i generations into the past is approximately $(1 - \binom{n}{2}/N)^i$. The process is placed on the coalescent time scale by letting 1 scaled unit of time be N generations. The probability of no coalescences in τ scaled time units in the limit is

$$\lim_{N\to\infty}\left(1 - \binom{n}{2}/N\right)^{[N\tau]} = e^{-\binom{n}{2}\tau}.$$

Thus the coalescence time, T_n the time while there are n lineages can be seen to be $\sim \exp(\binom{n}{2})$. Pairs of lineages are exchangeable so when a coalescence occurs the choice of pair to coalesce is uniform. The genealogy is completed when the last two lineages coalesce to give the most recent common ancestor of the sample. As the coalescent rate is quadratic with the number of lineages the sample, even as $n \to \infty$, is guaranteed to find a common ancestor in finite time.

A coalescent genealogy may also be constructed by simultaneously generating the genealogy and the mutations on the genealogy. It is helpful to consider this approach and the urn model discussed shortly as it allows a better understanding of the tree recursions we will encounter later. From the Poisson nature of the number of mutations on an edge it follows that the waiting time between mutations on a branch is $\sim \exp\left(\frac{\theta}{2}\right)$. Instead of just considering the time to the next coalescence event, the time to either the next mutation on one of the n edges or a coalescence event can be considered, this time is

$$\sim \exp\left(\binom{n}{2} + \frac{n\theta}{2}\right). \tag{2.29}$$

Given that an event has occurred, by a relative rate argument, the probability that it is a mutation or a coalescence are respectively

$$\frac{\frac{n\theta}{2}}{\binom{n}{2} + \frac{n\theta}{2}} = \frac{\theta}{n+\theta-1}, \quad \frac{\binom{n}{2}}{\binom{n}{2} + \frac{n\theta}{2}} = \frac{n-1}{n+\theta-1}. \tag{2.30}$$

As before, if a coalescence has taken place a pair of lineages to coalesce is chosen at random. If a mutation is chosen to happen, a mutation is placed on a randomly chosen lineage. For questions regarding the mutation pattern in the sample it can be seen that the times are unnecessary and that the order of events is sufficient.

Forwards in time this approach can be viewed as an urn model (Hoppe, 1984). Two balls of weight θ and one are initially placed in an urn. A ball is chosen at random with

probability proportional to its weight, and then replaced into the urn. If the θ ball is chosen, then, for the infinite alleles model, a weight one ball of a colour not present in the urn is also placed into the urn. If a weight one ball is chosen then another weight one ball of the same colour as the one picked is placed in the urn as well. We proceed until n balls of weight one are present in the urn. This approach gives the same distribution of alleles as the Coalescent and is also helpful in understanding the recursions to be discussed later.

2.3.1 The Coalescent and Variable Population Size

Incorporating variation in the size of the population is an important addition to the coalescent (Slatkin and Hudson, 1991; Griffiths and Tavaré, 1994b). It also, as we will see later, presents an exact analogy to the coalescent with natural selection when the population frequency of the selected type is known throughout time. Instead of the fixed population size N that has been considered to date, we now allow the population to vary so that its size at time t is $N(t)$. Intuitively, when the population size is small a pair of lineages are more likely to find their common ancestor, as there is a smaller choice of parents in the generation before. This change in the rate of coalescence can be incorporated via a change in the time scale of the coalescent. For a discrete generation model (e.g. the Wright Fisher model) t generations can be rescaled to

$$g(t) = \sum_{i=1}^{t} \frac{1}{N(i)}, \tag{2.31}$$

while in a continuous setting

$$g(t) = \int_0^t \frac{1}{N(s)} ds. \tag{2.32}$$

Scaling the population size by the population size at $t = 0$, the rate of coalescence when there are k lineages at a time t is

$$\binom{k}{2}\lambda(t), \tag{2.33}$$

where $\lambda(t) = N(0)/N(t)$. The distribution of the time of a coalescence in k lineages, for a sample of n, T_k given that the last coalescence occurred at time S_{k+1} ($S_{k+1} = T_n + \cdots + T_{k+1}$) has an inhomogeneous exponential distribution

$$\binom{k}{2}\lambda(S_{k+1} + T_k)\exp\left\{-\binom{k}{2}\int_{S_{k+1}}^{T_k}\lambda(u)du\right\} \tag{2.34}$$

The choice of pair of lineages to coalesce is left unchanged as the pairs are still exchangeable and the mutation process given edge lengths is the same as before. Thus the Coalescent with a varying population size is exactly the same as a neutral process but with scaled time.

Once again the genealogy and mutation process may be considered simultaneously. The time, t, to the next mutation or coalescence event back in time given that the last event occurred at s and there are currently k lineages is distributed as

$$\left(\binom{k}{2}\lambda(t) + \frac{k\theta}{2}\right)\exp\left\{-\int_s^t \binom{k}{2}\lambda(u) + \frac{k\theta}{2}du\right\} \tag{2.35}$$

Given that an event has occurred at t the probability that it is a mutation or coalescence are respectively

$$\frac{\frac{k\theta}{2}}{\binom{k}{2}\lambda(t) + \frac{k\theta}{2}}, \quad \frac{\binom{k}{2}\lambda(t)}{\binom{k}{2}\lambda(t) + \frac{k\theta}{2}}. \tag{2.36}$$

2.4 Gene Trees

A number of different types of genetic markers are used to identify and study regions under selection including microsatellites, see for example Payseur et al. (2002), and

single nucleotide polymorphisms (SNPs) utilised by Akey et al. (2002) and Sabeti et al. (2002) amongst others. In this thesis we focus on data generated from resequencing a small region. We assume that this region is fully linked and that the model of mutation is the infinitely-many-sites model.

Based on the assumption of infinitely-many-sites other authors have derived various properties of the mutation number and pattern. The segregating sites are the only informative sites about the genealogy and thus it is enough to consider only them. The expected number of segregating sites ($\mathbb{E}(S)$) in n sequences was derived by Watterson (1975) and is

$$\mathbb{E}(S) = \theta \sum_{i=1}^{n} \frac{1}{i}. \tag{2.37}$$

This was derived before the introduction of the Coalescent process but it is most easily derived in a Coalescent framework. The probability distribution of the number of segregating sites was also derived by Watterson (1975). The full distribution is the convolution of geometric distributions with p.g.f.

$$G_S(s) = \prod_{i=1}^{n-1} \frac{1}{(1 + \theta(1-s)/j)}. \tag{2.38}$$

The number of haplotypes K, i.e. the total number of distinct sequences in the sample of n was studied by Ewens (1972). For example the expected number of haplotypes is

$$\mathbb{E}(K) = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1}. \tag{2.39}$$

An urn model representation of the infinitely-many-sites model is described by Griffiths (1989). The pattern of mutation in a sample contains information about the genealogical history of that sample. A unique gene tree may be constructed for the data under the assumption of no recombination or recurrent mutation. The gene tree is the perfect phylogeny of the data, a unique representation of the full data up to the ordering of mutations on edges. The gene tree is partially time ordered in the sense that certain mutations have to occur in time before others. This tree is often a helpful way to visualise

the mutational pattern. The gene tree vastly reduces the coalescent tree space that can explain the data and this greatly aids inference. A representation of sequences in an example data set is shown in Figure 2.3. Sites on a haplotype that are of a nonancestral type, i.e. mutant, are marked with a cross. For example, at the vertical position of the first cross the first haplotype has the ancestral allele while haplotypes $2-7$ have the mutant allele at this site. The data may be represented as a $p \times q$ binary incidence matrix, showing which of the p haplotypes have the q observed segregating sites. The matrix has elements $S_{ij} = 1$ if the j^{th} segregating site is of the mutant type in sequence i, and 0 otherwise. The binary incidence matrix for the example data set is shown in Table 2.2.

While a tree can be deduced from first principles and drawn by hand from the binary incidence matrix it is a laborious and error prone process. Gusfield (1991) details an efficient algorithm to produce the perfect phylogeny, the gene tree, from the data. This is briefly detailed here.

1. Represent identical duplicate columns in the binary data matrix as a single column with a label corresponding to the identical columns. This collapses identical mutations, for example columns 10 and 12 are assigned to a single column labelled $(10, 12)$.

2. Considering each column as a binary number, sort the numbers into increasing order, with the smallest number in column one (note this is very slightly different from the original Gusfield algorithm. The numbers were originally sorted in decreasing order to be read from right to left).

3. Construct paths from the leaves of the tree to the root in the gene tree by labelling

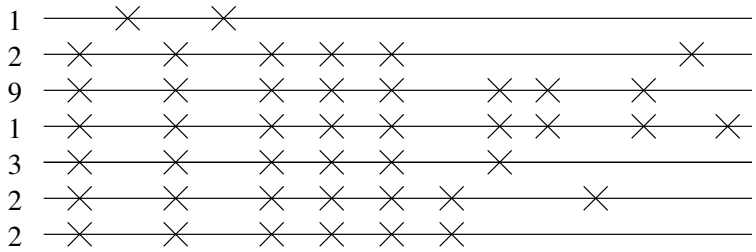

Figure 2.3: An example data set. Each line represents a haplotype. The number to the left of the haplotype is the multiplicity of that haplotype in the sample. Sites on a haplotype which are nonancestral are shown as crosses.

multiplicities	Segregating sites													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
2	1	0	1	0	1	1	1	0	0	0	0	0	1	0
9	1	0	1	0	1	1	1	0	1	1	0	1	0	0
1	1	0	1	0	1	1	1	0	1	1	0	1	0	1
3	1	0	1	0	1	1	1	0	1	0	0	0	0	0
2	1	0	1	0	1	1	1	1	0	0	1	0	0	0
2	1	0	1	0	1	1	1	1	0	0	0	0	0	0

Table 2.2: The binary incidence matrix for the example data set. 0 denotes the ancestral type and 1 the mutant type

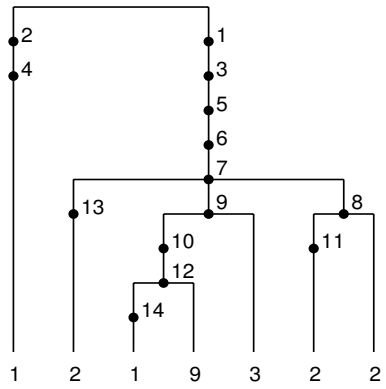

Figure 2.4: The gene tree for the example data set. The black dots are the mutations with their reference number written beside them, and the multiplicities of the various types are given at the bottom.

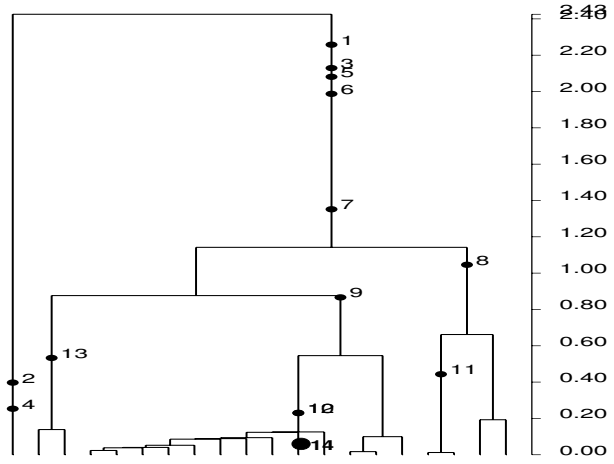

Figure 2.5: A possible coalescent tree for the example data. The horizontal lines are where coalescent events take place, joining the two lineages that find their common ancestor at this time. The axis on the righthand side gives the time in coalescent units.

nodes by mutation column labels, and reading left to right where 1's occur in rows.

Figure 2.4 shows the gene tree for the example data set. It is unique up to the ordering of the mutations $(1, 3, 5, 6, 7)$, $(2, 4)$, and $(10, 12)$. A possible coalescent history that could underlie the gene tree is shown in Figure 2.5. In a gene tree, mutations are vertices and the coalescences are not observed. In a coalescent tree, coalescences are the vertices and the mutations are on the edges. The infinitely-many-sites assumption is an elegant approximation that is often made. In species such as humans the mutation rate in many areas of the genome is suitably low that the assumption is not inappropriate.

The data might feature a small number of sites or sequences that violate the infinitely-many-sites assumption (i.e. the data set has a small of sites that fail the four gamete test (Hudson and Kaplan, 1985)). If this is the case, it is possible to prune the data set. Sites may be removed, or an extra column added in the binary mutation matrix (if a repeat mutation can be resolved in this way), or whole sequences removed. This would make the data compatible with the infinitely many sites assumption. However issues are raised by this treatment, firstly the results are questionable as important data might have been excluded. In general, an incompatibility might suggest that the 'tree' is extended back deeper in time than the subsequent analysis will show. The incompatibility after all represents another event on the tree and would deepen the tree if included. In general the application of any method outside of its assumptions should be treated carefully. As long as caution in interpreting the results is taken however, gene tree style methods can still be useful.

2.5 Previous Approaches to Modelling the Coalescent with Selection

Selection is difficult to incorporate into the Coalescent framework as the mutation and genealogical process are no longer independent, The relative birth rate of a parent is now dependent on its type and that of the rest of the population, thus the coalescent rate will depend on the allelic frequencies through time. This breaks down the elegance of the Coalescent making analytical results much harder to obtain. A brief overview of approaches to selection and the coalescent is given here.

The problem of incorporating non-deterministic selection into the Coalescent has been tackled in a number of different ways. In this section we will briefly describe how this has been done, and explain some of the advantages and disadvantages of the various methods. A number of other approaches to selection and linked variation have been developed and a few are mentioned here. Slatkin (2001) gives a method to generate stochastic trajectories backwards in time in an expanding population. A coalescent genealogy can be generated using this population trajectory, and the likelihood of the decay of linkage disequilibrium from the selected locus evaluated given this genealogy. Wiuf (2001) derived approximations to the coalescent for low frequency variants and examined the effects of population expansion and selection. In Donnelly et al. (2001) the observation that the stationary distribution of a nonneutral model with parent independent mutation is similar to that of the neutral distribution has been exploited to develop simulation and inference methods.

2.5.1 Ancestral Selection Graph (ASG)

The ASG, developed by Krone and Neuhauser (1997) and Neuhauser and Krone (1997), is the genealogical process of the diffusion with selection. This is derived by considering first a graphical representation of the Moran model with selection and mutation. In this representation a set of N lines represent the population of size N through time. Forward in time two types of event may be placed on a line: mutations and arrows connecting the line to another randomly chosen line. An arrow pointing to a line represents this line being replaced by an offspring from the line from which it points. In terms of ancestry an arrow signifies the replacement of that individual's ancestry with the ancestral information of the line it points from. The rate that arrows are placed from a line is determined by the allelic type of the line, a line of the selected type gives birth at a higher rate than a neutral line.

Forward in time this process is simple. A type is drawn for an individual from the stationary distribution and this line proceeds forward mutating to other types being killed and thus inheriting the type of the incoming arrow, and replacing other's ancestry by its own, by giving birth onto another line through an arrow event. A sample of individuals may be examined, their ancestry followed back up the graph, crossing any arrows that come into their line, and then following this ancestral line upward.

The difficulty arises when trying to construct the ancestry for a sample backward in time, as the rate a line should give birth to arrows is not known as its allelic type is unknown. Krone and Neuhauser (1997) over came this by letting all lines give birth to arrows at the highest rate. These arrows are then probabilistically labelled, a form of thinning, to determine whether an allelic type's ancestral information may cross across the arrow. When the ultimate ancestor of the sample (the topmost node in the graph,

the first node where the number of lineages is one) is found backward in time its type may be drawn from the stationary distribution of the allele frequency and this ancestral information flows forward in time down the graph crossing arrows where the labels permit.

The large population limit of this process is one where the usual coalescent process is added to by branching events representing the labelled arrow events (or unresolved birth events) thus a graph is created. When the ultimate ancestor of the sample is reached a type is drawn from the stationary distribution and the graph is collapsed forward in time, leaving the true ancestry. An example ASG is shown in Figure 2.6. The ancestry of a sample of 3 has been followed back and the ancestral type 1 (the unfit type) of the ultimate ancestor drawn. Following down the graph we do not cross the phantom branch as the line is of the unfit type. To find the probability of a sample's mutation configuration a recursion, related to those of Griffiths and Tavaré, is constructed by considering the generator (similar to that described in 2.2.2) of the diffusion process acting on a multinomial sample.

The ASG effectively integrates out the population trajectory of the selected allele by the introduction of the branching events. It has been adapted to frequency dependent selection (Neuhauser, 1999) and the graph process extended to incorporate recombination and general selection (Donnelly and Kurtz, 1999). A computational method to obtain the stationary distribution of complex mutation and selection models is developed for a single locus by Fearnhead (2003) and for unlinked sites by Fearnhead (2001). However the branching is extreme in the strong selection case and makes simulation difficult. This has been partly overcome by ignoring certain branching events that have no effect on the ancestry (Slade, 2000b). Interesting results for the ASG have been derived (Fearnhead, 2002) but analytical results are difficult to achieve and can lack an intuitive

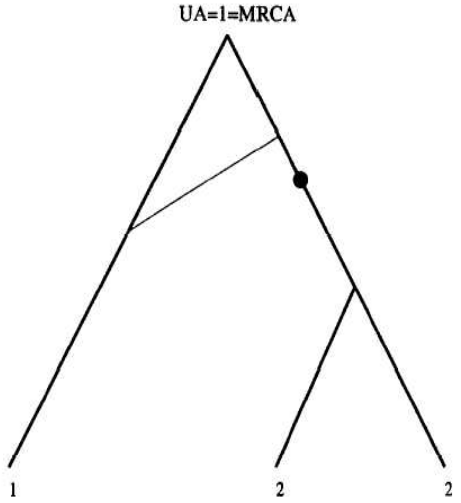

Figure 2.6: An example ASG taken from Neuhauser and Krone (1997). By the drawing of the type for the Ultimate ancestor (UA) the true genealogy can be found. The true genealogy is shown in bold having drawn an unfit type for the UA. The numbers at the tips give the types.

element that coalescent results brought to population genetics. While good progress has been made on inference (Slade, 2000a; Stephens and Donnelly, 2003) full likelihood of linked data in this setting is still not possible.

2.5.2 Expectations of the Coalescent and Path Integrals

Kaplan et al. (1988) develop an approach to find expectations of various statistics of the genealogy when selection acts at a single bi-allelic locus. The method has been taken forward, the diffusion process rigorously proved and boundary conditions given by Barton et al. (2004) and Barton and Etheridge (2004). Kaplan et al.'s 1988 method is described in part here. Let $Q(t) = (i, j)$ be the selected allelic configuration of the lineages ancestral to the sample in the t^{th} generation, when there are i of A_1, j of A_2. The

notation $|Q(t)|$ denotes the total number of lineages at t, i.e. $|Q(t)| = i + j$. The joint evolution of the frequency of the selected allele and the ancestral lineages, $Q(t) = (i, j)$ is studied. A Wright-Fisher model with recurrent mutation (see Section 2.1.2) is used to model the population through time. There are two possible ways that the sample allelic configuration, $Q(t)$, can change in the previous generation and these are now examined.

Case 1 $|Q(t-1)| \neq |Q(t)|$

In this case a coalescence has occurred between two of the ancestral lineages, two of the sampled genes have found a ancestral parent gene in the generation t. For example $Q(t-1) = (i, j-1)$, here two A_2 genes have found a parent. The fraction of genes contributed to the generation t by a parental gene of type A_2 in t is

$$\frac{X_t w_{22} + (1 - X_t) w_{12}}{N \overline{w}_t} + O\left(\frac{1}{N^2}\right) = \frac{f_{A_2}(A_2, t)}{N X_t} + O\left(\frac{1}{N^2}\right), \tag{2.40}$$

and probability of two sampled genes in $t-1$ having the same parent A_1 in t is

$$N X_t \frac{f_{A_2}(A_2, t)}{N X_t f(A_2, t)} + O\left(\frac{1}{N^2}\right) = \frac{1}{N X_{t-1}} + O\left(\frac{1}{N^2}\right). \tag{2.41}$$

In the derivation of these quantities recall from Section 2.1.2 that

$$f_{A_2}(A_2, t) = \frac{1 - v}{\overline{w}_t} \left(X_t^2 w_{22} + X_t (1 - X_t) w_{12} \right), \tag{2.42}$$

$$f_{A_1}(A_2, t) = \frac{u}{\overline{w}_t} \left(X_t (1 - X_t) w_{12} + (1 - X_t)^2 w_{22} \right). \tag{2.43}$$

As there are $\binom{j}{2}$ pairs of ancestral lineages the probability of a coalescence in the j A_2 lineages in generation t is approximately

$$\binom{j}{2} \frac{1}{N X(t)}, \tag{2.44}$$

similarly the probability of a coalescence in the i A_1 lineages in generation t is approximately

$$\binom{i}{2} \frac{1}{N(1 - X(t))}, \tag{2.45}$$

Case 2 $|Q(t-1)| = |Q(t)|$ but $Q(t-1) \neq Q(t)$

In this case an ancestral lineage has moved between backgrounds. For example $Q(t) = (i+1, j-1)$, a lineage has moved from the A_2 background to the A_1 background. The probability of one of the j ancestral A_2 lineages in generation $t-1$ having a parent A_1 in the generation before is

$$\begin{aligned} P(Q(t) = (i+1, j-1) \mid Q(t) = (i,j), X_t) &= j\frac{f_{A_2}(A_1, t)}{f(A_2, t)} + O\left(\frac{1}{N^2}\right), \\ &\approx j\frac{1 - X_{t-1}}{X_{t-1}}\frac{\mu_{12}}{N}, \end{aligned} \tag{2.46}$$

where μ_{12} is the population scaled mutation rate from A_1 to A_2, see Section 2.2.2 for a definition.

Using these cases, and placing the process on a diffusion time scale a set of linked differential equations can be constructed for the expected times in the various parts of the tree. These equations are related to the diffusion processes with killing discussed in Section 2.2.2. In that lineages are killed by coalescence and lineages in an particular type are killed by mutation and born into another type. The genealogy of a selectively neutral locus partially linked to the selective locus was studied by Hudson and Kaplan (1988), they find that the neutral locus recombines from the A_2 to A_1 background in the t^{th} generation with probability

$$(1 - X_t)\rho/N, \tag{2.47}$$

and with probability

$$X_t\rho/N \tag{2.48}$$

lineages recombine from the A_1 to the A_2 background, where ρ is the population scaled recombination rate between the selected and neutral locus.

2.6 Explicit Trajectory Models

While the ASG and the method of Barton and Etheridge (2004) integrate out the trajectory of the selected allele, a number of approaches chose to keep the trajectory explicitly in the model. This approach is particularly favoured for strong selection as the trajectory of the allele becomes increasing deterministic and thus better known. The trajectory of the selected allele can be simulated or approximated by a deterministic model. If the trajectory of the selected allele is known through time then the coalescent can be simulated backwards in time using the frequency of the selected allele and the ancestral allele as relative population sizes in a two deme subdivided population model (Kaplan et al., 1988).

2.6.1 Deterministic Selection

Many approaches (see for example Przeworski (2003)) choose to approximate the frequency of the selected allele through time with a deterministic trajectory. The deterministic equation is found by treating the infinitesimal mean, $\mu(x)$, as the rate of change in a differential equation, for example in a model of genic selection

$$\frac{dx}{dt} = \mu(x) = \beta/2x(1-x)$$

$$x(t) = \frac{1}{1 + \frac{1-p}{p}e^{\beta/2t}}. \tag{2.49}$$

where p is the initial frequency of the selected allele. Deterministic approximations are only valid when the infinitesimal mean is much greater than the infinitesimal variance, i.e. $\beta >> 1$. Figure 2.7 shows a comparison of a deterministic trajectory to stochastic trajectories for $\beta = 10$, while the stochastic trajectories are close to the deterministic approximation there is still variance in the paths.

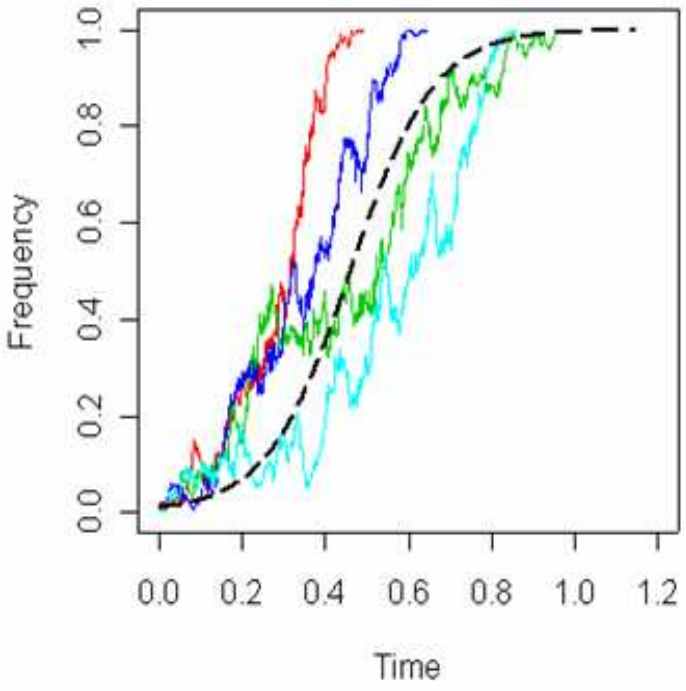

Figure 2.7: A comparison of a deterministic trajectory (shown in the dashed line) with 4 stochastic trajectories for $\beta = 10$

Chapter 3

The Wright-Fisher Diffusion Process and Reversibility

3.1 Properties of an Allele Segregating in the Population

The retrospective properties of the frequency trajectory of an allele polymorphic in the population have long been of interest in population genetics. These properties are still very much an area of study today as questions in population genetic inference are naturally concerned with the past history of alleles (Slatkin, 2002). Let $\{X(t),\ t \geq 0\}$ be the trajectory of a diffusion process modelling the frequency of an allele at a biallelic locus, where no recurrent mutation occurs, from when it arose at $t = 0$ by a mutation. This Wright-Fisher diffusion process model for the frequency has a generator

$$\mathcal{L} = \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2} + \mu(x)\frac{\partial}{\partial x}, \tag{3.1}$$

where $\sigma^2(x) = x(1-x)$. In general frequency dependent selection model $\mu(x) = \frac{1}{2}\beta(x)x(1-x)$, where the selection rate of the mutant allele is $\beta(x)$, for a model of

genic selection $\beta(x) = \beta$. The initial state of the process is $X(0) = p$, where p will be an arbitrarily small frequency, representing the mutation arising in one gene in a large population of genes, i.e. $p = 1/N$ as $N \to \infty$. Mutation rates in many organisms are known to be low and thus a model of genetic drift and selection alone, as in this case, is often thought to be appropriate. Importantly this model is the marginal diffusion process model of the frequency of an allele at one site in the infinitely-many-sites model. The age of an allele found at a frequency x in the population is well defined in this model as the time since the allele was at a frequency p, as $p \to 0$. However this process can be conceptually difficult to understand as the introduction of the allele by mutation is a unique event, rather than recurrent, and thus average properties of $\{X(t),\, t \geq 0\}$ might be difficult to discuss. The boundaries 0 and 1 are the (only) absorbing states, since there is no recurrent mutation in the model, and so the process has no stationary distribution. A number of different approaches have been used to study the retrospective properties of $\{X(t),\, t \geq 0\}$ and a selection of these will be reviewed. The retrospective properties of the infinitely-many alleles model have also been an area of thorough investigation (Watterson, 1976; Watterson and Guess, 1977; Maruyama and Fuerst, 1983). While this review restricts itself to discussion of retrospective properties of an allele in a diffusion process setting the properties have also been studied using discrete state space models (Kelly, 1977; Pakes, 1979; Tavaré, 1980; Pakes and Tavaré, 1981; Donnelly, 1984) and coalescent approaches (Griffiths and Tavaré, 1998; Wiuf and Donnelly, 1999; Stephens, 2000; Griffiths and Tavaré, 2003).

The transient pseudo-distribution, $f(y)$, sometimes termed the distribution under irreversible mutation or the population frequency spectrum, of the frequency of the allele was derived by Wright (1938), and subsequently by Kimura (1964). The transient pseudo-distribution is an improper distribution, as the normalisation constant is infinity. $f(x)dx$ is proportional to the average amount of time spent in a frequency interval

$[x, x + dx]$ before absorption at 0 or 1 for process with generator (3.1) as $p \to 0$ and an explicit expression is

$$f(y) \propto m(y)u_0(y), \;\; 0 < y < 1, \tag{3.2}$$

where $m(x)$ is the speed measure of the diffusion process and $u_0(x)$ the probability of hitting 0 before 1 from a frequency x (see Section 2.2.1 for explicit expressions). (3.2) can be derived from as limit of the Green function, $G(p, y)$, of the process

$$\lim_{p \to 0} G(p, y) = m(y)u_0(y). \tag{3.3}$$

(3.2) and the Green function are also interpretable as the stationary distribution of the return process analogue of $\{X(t) \mid t \geq 0\}$. This result was shown by Ewens (1963), using a modified forward equation, and used by Ewens (1964) in various applications. Return processes are discussed later in this Section. Figure 3.1 shows the population frequency pseudo-distribution for a number of different selection values of a derived allele for a genic model of selection. For all β the pseudo-distribution $f(x)$ concentrates on the tail-ends of of the interval [0,1]. For neutrality (i.e. a selection coefficient $\beta = 0$) $f(x) = \frac{1}{x}$, nearly all neutral alleles are lost almost immediately and only rarely will an allele reach higher frequencies by drift alone. For higher strengths of selection the curve becomes increasing 'U' like, as $\lim_{\beta \to \infty} f(x) = [x(1 - x)]^{-1}$ (see Section 5.3.3 for a derivation of this). An allele conferring a large selective advantage spends the vast majority of time at low or high frequencies, even for large β the allele is often lost quickly from the population, while at high frequencies the derived allele can take a long time to replace the few remaining ancestral alleles.

The time before the eventual loss or fixation of an allele has been studied extensively by a number of authors (see Crow and Kimura (1970) and references therein). Another area of study concerned the time before a particular boundary is reached, i.e. the time conditional on the point of absorption, in effect paths destined for the other

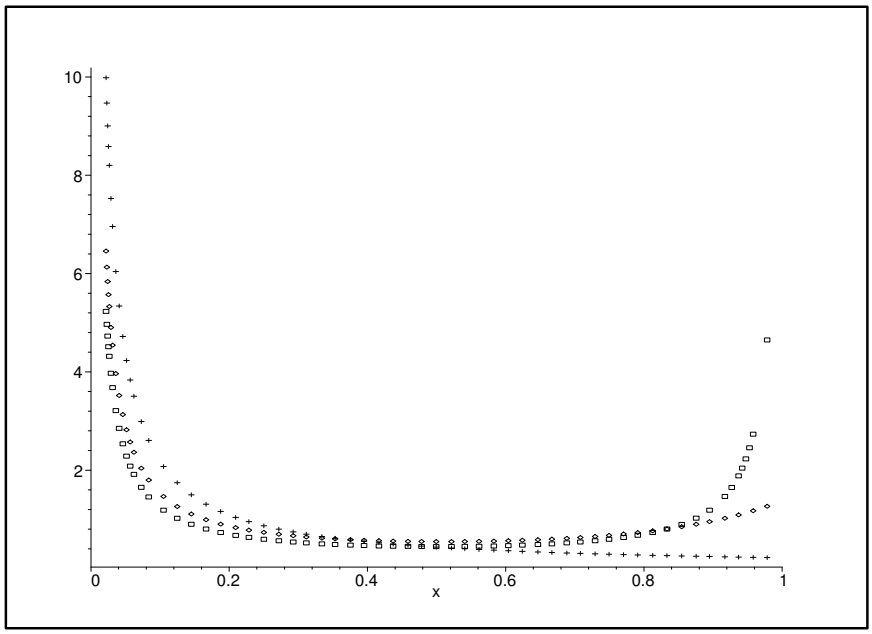

Figure 3.1: The population frequency spectrum of a derived mutation under a variety of genic selection coefficients $\beta = 0$ (+),$\beta = 10$ $\diamond$,$\beta = 100$ $\square$

boundary are ignored (Kimura and Ohta, 1969). This question is of interest as the time for an allele to fix in a population has a direct bearing on the rate of divergence between recently split species. The forward and backward equations conditional on hitting the boundary 1 before 0 were derived in a population genetics setting by Ewens (1973). Maruyama (1972) using the backward equation and techniques developed in Maruyama and Kimura (1971), derived the expectation and variance of the amount of time spent by paths in a frequency interval $[y, y + dy]$ given $x(0) = p$ and eventual fixation. Maruyama terms the expected time in a frequency interval the conditional sojourn time, but in a more general setting it would be called the Green function of the conditional process described by Ewens (1973). The sojourn times of the conditional process for a model of genic selection depend on $|\beta|$ only, thus, although the probability of a negatively selected allele reaching fixation is much smaller than that of a positively selected allele, conditional on reaching fixation the amount of time spent at intervening frequencies is the same.

Kimura and Ohta (1973) studied the expected age of a neutral allele found at frequency x, $0 < x < 1$, in the population. Let $f(p, x, t)$ be the transition density of the diffusion process (defined in Section 2.2.2) and T_p denote the time since the frequency of the allele was p (note that this does not have to be the last time the frequency was p). Then, for an allele whose frequency is currently x

$$\begin{aligned} \mathbb{E}_x(T_p) &= \mathbb{E}\left(T \mid x(0) = p, x(T) = x\right) && (3.4) \\ &= \frac{T_x^{(1)}}{T_x^{(0)}}, && (3.5) \end{aligned}$$

where

$$T_x^{(i)} = \int_0^\infty t^i f(p, x, t) dt, \quad i = 0, 1, \cdots \tag{3.6}$$

Kimura and Ohta (1973) use a modified forward equation to obtain expressions for T_1 and T_0. They find the expected T of a neutral allele to be

$$-2\big((1-p)p^{-1}\log(1-p) + x(1-x)^{-1}\log x + 1\big). \tag{3.7}$$

$\mathbb{E}_x(\text{Age})$, the expected age of the allele, is the limit of (3.4) as $p \to 0$ thus

$$\mathbb{E}_x(\text{Age}) = -2x(1-x)^{-1}\log x. \tag{3.8}$$

They find the second moment of the age to be

$$x(1-x)^{-1}\log x - \int_0^x (1-z)^{-1}\log z dz. \tag{3.9}$$

They noted

> One additional property of neutral alleles which may be of interest from a mathematical standpoint is that the average age of a mutant allele having a current frequency y (assuming $p = 0$) is equal to the average time until extinction of the same allele (excluding the cases of its eventual fixation).

In upper part of Figure 3.2 a sample path of an allele currently at frequency x is shown. A sample path of an allele from a frequency x forward in time to loss is shown in lower part of Figure 3.2. Kimura and Ohta's (1973) observation is that the expected time of both of these events is the same, this observation is indicative of the reversibility results that followed later. Maruyama (1974) considered the age of an allele found at x in the population in two cases, the first is that of one way mutation away from the allele, $\mu_{21} > 0, \mu_{12} = 0$ for the process with the generator given in Section 2.2.2. The second case examined by Maruyama is that of no recurrent mutation but genic selection acting on the allele. When one way mutation acts on the frequency of the allele there are two possible interpretations of the age as the boundary at 1 is no longer absorbing. The first is that the age of the allele excludes those trajectories that have reached fixation at some point in the past, the second interpretation is that the age of the allele includes those trajectories that have reached fixation one or more times in the past. Expressions for the expected age and second moment of the age in the various cases were derived by Maruyama using the backward equation and sojourn times in an approach similar to that of Maruyama (1972). Maruyama (1974) notes that the age of an allele in the case of genic selection is the same irrespective of the sign of β. This is a generalisation of the observation that the time for an allele to fix in the population, conditional on it fixing, depends only on $|\beta|$ made in Maruyama (1972).

A diffusion process conditional on fixation, starting from near 0, and a diffusion process, initially close to 1, conditioned on loss were found to have identical sojourn times by Maruyama and Kimura (1974). This result is, once again, indicative of the underlying reversibility of the process, in this case for paths conditioned on reaching 1 before 0. The approach of Maruyama (1974) was generalised by Maruyama and Kimura (1975) to allow moments of the integral along sample paths of any function of gene frequency, $w(\zeta)$, to be evaluated conditional on the current frequency x. The reader is reminded

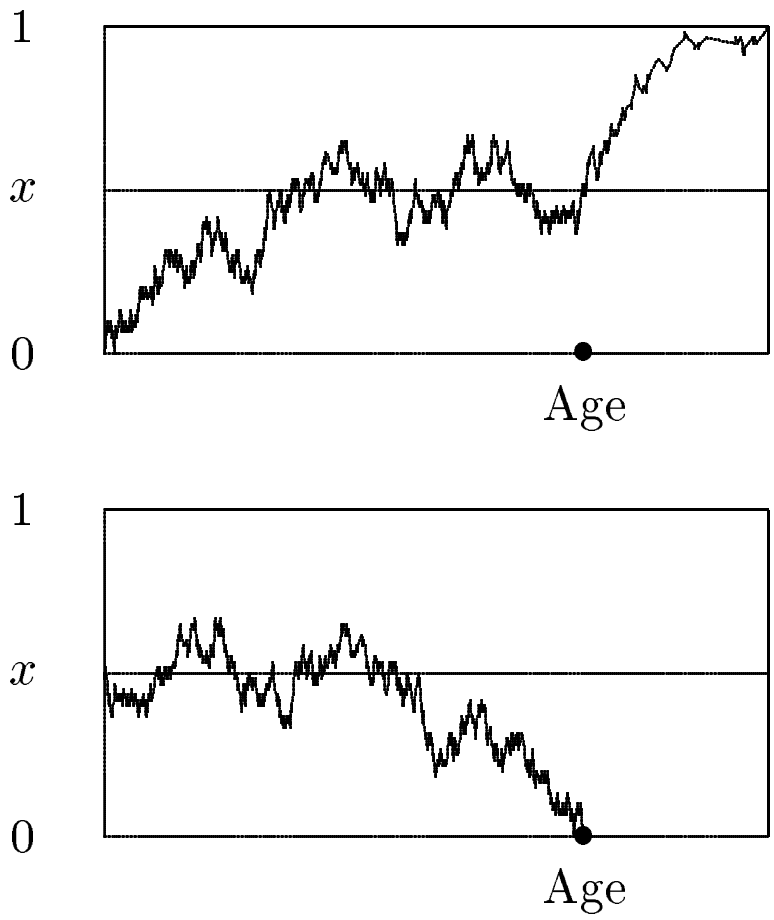

Figure 3.2: The upper half of the figure shows the trajectory of an allele forward in time from when it is introduced into the population, passing through a frequency interval $[x, x + dx]$, before going to fixation in this case. The lower half of the figure shows a trajectory of an allele forward in time starting from an initial frequency x, going to loss.

that T is the current time. In their general formulation the first moment of the integral of $w(\zeta)$ along paths, a generalised version of (3.4), is

$$\mathbb{E}\left(\int_0^T w\big(x(t)\big)dt \mid x(0) = p, x(T) = x\right) = F_x^{(1)}(p)/F_x^{(0)}(p) \tag{3.10}$$

where, by considering sample paths in the frequency interval $[\zeta, \zeta + d\zeta]$ at time t_1 on

the way to x from p,

$$\begin{aligned} F_x^{(1)}(p) &= \int_0^\infty \int_{t_1}^\infty f(p,\zeta,t_1)w(\zeta)f(\zeta,x,t_2-t_1)dt_2dt_1, && (3.11)\\ &= \int_0^1 G(p,\zeta)w(\zeta)G(\zeta,x)dx, && (3.12)\\ F_x^{(0)}(p) &= \int_0^\infty \int_{t_1}^\infty f(p,\zeta,t_1)f(\zeta,x,t_2-t_1)dt_2dt_1, && (3.13)\\ &= \int_0^\infty f(p,x,t)dt, && (3.14)\\ &= G(p,x), && (3.15) \end{aligned}$$

where $G(p,x)$ is the Green function of the process. To obtain (3.4), the time since the allele was at a frequency p given that the current frequency is x, $w(\zeta)$ is set equal to 1. (3.10) may be rewritten as

$$\mathbb{E}\left(\int_0^T w\big(x(t)\big)dt \mid x(0)=p, x(T)=x\right) = \int_0^1 w(\zeta)\frac{G(p,\zeta)G(\zeta,x)}{G(p,x)}d\zeta \qquad (3.16)$$

thus

$$G_x(p,\zeta) = \frac{G(p,\zeta)G(\zeta,x)}{G(p,x)} \qquad (3.17)$$

is the conditional Green function (or conditional sojourn time), i.e. the expected amount of time spent in a frequency interval $[\zeta,\zeta+d\zeta]$ conditional on $x(0)=p, x(T)=x$. For a general introduction to the conditional Green function see Maruyama (1977). In the limit as $p\to 0$ then (3.16) is the general formula for the expectation of the integral along sample paths of $g(x)$ since the allele arose till the current time T. This approach helps to elucidate the use of solutions to the backward equation conditional on the current frequency used in previous papers by the two authors. A difficulty noted by Sawyer (1977) is that (3.10) is not exact unless the present time T is defined.

The papers discussed so far stimulated other authors to examine the problem of retrospective properties of processes in population genetics. A diffusion process with killing

(see Section 2.2.2) was introduced to study expected values of functionals of the retrospective path by Sawyer (1977). Sawyer (1977) offered an interpretation of the current time T in (3.10). In his Bayesian formulation the age of the allele, τ, is found by

$$\mathbb{E}_x(Age) = \lim_{p\to 0}\lim_{T\to\infty} \mathbb{E}\left(\int_{T-\tau}^{T} dt \mid x(0) = p, x(T) = x\right), \tag{3.18}$$

where the prior on τ is uniformly distributed on $[0, T]$, i.e. the mutation is equally likely to have occurred at any time in the past. Levikson (1977) studied the distribution of the age of the allele utilising both Markov Chain and Diffusion return processes. In a Diffusion return process, $\{X_r(t) \mid t \geq 0\}$, when a sample path reaches either boundary it is immediately returned to some fixed interior point, p. Levikson (1977) postulated that the distribution of the age of the return process $A(s)$ found at a frequency x, i.e. the last time the process was returned, is

$$\mathbb{P}(A(s) \in (t, t + dt) \mid X_r(s) = x) = \frac{f(p, x, t)dt}{\int_0^\infty f(p, x, t)dt}, \tag{3.19}$$

as $p \to 0$ this is the distribution of the age of the allele. The use of return processes is thought desirable by many authors as discussion of average properties is felt to be more valid for a process that repeats itself than a single unique event. Ewens (1979) notes that the return process for a discrete state space model is similar to a process with a small recurrent mutation rate ϵ, as $\epsilon \to 0$. This approach is similar to one used by Griffiths (2003) discussed later.

A key idea in understanding many of the results in stochastic process theory as applied to genetics is the reversibility of the process. A process is defined by Watterson (1977) as reversible if:

> given the present state of a stochastic process, the statistical properties of its future behaviour are the same as those for its past history treated as a stochastic process with time running forward.

A continuous time Markov Chain $p_{ij}^{(t)} = P(X(t) = j \mid X(0) = i)$ is reversible if and only if for all i and j

$$\pi_i p_{ij}^{(t)} = \pi_j p_{ji}^{(t)}, \tag{3.20}$$

where π_i is the stationary distribution of the process at state i. (3.20) is termed the detailed balance equation (see Feller (1968)). (3.20) is equivalent to the simpler equation that the Markov Chain is reversible if for all i and j

$$\pi_i q_{ij} = \pi_j q_{ji}, \tag{3.21}$$

where the matrix $\mathbf{Q}$, with elements $\{q_{ij}\}$, is the rate matrix of the Markov Chain. Processes in population genetics are often not strictly reversible. The forward process without knowledge of the boundary it arose from is reversible. However, in the case when the allele of interest is known to be derived (i.e. $X(0) = 0$) then the forward process must be conditioned on $X(\infty) = 0$ in order to match the backward process. The term reversibility in population genetics is often abused to include those processes describing derived alleles where the statistical properties of the forward path are identical to those of the path backward in time conditional on the loss of the allele forward in time. These processes should perhaps be termed 'quasi-reversible' but unfortunately this distinction has not been in the literature.

Watterson (Watterson, 1976, 1977) first used reversibility to study the distribution of the age of an allele. Watterson (1977) observed that a return process has a stationary distribution as the boundaries are no longer absorbing, see Ewens (1963) for the derivation. The transition density of many one dimensional diffusion processes may be written as

$$f(x, y, t) = m(y)h(x, y, t) \tag{3.22}$$

where $m(y)$ is the speed measure of the diffusion process and $h(x, y, t)$ is a symmetric function in x and y, i.e. $h(x, y, t) = h(y, x, t)$. Often $h(x, y, t)$ can be written in a

spectral expansion representation in terms of a set of functions orthogonal to $m(x)$, see Section 13 of Chapter 15 of Karlin and Taylor (1981). Let $\mathbb{B}(\cdot, a, b)$ denote a beta distribution with parameters a and b. The transition function of the one dimensional neutral Wright-Fisher diffusion process, with no mutation, can be written as

$$f(x,y,t) = m(y)\sum_{k=2}^{\infty} d_k(t)\sum_{\alpha=1}^{k-1}\binom{k}{\alpha}x^{\alpha}(1-x)^{k-\alpha}\mathbb{B}(y;\alpha+1,k-\alpha+1), \quad (3.23)$$

where $d_k(t)$ is the distribution of the number of lines, k, at time t in a pure death process where lines die at rate $\binom{k}{2}$, $k = 2, 3, \cdots$, with an entrance boundary at ∞. This death process is called the lines of descent process (Griffiths, 1980) and is the genealogical process underlying the neutral diffusion process (see Tavaré (1984) for a discussion of the line-of-descent process and its relationship to the coalescent). For a model of genic selection the transition density may be written as

$$\begin{aligned} f(x,y,t) &= m(y)\sum_{k=2}^{\infty}\sum_{\alpha=1}^{k-1} d_{k-\alpha,\alpha}(t)\binom{k}{\alpha}x^{\alpha}(1-x)^{k-\alpha}c(k,\alpha)y^{\alpha}(1-y)^{k-\alpha}, \\ c(k,\alpha)^{-1} &= \int_0^1 m(\zeta)\zeta^{\alpha}(1-\zeta)^{k-\alpha}d\zeta \end{aligned} \quad (3.24)$$

where $d_{k-\alpha,\alpha}(t)$ is a process with births and deaths, with an entrance boundary of ∞ partitioned into x and $1-x$ (Barbour et al., 2000). The diffusion process analogue of the detailed balance equation (Keilson, 1965)

$$m(x)f(x,y,t) = m(y)f(y,x,t) \quad (3.25)$$

is satisfied. $f(x,y,t)$ is reversible with respect to the speed measure of the diffusion, i.e. the stationary distribution of the return process. An analogous form of (3.21) also exists for diffusion processes relating the generators of the forward and time reversed processes. Using (3.19) and (3.25) Watterson (1977) showed that the distribution of the age of the allele could be obtained by a reversibility argument i.e.

> the past age density equals the future absorption time density, (Watterson, 1977).

Let $\{X_m(t) \mid t \geq 0\}$ be a diffusion process with population scaled mutation rate μ_{12} and μ_{21} between alleles. The generator of this process is given in Section 2.2.2. This process has stationary distribution

$$\pi(x) = \frac{x^{\mu_{21}-1}(1-x)^{\mu_{12}-1}s(x)}{\int_0^1 x^{\mu_{21}-1}(1-x)^{\mu_{12}-1}s(x)dx}. \tag{3.26}$$

$\{X_m(t) \mid t \geq 0\}$ is reversible with respect to the stationary distribution $\pi(x)$. Let $\{X_u(t) \mid t \geq 0\}$ be the process with generator (3.1) where the allele is not known to be ancestral or derived, i.e. $X(0) = 0, 1$. Thus an interpretation of the $\{X_u(t) \mid t \geq 0\}$ reversibility with respect to $m(x)$ is that $\{X_u(t) \mid t \geq 0\}$ is the limiting process of $\{X_m(t) \mid t \geq 0\}$ as $\mu_1, \mu_2 \to 0$, and that $m(x)$ is the limiting un-normalised distribution of $\pi(x)$. A different interpretation of $m(x)$ is that it is the transient pseudo-distribution (3.2) of the diffusion process when the state of the allele as ancestral or derived is unknown i.e.

$$f(x \mid x(0) = 0, 1) \propto m(x)u_0(x) + m(x)u_1(x) = m(x). \tag{3.27}$$

Thus, even if the mutation is a unique event the process $\{X(t) \mid t \geq 0\}$ can be said to be reversible with respect to the sample path average transient pseudo-distribution. Once the process has been reversed then the forward path can be conditioned on the allele eventually being lost from the population, $X(\infty) = 0$ thus matching the backward path having initial condition $X(0) = 0$.

Nagasawa and Maruyama (1979) and Tavaré (1979) formalised the reversibility argument of Watterson (1977). Nagasawa and Maruyama (1979) discuss the links between the forward in time methods (i.e. the conditional Green function (3.17)) and the reversed process, and offer an interpretation of the current time T in (3.10). They noted that T can be thought of as the last time, before loss or fixation, that the state x is visited, but that this is approximately the same for large T as for the current time, see van Herwaarden and van der Wal (2002) for discussion. This explanation however perhaps

lacks the elegance of the Bayesian interpretation offered by Sawyer (1977).

Let $\{X^*(t),\ t \geq 0\}$ denote the trajectory of the time reversed process where $t = 0$ is the present day and $X(0) = x$. Nagasawa and Maruyama (1979) derive the transition density and generator of the time reversed diffusion process. They find that the transition density of the time reversed diffusion process, $f^*(x, y, t)$, is the identical to the forward process conditioned on hitting zero, thus

$$f^*(x, y, t) = f(x, y, t)\frac{u_0(y)}{u_0(x)}. \tag{3.28}$$

The generator of $\{X^*(t),\ t \geq 0\}$, $\mathcal{L}^*$, is similar to $\mathcal{L}$ with $\mu(x)$ replaced by

$$\mu^*(x) = \mu(x) - \frac{s(x)\sigma^2(x)}{\int_x^1 s(u)du}. \tag{3.29}$$

For example, in the case of a model of genic selection, i.e. $\beta(x) = \beta$,

$$\mu^*(x) = -\frac{1}{2}\beta x(1-x)\coth\big(\frac{1}{2}\beta(1-x)\big), \tag{3.30}$$

which is a symmetric function of β, i.e. it depends only on $|\beta|$. Thus the observation of Maruyama (1974) that the age of an allele, found at a frequency x, experiencing genic selection depends on $|\beta|$ alone can be understood in terms of the conditional process and seen to hold for all properties of the retrospective path $\{X^*(t),\ t \geq 0\}$. Another interpretation of the reversibility of the process, introduced here, is gained by considering a process, with a transition density $f^*(x, y, t)$, which can be seen to satisfy

$$m(x)u_0(x)f(x, y, t) = m(y)u_0(y)f^*(y, x, t). \tag{3.31}$$

(3.31) is not the detailed balance equation in that the transition density of the starred process is not the transition density of the forward process. (3.31) could be thought of as a quasi-detailed balance equation in that the process, with transition density $f(x, y, t)$ is quasi-reversible with respect to $m(x)u_0(x)$, the stationary distribution of the return

process (or in a nonrecurrent setting the transient pseudo-distribution for a derived allele).

A different approach to the problem of the unique event of the allele arising and its path before loss or fixation was developed by Sawyer and Hartl (1992). The authors considered a model of an infinite number of unlinked sites, in a Wright-Fisher population, with population size N. A mutation arises in the population with probability v_N per generation. v_N is related to the population scaled mutation rate μ, $Nv_N \propto \mu$. Once a mutation has taken place at a site no subsequent mutation occurs at this site. The frequency of the mutant allele at the i^{th} site k generations after the mutation arose is $X^N_{i,k}$, and as the sites are unlinked the processes $\{X^N_{i,k}\}$ are noninteracting. The process is taken to the diffusion limit and has a generator (3.1). Sawyer and Hartl wish to integrate the expectation of a function $w(y)$, where $w(0) = w(1) = 0$, over y weighted by the number of sites where the frequency of the allele is in the interval $[y, y + dy]$ at the current time ($0 \leq y \leq 1$). This can be evaluated by

$$\lim_{N\to\infty} \sum_{j=1}^{N-1} w(j/N) v_N \sum_{i=1}^{\infty} \sum_{k=0}^{\infty} P(X^N_{i,k} = j/N \mid X_{i,0} = 1/N) \tag{3.32}$$

$$= \mu \int_0^1 w(y) m(y) u_0(y) dy. \tag{3.33}$$

This approach leads to the Poisson random field method to analyse data from unlinked sites in the genome and evaluate the average selection coefficient of a nonsynomous mutations at unlinked sites in the genome (Sawyer and Hartl, 1992). The Poisson random field approach has been extended to estimate the distribution of selective effects (Bustamante et al., 2003); there the selection parameter β, of an allele is drawn from a distribution and Bayesian and Maximum likelihood methods to estimate the parameters of the distribution of selective effects are considered. The method of Sawyer and Hartl (1992) of considering infinitely-many-unlinked sites is a product state space approach and perhaps gives a clearer biological intuition than the return process and more

rigorous state space for the process averaged over in deriving the transient pseudo-distribution (3.2). The distribution of the age of an allele found at frequency x in the population will now be informally derived using the infinitely-many-unlinked sites framework. $\delta_{a,b}$ denotes the Kronecker delta ($\delta_{a,b} = 1$ if $a = b$ and is zero otherwise). Letting $w(j/N, k) = \delta_{j/N,x}$ only the two inner summations in (3.32) remain

$$\lim_{N\to\infty} v_N \sum_{i=1}^{\infty}\sum_{k=0}^{\infty} P(X_{i,k}^N = x \mid X_{i,0} = 1/N) = \mu \lim_{p\to 0}\int_0^{\infty} f(p, x, t)dt, \tag{3.34}$$

thus from (3.33) this is equal to

$$\mu m(x)u_0(x)dx. \tag{3.35}$$

This can be interpreted as the expected number of alleles at independent sites (having arisen at some point in the past) that are currently in the frequency interval $[x, x + dx]$, this is analogous to $f(x)dx$. If instead of $w(j/N) = \delta_{j/N,x}$, we let $w(j/N, k) = \delta_{j/N,x}\delta_{k,t}$ then (3.32) becomes

$$\lim_{N\to\infty} v_N \sum_{i=1}^{\infty} P(X_{i,t}^N = x \mid X_{i,0} = 1/N) = \mu \lim_{p\to 0} f(p, x, t)dx. \tag{3.36}$$

This is the expected number of alleles at independent sites that have arisen at a time t in the past and currently are in the frequency interval $[x, x + dx]$. By a relative rate argument the distribution of the age of an allele found at a current frequency x in the population is

$$\lim_{p\to 0} \frac{\mu f(p, x, t)dx}{\mu m(x)u_0(x)dx}. \tag{3.37}$$

While the infinitely-many-unlinked sites model of Sawyer and Hartl (1992) uses forward in time arguments, a backward in time argument using reversibility based on the average occupancy of a frequency interval, (3.35), would be possible. This form of reversibility would have much in common with the quasi-reversibility possible using the transient pseudo-distribution as an unnormalised stationary distribution, see (3.31).

Reversibility of a diffusion process with killing, see Section 2.2.2, is considered by Griffiths (2003) and Patterson (2004). Both consider a process that describes the time to the first coalescence in k lineages that coalesce at rate $\binom{k}{2}/X^*(t)$, where $\{X^*(t) \mid t \geq 0\}$ is a diffusion process, with a forward in time generator as in (3.1). The generator of the process with killing is $\mathcal{L}^* - \binom{k}{2}/x$. This process is similar to that described in Barton et al. (2004) and Barton and Etheridge (2004), discussed in Section 2.5.2, but their diffusion process has recurrent mutation and thus a stationary distribution. Patterson studies the case of $\mu(x) = 0$ while Griffiths (2003) considers the general case. Griffiths (2003) discusses the results found in relation to those of Griffiths and Tavaré (2003) derived in a coalescent framework.

3.2 Properties of an Allele Segregating in a Sample

Results of Griffiths (2003) are now discussed in detail, as they underpin much of the work in this thesis. In Griffiths' work the sample path $\{X(t) \mid t \geq 0\}$ is a single path and reversibility with respect to $m(x)$ is given the interpretation of being the sample path average rate of occupance of $[x, x + dx]$, i.e. interpretation given in (3.27). This sample path average approach utilised by Griffiths is close to that studied by Kimura and Maruyama in a number of papers, in that average properties are discussed without recourse to a processes to average over as in the Poisson random field or return process approach. The population frequency is often unknown; more frequently the allele is known to be present in n out of $n + m$ genes. Griffiths (2003) considers retrospective properties of $\{X(t) \mid t \geq 0\}$ in a general diffusion model, where only a sample from the population is known.

Consider the rate that sample paths of diffusion process, at time t, produce the sample

n of the selected allele out of $n+m$ genes. The joint probability of the sample and density of the frequency x at time t is

$$\binom{n+m}{n} x^n (1-x)^m f(p,x;t). \tag{3.38}$$

Integrating with respect to x gives

$$P(n, n+m \mid p, t) = \int_0^1 \binom{n+m}{n} x^n (1-x)^m f(p,x;t) dx, \tag{3.39}$$

this is the probability of obtaining the sample at time t. The density of the population frequency at time t given the sample frequency is

$$\begin{aligned} f(x \mid n, n+m, p, t) &= \frac{f(x, n, n+m \mid p, t)}{P(n, n+m \mid p, t)}. \\ &= \frac{\binom{n+m}{n} x^n (1-x)^m f(p,x;t)}{\int_0^1 \binom{n+m}{n} x^n (1-x)^m f(p,x;t) dx} \\ &= \frac{m(p)}{m(p)} \frac{\binom{n+m}{n} x^n (1-x)^m f(p,x;t)}{\int_0^1 \binom{n+m}{n} x^n (1-x)^m f(p,x;t) dx}. \end{aligned} \tag{3.40}$$

Using the reversibility of the process (3.25) the probability of obtaining the sample along paths can now be considered forward in time from x to p,

$$\frac{\binom{n+m}{n} x^n (1-x)^m m(x) f(x,p;t)}{\int_0^1 \binom{n+m}{n} x^n (1-x)^m m(x) f(x,p;t) dx}. \tag{3.41}$$

Integrating (3.39) with respect to t the total rate at which diffusion paths give the sample, can be found

$$f(n, n+m \mid p) = \int_0^\infty \int_0^1 \binom{n+m}{n} x^n (1-x)^m m(x) f(x,p;t) dx dt \tag{3.42}$$

We now find the joint density at a time t of the population frequency given the sample configuration. Through conditioning we arrive at

$$\begin{aligned} & \frac{\binom{n+m}{n} x^n (1-x)^m m(x) f(x,p;t)}{\int_0^\infty \int_0^1 \binom{n+m}{n} x^n (1-x)^m m(x) f(x,p;t) dx dt}, \\ = \; & \frac{\binom{n+m}{n} x^n (1-x)^m m(x) f(x,p;t) u_0(x)}{u_0(x) \int_0^\infty \int_0^1 \binom{n+m}{n} x^n (1-x)^m m(x) f(x,p;t) dx dt}. \end{aligned} \tag{3.43}$$

Let $\phi_x(t)$ denote the density of the time to be absorbed at 0 from a frequency x conditional on absorption at 0. The following two limits are now used

$$\lim_{p\to 0} \frac{f(x,p;t)}{u_0(x)} = \phi_x(t), \tag{3.44}$$

and

$$\lim_{p\to 0} \int_0^\infty f(x,p;t)dt = u_0(x). \tag{3.45}$$

The first limit states that the distribution of the time to 0 is the limit of probability conditional on hitting zero ($X(\infty) = 0$) of being at a frequency p at a time t as $p \to 0$. The second limit is that the probability of hitting 0 before 1 is the limit the transition function density at a frequency p, as $p \to 0$, integrated over all t. The limit of (3.43) using (3.44) and (3.45) as $p \to 0$ is

$$\frac{\binom{n+m}{n} x^n (1-x)^m m(x) u_0(x) \phi_x(t)}{\int_0^1 \binom{n+m}{n} x^n (1-x)^m m(x) u_0(x) dx}. \tag{3.46}$$

The density of the age of the mutation given the sample configuration is found by integrating (3.46) over x

$$\frac{\int_0^1 x^n (1-x)^m m(x) u_0(x) \phi_x(t) dx}{\int_0^1 x^n (1-x)^m m(x) u_0(x) dx}, \tag{3.47}$$

and the expectation of the age of the mutation given the sample is

$$\frac{\int_0^1 x^n (1-x)^m m(x) u_0(x) \mathbb{E}_x(Age) dx}{\int_0^1 x^n (1-x)^m m(x) u_0(x) dx}. \tag{3.48}$$

This can be numerically solved using Gaussian quadrature, and has been implemented by myself for the analysis in (Römpler et al., 2004). The density of the population frequency given the sample is found by integrating (3.46) over t

$$f(x \mid n, n+m) = \frac{\binom{n+m}{n} x^n (1-x)^m m(x) u_0(x)}{\int_0^1 \binom{n+m}{n} x^n (1-x)^m m(x) u_0(x) dx}. \tag{3.49}$$

A Bayesian interpretation of (3.49) is that $m(x)u_0(x)$ is an improper prior on the population frequency x, this prior is the pseudo-transient distribution (3.2).

The Figures 3.3, 3.4 and 3.5 show the frequency spectrum given a sample of 10 genes with different numbers ($5, 3$ and 7) of the selected allele found in the sample. In each of these figures a $\mathbb{B}(\cdot, \cdot)$ distribution is shown as a solid line for comparison to the posterior when under the prior all frequencies were equally likely. In Figure 3.3 the neutral ($\beta = 0$) line peaks at just above 0.4. Under neutrality $m(x)u_0(x) = \frac{1}{x}$, thus the frequency distribution is $\mathbb{B}(n, m + 1)$, this peaks lower than expected if all frequencies were equally likely. The two lines for strong selection ($\beta = 50, 100$) both peak close to 0.5 as the population frequency distribution $\rightarrow \frac{1}{x(1-x)}$ as $\beta \rightarrow \infty$ thus they peak at approximately the same point as the $\mathbb{B}(6, 6)$ shown in a sold line, however they are more diffuse distributions.

In Figure 3.4 the distributions given the sample of 3 mutant alleles out of 10 sampled are shown. Their peaks are shifted downwards with respect to $\mathbb{B}(4, 8)$ as a selected allele spends little time in the intermediate frequencies and a neutral allele is unlikely to reach them. Figure 3.5 shows how a sample of 7 out of 10 affects the frequency spectrum. The neutral line is shifted down as it is unlikely that a neutral allele would be at a frequency as high as 0.70. The strong selection lines are shifted to high frequencies because the selected allele spends more time at high and low frequencies than at those in the middle.

The Figure 3.6 shows how larger samples quickly overcome the shift towards lower frequencies in a neutral frequency spectrum given the sample. The prior implies the frequency of 0.5 is unlikely to be reached by a neutral trajectory but with a sample of 10 out 20 the posterior frequency is quite likely to be in the region of 0.5.

$f(x \mid n, n + m)$ can also be derived using forward in time arguments and the Green

function, which is completely equivalent to the approach outlined above. The amount of time spent in a small frequency interval $(x, x+dx)$ given $x(0) = p$ is $G(p,x)dx$ (see equation (2.26)). Thus the rate that a particular frequency x gives rise to the sample is the product of the expectation of the amount of time spent in $(x, x+dx)$ and the probability that the sample was taken when the population frequency was x. As the initial frequency $p \to 0$ this is

$$2\binom{m+n}{n} x^n (1-x)^m u_1(p) m(x) \int_x^1 s(w)dw, \quad p < x \leq 1, \tag{3.50}$$

noting that the Green function for $x < p$ is negligible as $p \to 0$. The rate that any path generates the sample is the integral over (3.50) with respect to x

$$2\binom{m+n}{n} u_1(p) \int_0^1 x^n (1-x)^m m(x) \int_x^1 s(w)dwdx \tag{3.51}$$

The density of the frequency conditional on the sample is given by a relative rate argument, i.e. the rate that the frequency x gives rise to the sample compared to any population frequency

$$\frac{2\binom{m+n}{n} x^n (1-x)^m u_1(p) m(x) \int_x^1 s(w)dw}{2\binom{m+n}{n} u_1(p) \int_0^1 x^n (1-x)^m m(x) \int_x^1 s(w)dwdx}. \tag{3.52}$$

Obtaining this result via the Green function is similar to the way this result could be derived in a Poisson random field approach. Although this result is easier to gain access to, it lacks the elegance of the above approach that allows the age conditional on the sample and the frequency spectrum of the sample to be found simultaneously.

3.3 The Moran Model and Reversibility

The diffusion process can be approximated with a Moran model having a population size of N genes to simulate a trajectory. Then, forward in time, the number of copies

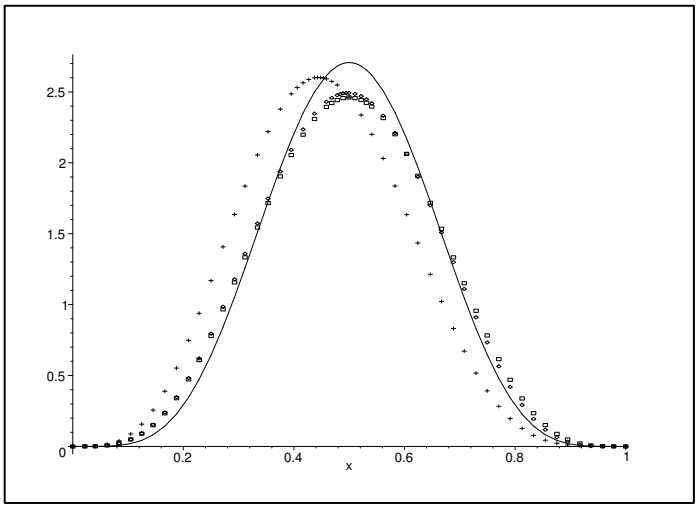

Figure 3.3: The frequency spectrum of a derived mutation conditional on observing a sample of 10 containing 5 copies of the selected allele under a variety of selection coefficients: $\beta = 0$ (+), $\beta = 10$ ($\diamond$), $\beta = 100$ ($\square$) and the solid line is a $\beta(6, 6)$ distribution.

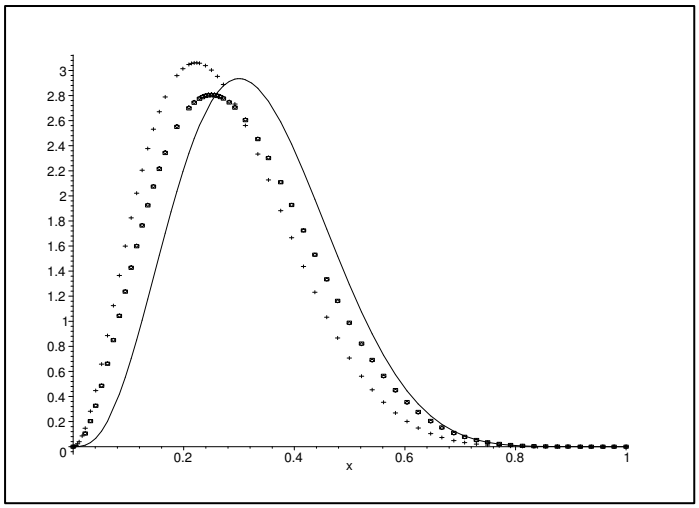

Figure 3.4: The frequency spectrum of a derived mutation conditional on observing a sample of 10 containing 3 copies of the selected allele under a variety of selection coefficients: $\beta = 0$ (+), $\beta = 10$ ($\diamond$), $\beta = 100$ ($\square$) and the solid line is a $\beta(4, 8)$ distribution.

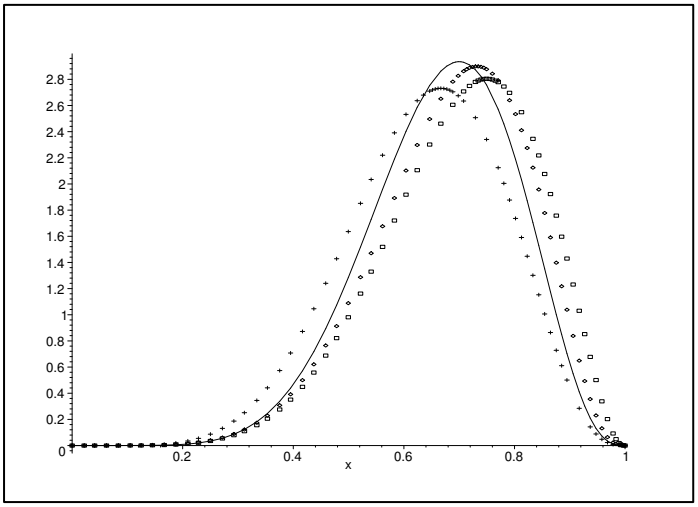

Figure 3.5: The frequency spectrum of a derived mutation conditional on observing a sample of 10 containing 7 copies of the selected allele under a variety of selection coefficients: $\beta = 0$ (+), $\beta = 10$ ($\diamond$), $\beta = 100$ ($\square$) and the solid line is a $\beta(8, 4)$ distribution.

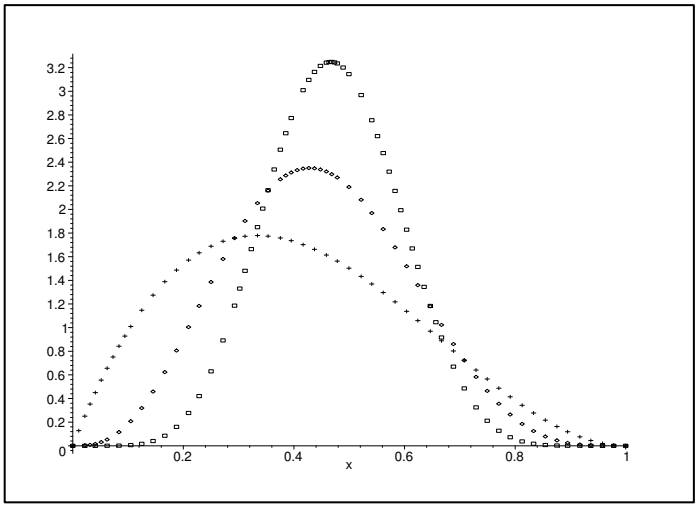

Figure 3.6: The frequency spectrum of a neutral derived mutation conditional on observing a variety of samples: 2 out of 4 $(+)$, 5 out of 10 ($\diamond$), and 10 out of 20 ($\square$).

of the mutant allele $\{Z(t), t \geq 0\}$ is governed by a continuous time birth and death process. A fixed population size of N genes is kept so that at time t the number of mutant genes is $Z(t)$ and the number of non-mutant genes is $N - Z(t)$. The states $z = 0$ and $z = N$ are absorbing states. The Moran model approximating the general diffusion process, see for example Griffiths (2003), has birth and death rates that are respectively, with $x = j/N$,

$$\lambda_j = \frac{1}{2}N(N\sigma^2(x) + \mu(x)),\ \mu_j = \frac{1}{2}N(N\sigma^2(x) - \mu(x)). \tag{3.53}$$

In the case of frequency dependent selection

$$\lambda_j = \lambda^{(N)}(1 + \frac{1}{2}s_N(j))j(N-j)/N,\ \mu_j = \lambda^{(N)}(1 - \frac{1}{2}s_N(j))j(N-j)/N,$$

where $\lambda^{(N)} = N/2$ and $s_N(j) = \beta(x)/2N$. Reproduction in this model is at rate $\lambda^{(N)}(1 - s_N(j)/2)$ per non-mutant gene, where a non-mutant gene is chosen at random to reproduce and a gene (of either type) is chosen at random to die, and similarly with rate $\lambda^{(N)}(1 + s_N(j)/2)$ for mutant genes. In the Moran model the probability of absorption into state 0 from an initial state z (Karlin and Taylor, 1975) is

$$u_0^N(z) = \frac{\sum_{i=z}^{N-1}\prod_{j=1}^{i}\left(\frac{\mu_j}{\lambda_j}\right)}{1 + \sum_{i=1}^{N-1}\prod_{j=1}^{i}\left(\frac{\mu_j}{\lambda_j}\right)}. \tag{3.54}$$

We choose to approximate $\{X^*(t), t \geq 0\}$ by first approximating $\{X(t), t \geq 0\}$ by a Moran Model as $\{Z(t), t \geq 0\}$ and then reversing the Moran Model $\{Z^*(t), t \geq 0\}$. An alternative approach would be to approximate $\{X^*(t), t \geq 0\}$ directly by simulating from a Moran Model with rates containing $\mu^*(x)$ instead of $\mu(x)$. Informally, both of these methods of approximating the reversed diffusion process are the same in the large population limit. Once again as in the previous section quasi-reversibility is needed to allow this method to be valid. In showing this reversibility Griffiths (2003) is summarised.

In the Moran model the probability of moving from i to j in time t is $p_{i,j}^{(t)}$, this is the analogy of the transition function for the diffusion process. To show the quasi-reversibility of the Markov Chain we need to show for $0 < z < N$ that

$$\begin{aligned} p_{1,z}^{(t)} &= p_{z,1}^{(t)} \mu_1 \rho_z, && (3.55) \\ &= \rho_z u_0^N(z) g_z(t), && (3.56) \\ \text{where } \rho_z &= \frac{\lambda_1 \cdots \lambda_{z-1}}{\mu_1 \cdots \mu_z}. \end{aligned}$$

$g_z(t)$ is the distribution of the time to loss conditional on loss. Informally $\mu_1 \rho_z$ is equivalent to the ratio of the stationary distributions if they existed, and this is a detailed balance equation similar to equation (3.20) in the presence of a pseudo-transient distribution.

Consider a Markov Chain (this process is called the ϵ-chain) with rate ϵ of escaping the absorbing states i.e.

$$\lambda_0^\epsilon = \epsilon, \qquad \mu_N^\epsilon = \epsilon. \tag{3.57}$$

The remaining birth and death rates are the same as in the standard Moran model

$$\lambda_z^\epsilon = \lambda_z, \qquad \mu_z^\epsilon = \mu_z \quad 0 < z < N. \tag{3.58}$$

This ϵ-chain has a stationary distribution, as it has non-absorbing boundaries. The stationary distribution takes the form

$$\pi_z^\epsilon = C \frac{\lambda_0^\epsilon \cdots \lambda_{z-1}^\epsilon}{\mu_1^\epsilon \cdots \mu_{z-1}^\epsilon}, \tag{3.59}$$

where C is a constant. It is also reversible with respect to the stationary distribution, and the detailed balance equation is

$$\begin{aligned} p_{1,z}^{(t),\epsilon} &= p_{z,1}^{(t),\epsilon} \frac{\pi_z^\epsilon}{\pi_1^\epsilon} && (3.60) \\ &= p_{z,1}^{(t),\epsilon} \mu_1 \rho_z, && (3.61) \end{aligned}$$

letting $\epsilon \to 0$ (3.56) is obtained, thus showing that the chain with absorbing states is reversible.

As the allele is derived in the frequency must have arisen from zero looking backwards in time, and thus the reversed process must be conditioned on reaching 0 before N forward in time. The probability of going to loss from a state z is (3.54). We need now only weight the transition probabilities conditional on going to loss, to have a Markov chain that approximates $\{X^*(t), t \geq 0\}$. The birth and death process $\{Z^*(t), t \geq 0\}$ has rates

$$\lambda_z^* = \lambda_z \frac{u_0^N(z+1)}{u_0^N(z)}, \quad \mu_z^* = \mu_z \frac{u_0^N(z-1)}{u_0^N(z)}, \quad 0 < z < N. \tag{3.62}$$

As

$$u_0^N(z) = \frac{\lambda_z}{\lambda_z + \mu_z} u_0^N(z+1) + \frac{\mu_z}{\lambda_z + \mu_z} u_0^N(z-1) \tag{3.63}$$

it follows that

$$\lambda_z^* + \mu_z^* = \lambda_z + \mu_z, \tag{3.64}$$

thus the overall rate of events is left unchanged. From this point on we are only concerned with reversed process, so abuse notation by omitting the * superscript. Now $\{Z(t),\ t \geq 0\}$ and $\{X(t),\ t \geq 0\}$ will denote reversed processes. A range of trajectories with different genic selection values can be seen in Figure 3.7. The slow convergence to a deterministic trajectory can be seen through the sequence of $a - f$. The trajectory behaves deterministically at intermediate frequencies for $\beta >> 1$. At lower frequencies the trajectory is still stochastic at $\beta = 100$, so the variance in the age of the mutation will still be appreciable even at high β.

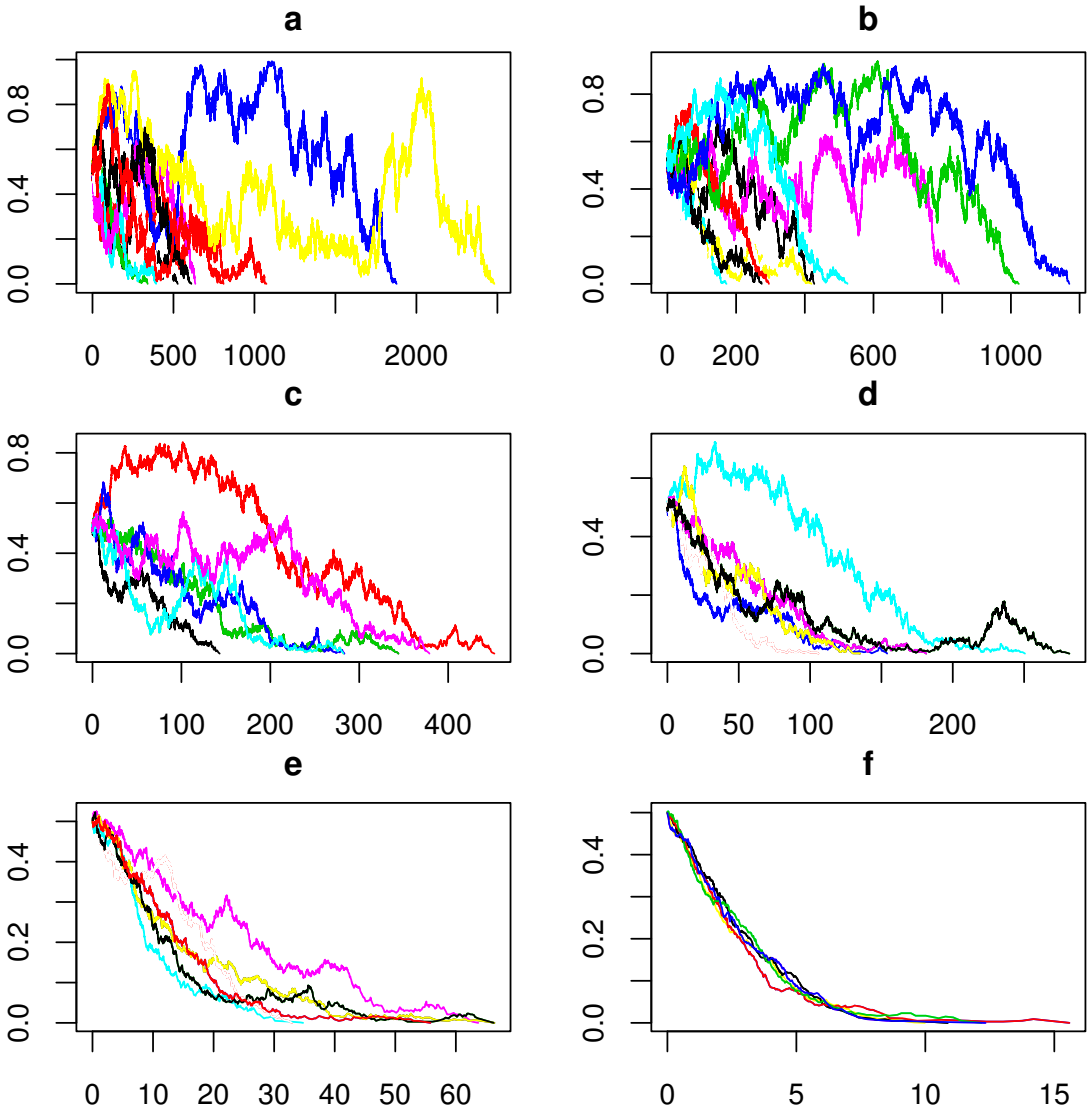

Figure 3.7: A sample of twenty trajectories generated by a backward in time Moran model conditioned on loss from a present day frequency of .5 in a population of 1000 for a variety of selection coefficients a) neutral $\beta = 0$, b) $\beta = 0.1$, c) $\beta = 1$ d) $\beta = 25$ e) $\beta = 50$, f) $\beta = 100$. Time is given along the x-axis and number with the selected mutation in the population on the y-axis

Chapter 4

Simulation of Coalescent Genealogies with Selection

4.1 The Coalescent and the Moran Model

Consider a sample of n genes carrying the selected allele (with selection coefficient $s_N(j)$ when there are j alleles in the population) and m alleles of the ancestral non-selected type. In the Moran model let $\left(A_n^N(t), A_m^N(t)\right)$ be the process describing the number of ancestors of a sample of (n, m) genes at time t in the past, measuring time backward from the present day. We now outline a derivation of the behaviour of this process due to Griffiths (2003), similar outlined proofs for the Wright-Fisher Model can be found in Section 2.5.2. Conditional on $\left(A_n^N(t) = \zeta, A_m^N(t) = \eta\right)$ and that there are currently z copies of the selected allele the probability of a coalescence within the selected ancestors of the sample is

$$\lambda^N(1 + s_N(j)) \left\{ z.\frac{z-1}{N}.\frac{\binom{\zeta}{2}}{\binom{z}{2}} + z.\left(1 - \frac{z}{N}\right).\frac{\binom{\zeta}{2}}{\binom{z+1}{2}} \right\} dt, \tag{4.1}$$

$$\sim \binom{\zeta}{2} x(t)^{-1} dt, \tag{4.2}$$

where $x(t) = z/N$. The first term of (4.1) relates to a birth by one the selected allele genes that replaces one of the other genes of the selected type in the population, and that birth creates a coalescence within the sample. The second term of (4.1) represents the birth of a selected offspring which replaces one of the non-selected genes in the population resulting in a coalescence in the sample. The probability of a coalescence in the nonselected sample is by a similar argument

$$\sim \binom{\eta}{2}(1 - x(t))^{-1}ds \tag{4.3}$$

Thus the coalescent with selection can be viewed as a variable population size model (see Section 2.3.1) with the relative sizes back in time being determined by the segregation of the selected mutation (i.e. the trajectory). The trajectory therefore holds all the information on selection.

Conditional on the process $X(t)$ the ancestral processes are distributed with $0 < t < \tau$ (where τ is the time of loss of the mutation from the population) as

$$\begin{aligned} A_n(t) &= A_n^\circ\left(\int_0^t \frac{1}{X(u)}du\right) \\ A_m(t) &= A_m^\circ\left(\int_0^t \frac{1}{1 - X(u)}du\right) \end{aligned} \tag{4.4}$$

where A_n° and A_m° are independent death processes with rates when there are k lineages of $\binom{k}{2}$ i.e. the original coalescent process.

4.2 Simulating a Coalescent History

This section describes a method to generate a coalescent genealogy for a sample of $n + m$ fully-linked sequences where n sequences contain a selected site with selection coefficient β, and m sequences are selectively neutral. The population scaled mutation rate, in a diploid population, for the whole region is $\theta = 4N_e\mu$ where N_e is the effective

population size and μ is the mutation rate per sequence per generation. Time is measured in a coalescent scale of $2N_e$ generations. Note that effective population size, N_e is different from the population size N in an approximating Moran model. N is chosen to be large enough to obtain a good diffusion process approximation, consistent with the amount of computation and storage required. A simple 'rule of thumb', automated in the software, is to use an N that gives a $\beta/N << 1$, with a minimum N of 300. In general too smaller N will not approximate the diffusion process well, especially near the boundaries, and may lead to an incorrect likelihood surface.

Let X be the population frequency of the selected allele at the present time. Griffiths (2003) shows that the density of X, conditional on n and m is

$$f_{n,m}(x) = \frac{\binom{n+m}{n} x^n (1-x)^m m(x) u_0(x)}{\int_0^1 \binom{n+m}{n} x^n (1-x)^m m(x) u_0(x) dx}, \tag{4.5}$$

where $u_0(x)$ is the probability that $\{X(t),\ t \geq 0\}$ is absorbed at 0 when $X(0) = x$, and $m(x)$ is the speed measure of the process. Explicit expressions are $u_0(x) = \int_x^1 s(y)dy \big/ \int_0^1 s(y)dy$, and $m(x) = \left[\sigma^2(x)s(x)\right]^{-1}$. The distribution of $\{X(t),\ t \geq 0\}$ is thus a diffusion process with a random initial frequency $X(0)$ having density (4.5). The coalescent structure of the two subgroups of n and m genes behaves as a subdivided population model with respective variable population size proportions of $X(t)$ and $1 - X(t)$ at time t (Kaplan et al., 1988). That is, if a and b are the number of ancestors of the two groups of genes at time t, then coalescence rates in the two groups are respectively $\binom{a}{2} X(t)^{-1}$ and $\binom{b}{2} \big(1 - X(t)\big)^{-1}$.

Events in a coalescent history can be either mutations or coalescences within the respective selected and nonselected ancestral subsamples. The time to the next event t given that the last event occurred at time u has a non-homogeneous exponential distribution, this non-homogeneity is due to the dependence on the trajectory of the selected

allele. The density function of this distribution is

$$g_{12}(t \mid n, m, u) \;=\; \gamma(t, n, m) \exp\left(- \int_u^t \gamma(s, n, m) ds \right), \; u < t < \infty \tag{4.6}$$

where at time s, $\gamma(s, n, m) = \gamma_1(s, n) + \gamma_2(s, m)$ is the total event rate. $\gamma_1(s, n) = \binom{n}{2} X(s)^{-1} + n\theta/2$, is the event rate in the selected subsample, $\gamma_2(s, m) = \binom{m}{2}\big(1 - X(s)\big)^{-1} + m\theta/2$ is the event rate for the nonselected subsample. An event time is generated by decomposing the distribution into the time to the next event in each subsample, then taking the minimum of these times. The density of the time s back to the next event in a subsample is

$$g_i(s \mid u, n) = \gamma_i(s, n) \exp\left(- \int_u^s \gamma_i(v, n) dv \right), \; u < s < \infty, \; i = 1, \;\; 2. \tag{4.7}$$

To generate a time to the next event in either subsample (4.7) is further decomposed. There are two simulated times for each subsample. For example, in the selected subsample times to the next mutation or coalescence have respective rates at a time v of

$$\frac{n\theta}{2}, \quad \binom{n}{2} \frac{1}{X(v)}. \tag{4.8}$$

The time to the next mutation is simple to generate. For the time of the next coalescence the appropriate cumulative distribution function is partially inverted and then solved numerically. The integral $\int_0^v X(t)^{-1} dt$ is stored at jump positions v in the approximating Moran model as a lookup table. A uniform random number U in $[0, 1]$ is generated and the value of s found by solving

$$\frac{-\ln(1 - U)}{\binom{n}{2}} = \int_0^s \frac{1}{X(v)} dv - \int_0^u \frac{1}{X(v)} dv, \tag{4.9}$$

by a bisection lookup routine, then interpolating between jumps of the Moran model. The assumption of a linear path between jumps allows the integral to correctly tend to infinity as the frequency tends to zero, forcing all coalescences in the selected subsample to happen before the removal of the selected mutation from the population. When

an event time t has been generated, an event is chosen to happen with probability proportional to its relative rate at the time t. The probabilities that the event is a mutation or a coalescence in the selected subsample are respectively

$$\frac{\frac{n\theta}{2}}{\gamma(t,n,m)}, \qquad \frac{\binom{n}{2}\frac{1}{X(t)}}{\gamma(t,n,m)}, \tag{4.10}$$

while the probabilities of a mutation or a coalescence in the nonselected subsample are

$$\frac{\frac{m\theta}{2}}{\gamma(t,n,m)}, \qquad \frac{\binom{m}{2}\frac{1}{1-X(t)}}{\gamma(t,n,m)}. \tag{4.11}$$

If a mutation occurs, one of the lineages in the subsample is chosen uniformly at random to have a mutation placed on it. If a coalescence occurs, two lineages are chosen uniformly at random in the subsample to coalesce, decreasing the number of lineages by one.

The selected subsample is guaranteed to find a common ancestor before τ, the time of the selected mutation, as the coalescence rate approaches infinity since $X(\tau) = 0$ and $\int_{\tau-u}^{\tau} X(t)^{-1}dt = \infty$, $u \geq 0$. When the trajectory of the selected mutation finishes at time τ, the selected mutation is placed on the common ancestor lineage of the selected subsample. Once the selected mutation has been added to the genealogy, the one remaining lineage of the selected subsample is added to the nonselected sample. The coalescent process returns to a neutral process for the remainder of the time back to the common ancestor.

4.3 Extensions

The scheme described here is for an infinitely-many-sites model but has been extended to a finitely-many-sites model, for sites other than the selected site (Spencer and Coop, 2004). Using the derivations, due to Kaplan et al. (1988) and Hudson and Kaplan

(1988) see Section 2.5.2, repeat mutation at the selected site and recombination may be included in the model described above. Considering the two variable sized deme analogy to selection when the trajectory is known, repeat mutation and recombination act as migration between the two demes allowing ancestral lineages to move between backgrounds. Let μ_{12} and μ_{21} be the population scaled mutations rates from the non-selected to the selected background and from the selected to nonselected background respectively. In the case of repeat mutation a lineage, at time t, moves from the selected to nonselected background with probability $\mu_{21}(1-X(t))/X(t)dt$ and from the nonselected background to the selected background with probability $\mu_{12}X(t)/(1-X(t))dt$. In the infinitely-many-sites limit, $\mu_{12}, \mu_{21} \to 0$, the only mutation between the 'demes' in the sample history occurs as $X(t) \to 0$ thus the selected mutation is placed upon the last remaining selected lineage at time τ. For a recombining sequence a lineage, at time t, a neutral allele at a locus L moves from selected to nonselected background with probability $\rho(1-X(t))/2$ and in the reverse direction with probability $\rho X(t)/2$, where ρ is the population scaled recombination rate between the locus L and the selected locus. The time to the next event in a sample of n selected and m nonselected lineages when recombination and repeat mutation are included is an inhomogeneous exponential with a rate at time t of

$$\gamma(t,n,m) + n\mu_{12}\frac{1-X(t)}{X(t)} + m\mu_{21}\frac{X(t)}{1-X(t)} + \rho(n+m)/2 \qquad (4.12)$$

and the probability of the various events is given by a relative rate argument as before. A method to simulate a recombining sequence with selection using the method described above has been implemented by a colleague and myself (Spencer and Coop, 2004).

Chapter 5

Importance Sampling, Selection and the Coalescent

5.1 Monte Carlo Methods

Let $h(x)$ be a function of X under some model. We wish to evaluate the expectation of $h(x)$ over a distribution $\pi(x)$

$$\mathbb{E}_\pi[h(X)] = \int h(x)\pi(x)dx. \tag{5.1}$$

In many cases the integration can not be performed analytically. The naive Monte Carlo strategy would be to sample from $\pi(x)$ and take the expectation over the sample. However $h(x)$ may be zero, or near zero, for much of the support of $\pi(x)$, thus this naive method will be highly inefficient. A large body of methods collectively known as Monte Carlo methods have been developed to deal with evaluating such integrals (see for example Liu (2001)). These methods have been successfully applied to population genetic data (see Stephens (2001) for a review).

Many approaches use Monte Carlo Markov Chains (MCMC's) to explore the posterior (Felsenstein et al., 1999; Wilson et al., 2003; Drummond et al., 2002) and these methods have been extended to a number of scenarios. MCMC methods applied to population genetic data often do not make use of the infinitely-many-sites mutation model, a site is allowed to mutate more than once. This allows the methods to be applied to species where the infinitely-many-sites mutation model is clearly not applicable, such as viruses. This crucially is also where they suffer, the infinitely-many sites assumption allows a vast reduction in tree space and thus in situations where the infinitely-many-sites assumption is reasonable often they take far longer for little noticeable difference in results.

5.2 Importance Sampling

Importance sampling and sequential importance sampling will first be explained briefly in a general setting. Given the ineffectiveness of naive simulation, sample points should be concentrated to the areas of space that contain much of the mass of the distribution $h(x)\pi(x)$. This can be achieved by sampling from a proposal or trial distribution, $q(x)$, that is a reasonable approximation to $h(x)\pi(x)$ (see Liu (2001) for a general introduction). The trial distribution must be non-zero for all of the support of $h(x)\pi(x)$. In importance sampling the reweighted function $h(x)\pi(x)/q(x)$ is integrated over the distribution $q(x)$

$$\mathbb{E}_\pi[h(X)] = \int h(x)\frac{\pi(x)}{q(x)}q(x)dx.$$

An approximation to this is

$$\mathbb{E}_\pi[h(X)] \approx \frac{1}{M}\sum_{i=1}^{M} h(X_i)\frac{\pi(X_i)}{q(X_i)},$$

where $X_i, \cdots, X_M$ are independent samples from $q(x)$. $\pi(x)/q(x)$ is termed the importance weight. If $\pi(\mathbf{x})$ is a complicated multidimensional distribution on $\mathbf{x} = (x_1, \cdots, x_d)$ then the distribution and the proposal distribution may be decomposed into a product of conditional distributions

$$\begin{aligned} \pi(\mathbf{x}) &= \pi(x_1)\pi(x_2 \mid x_1) \cdots \pi(x_d \mid x_1, \cdots, x_{d-1}), \\ q(\mathbf{x}) &= q(x_1)q(x_2 \mid x_1) \cdots q(x_d \mid x_1, \cdots, x_{d-1}), \end{aligned}$$

each x_i, $i = 1, \cdots, d$, may then be sampled sequentially from $q(x_i \mid x_1, \cdots, x_{i-1})$. The overall sequential importance weight is

$$\frac{\pi(\mathbf{x})}{q(\mathbf{x})} = \left(\frac{\pi(x_1)}{q(x_1)}\right)\left(\frac{\pi(x_2 \mid x_1)}{q(x_2 \mid x_1)}\right) \cdots \left(\frac{\pi(x_d \mid x_1, \cdots x_{d-1})}{q(x_d \mid x_1 \cdots x_{d-1})}\right). \tag{5.2}$$

This sequential importance sampling approach (Liu, 2001) is particularly helpful when $\pi(\mathbf{x})$ can be written as a recursive Markov system, as the coalescent can be, as this simplifies the sampling greatly.

5.2.1 The Likelihood of a Gene Tree

The likelihood function $L(\beta) = P(\mathcal{D} \mid \beta)$ is the probability of the data $\mathcal{D}$, represented as a gene tree (under the assumption of infinitely-many-sites), treated as a function of β. In this section we assume for simplicity that θ is known and that the emphasis is on estimating selection. In reality, profile likelihood surfaces for β for fixed values of θ may need to be considered.

The likelihood can be expressed as

$$L(\beta) = \int P(\mathcal{D} \mid \mathcal{H}) P_\beta(\mathcal{H}) d\mathcal{H}, \tag{5.3}$$

where $\mathcal{H}$ is the full coalescent history including information about the time of mutations and coalescences. $\mathcal{H}$ can be regarded as missing data in the representation (5.3).

In the infinitely-many-sites model $P(\mathcal{D} \mid \mathcal{H})$ is merely an indicator function of whether $\mathcal{H}$ is compatible with the gene tree. The distribution of a coalescent history $P_\beta(\mathcal{H})$ for a given β value is described in Chapter 4.

A naive Monte Carlo approximation to this would be to simulate M trees as described in Chapter 4, then count how many match the data to find a likelihood estimate

$$L(\beta) \approx \frac{1}{M} \sum_{i=1}^{M} P(\mathcal{D} \mid \mathcal{H}^{(i)}), \tag{5.4}$$

where $\mathcal{H}^{(1)}, \cdots, \mathcal{H}^{(M)}$ are independent samples from $P_\beta(\mathcal{H})$. However, the space of possible genealogies is vast and to evaluate the likelihood of a particular parameter by simply sampling from tree space and see how many are compatible with the data is not feasible. Most if not all of the histories generated would be incompatible with the data, thus this method would be highly inefficient. A computationally efficient method is to focus on genealogies that are compatible with the data (see Stephens (2001) for a general introduction to computational inference and coalescent genealogies). This approach is followed in importance sampling. The likelihood can be expressed as

$$L(\beta) = \int \frac{P_\beta(\mathcal{H})}{Q_\beta(\mathcal{H})} Q_\beta(\mathcal{H}) d\mathcal{H}, \tag{5.5}$$

where $Q_\beta(\mathcal{H})$ is a proposal distribution that has weight only on genealogies where $P(\mathcal{D} \mid \mathcal{H}) = 1$. Thus samples from $Q_\beta(\mathcal{H})$ are always compatible with the data. The ratio $P_\beta(\mathcal{H})/Q_\beta(\mathcal{H})$ is the importance weight. An approximation to (5.5) is

$$L(\beta) \approx \frac{1}{M} \sum_{i=1}^{M} \frac{P_\beta(\mathcal{H}^{(i)})}{Q_\beta(\mathcal{H}^{(i)})}, \tag{5.6}$$

where $\mathcal{H}^{(1)}, \cdots, \mathcal{H}^{(M)}$ are independent samples from $Q_\beta(\mathcal{H})$. $Q_\beta(\mathcal{H})$ is chosen as the distribution of a reverse time Markov process on histories with an initial state the sample data. Sample paths under $Q_\beta(\mathcal{H})$ are thus always compatible with the observed data configuration. If the embedded Markov chain of history changes is $\{\mathcal{H}_j; j =$

$0, 1, \ldots, m\big\}$, with m the number of events in the history, then the importance weight can be expressed as

$$\frac{P_\beta(\mathcal{H})}{Q_\beta(\mathcal{H})} = \left[\prod_{j=1}^{m} \frac{P_\beta(\mathcal{H}_{j-1} \mid \mathcal{H}_j)}{Q_\beta(\mathcal{H}_j \mid \mathcal{H}_{j-1})}\right] \cdot P(H_m), \tag{5.7}$$

where $\mathcal{H}_0 = \mathcal{D}$, the initial data. $P_\beta(\mathcal{H}_{j-1} \mid \mathcal{H}_j)$ and $Q_\beta(\mathcal{H}_j \mid \mathcal{H}_{j-1})$ are transition probabilities of history state changes. $P_\beta(\cdot \mid \cdot)$ is the distribution of a Markov chain forward in time. While the coalescent is usually considered backwards in time, the transition probabilities of the history state chain are only known forward in time. $Q_\beta(\mathcal{H}_j \mid \mathcal{H}_{j-1})$ are transition probabilities in reverse time, taking the form of approximately time-reversed distributions $P_\beta(\mathcal{H}_{j-1} \mid \mathcal{H}_j)$. See Stephens and Donnelly (2000) and Section 2 and 3 of De Iorio and Griffiths (2004a) for discussion. The ratios in (5.7) are sequential importance sampling weights.

We can regard the stochastic process generating $\mathcal{D}$ as consisting of two levels of missing data:

(a) the trajectory of the selected mutation denoted by $\big\{X(t),\ t \geq 0\big\}$; and

(b) the coalescent history $\mathcal{H}$ giving rise to the data.

That is, from (a),

$$\begin{aligned} P(\mathcal{D} \mid \beta) &= \mathbb{E}\Big(P\big(\mathcal{D} \mid \big\{X(t),\ t \geq 0\big\}\big)\Big) \\ &\approx \frac{1}{R}\sum_{i=0}^{R} P(\mathcal{D} \mid \{X(t), t \geq 0\}^{(i)}), \end{aligned} \tag{5.8}$$

where $\Big\{\big\{X(t), t \geq 0\big\}^{(i)}, i = 1, \ldots, R\Big\}$ are independent copies of the selected allele's trajectory, generated by the technique in Section 3.3 with a given value of β. The distribution of $\mathcal{D}$ given $\big\{X(t), t \geq 0\big\}$ does not depend on β; all the information about β is contained in the trajectory $\big\{X(t), t \geq 0\big\}$. In a further decomposition from (b) into

the coalescent history

$$\begin{aligned} P(\mathcal{D} \mid \beta) &= \mathbb{E}\Big[P\big(\mathcal{D} \mid \mathcal{H}\big) P\big(\mathcal{H} \mid \big\{X(t), t \geq 0\big\}\big)\Big] \\ &\approx \frac{1}{RM} \sum_{i=1}^{R} \sum_{j=1}^{M} P(\mathcal{D} \mid \mathcal{H}^{(i,j)}) P\big(\mathcal{H}^{(i,j)} \mid \{X(t), t \geq 0\}^{(i)}\big), \end{aligned} \tag{5.9}$$

for R independent trajectory copies, and M independent history copies for each trajectory. The probability $P\big(\mathcal{D} \mid \{X(t), t \geq 0\}\big)$ is calculated by sequential importance sampling on the coalescent history $\mathcal{H}$ given $\{X(t), t \geq 0\}$.

5.2.2 History States of a Neutral Gene Tree

We will first discuss the history states for a neutral gene tree, the only level of missing data is the history $\mathcal{H}$. The notation and recursion used here to describe the history states was developed and utilised by Griffiths and Tavaré in their series of papers on computational inference and gene trees. This description given here is to aid the understanding of the recursion for a gene tree when the trajectory is known. In the evolution of a gene tree a history state $\mathcal{H}_j$ is described by $(\mathcal{T}, \mathbf{n})$, where $\mathcal{T}$ is the ancestor's gene tree topology back in time from when $\mathcal{H}_j$ is observed, and $\mathbf{n} = (n_1, \ldots, n_k)$ are multiplicities of k types present in the history state $\mathcal{H}_j$. The topology $\mathcal{T}$ is described by mutation paths from the leaves of the tree at the current time, to the ancestral root of the tree, labelled 0. Figure 5.1a illustrates a gene tree configuration. In Figure 5.1a the path to the root for the first type on the left of the tree is (4,2,0), and $\mathbf{n} = (1, 2, 1, 9, 3, 2, 2)$.

The time between events is distributed exponentially as before, see (2.29). As noted in Section 2.3 the order of events in a genealogy in a time homogeneous model, i.e. the standard neutral coalescent, is sufficient to describe the mutational pattern seen in the sample. Thus the probability of a gene tree, in the standard neutral setting, is dependent

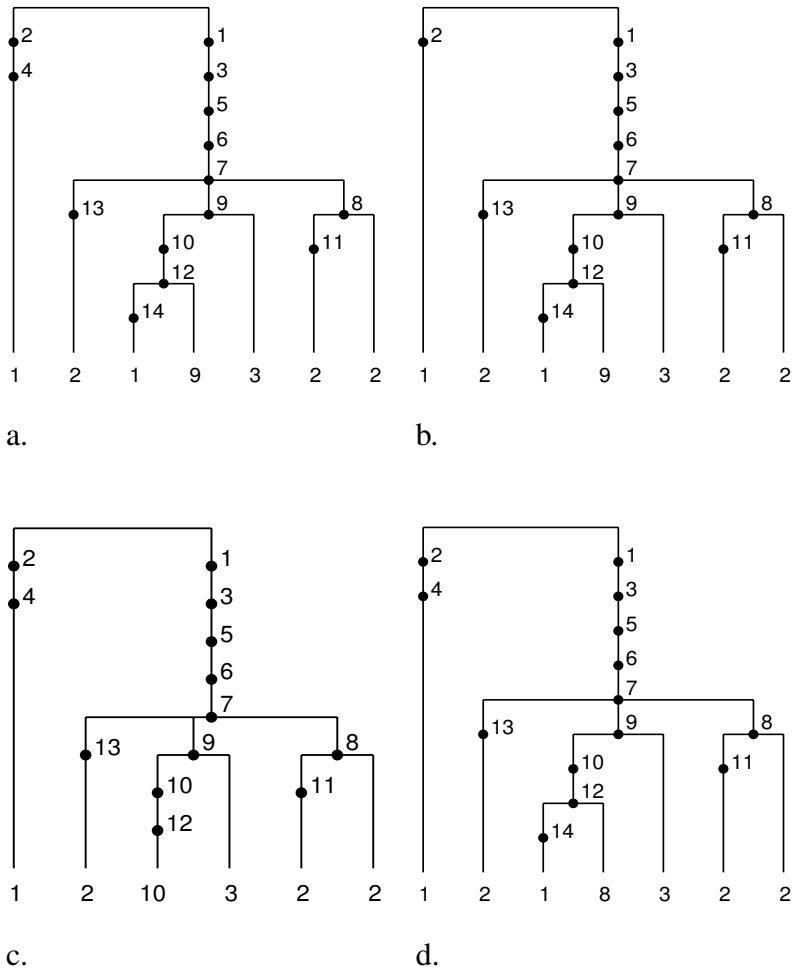

Figure 5.1: A set of four trees showing the original gene tree and three permissable moves and their notation. The dots on the tree are mutations and the numbers at the bottom of the tree denote the multiplicities of the types. a. The original gene tree $(\mathcal{T}, \mathbf{n})$, b. the removal of a mutation (4) that leaves haplotype i distinct $(\mathcal{T}'_{i-}, \mathbf{n})$, c. the removal of a mutation (14) that leaves the haplotype i identical to haplotype j $(\mathcal{T}''_{i-,j+}, \mathbf{n} + \mathbf{e}_j - \mathbf{e}_i)$, d. a coalescence occurs in the 4^{th} type from the left $(\mathcal{T}, \mathbf{n} - \mathbf{e}_i)$.

on only the order of events. The probability of embedded history state jump chain, i.e. the discrete Markov chain of history events embedded within the history states with time information, is a recursion summing over the next event in the history chain, see for example Griffiths (2002). Let $P\left[(\mathcal{T}, \mathbf{n}) \mid (\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}})\right]$ be the transition probability matrix of a history change forward in time in the tree from a state $(\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}})$, given an

event. $\mathcal{A}$ is the operator of the transition probability matrix of the history state chain, then the probability of the current history state, $p(\mathcal{T}, \mathbf{n})$, may be written as

$$\mathcal{A}p(\mathcal{T}, \mathbf{n}) = \sum_{(\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}})} p\big[(\mathcal{T}, \mathbf{n}) \mid (\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}})\big] p(\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}}), \tag{5.10}$$

considering the possible changes in the history this is

$$\begin{aligned} \mathcal{A}p(\mathcal{T}, \mathbf{n}) \;=\; & \left(\binom{n}{2} + \frac{n\theta}{2}\right)^{-1} \Big\{ \\ & \frac{n\theta}{2} \sum_i \frac{1}{n} p(\mathcal{T}'_{i-}, \mathbf{n}) \\ & + \frac{n\theta}{2} \sum_{i \to j} \frac{(n_j + 1)}{n} p(\mathcal{T}''_{i-,j+}, \mathbf{n} + \mathbf{e}_j - \mathbf{e}_i) \\ & + \binom{n}{2} \sum_{\{i : n_i \geq 2\}} \frac{(n_i - 1)}{n - 1} p(\mathcal{T}, \mathbf{n} - \mathbf{e}_i) \Big\}, \end{aligned} \tag{5.11}$$

where $\mathbf{e}_i$, the i^{th} unit vector, represents a multiplicity of 1 for an allele of type i. The probability of a mutation or a coalescence given that an event occurred is found by a relative rate argument, see Section 2.3. Examples of possible state changes of a tree back in time from a history state Figure 5.1a are shown in Figure 5.1b to d. The first and second terms on the right of (5.11) relate to removal of a mutation. There are two cases to consider when a mutation is removed. In the first case removal of the mutation on the singleton lineage of the i^{th} haplotype leaves the lineage distinct. The resulting gene tree topology is denoted by $\mathcal{T}'_{i-}$. The probability that a mutation occurred on this particular lineage forward in time, given that a mutation did occur, is $1/n$. Removal of mutation 4 in the example data set leaves the lineage distinct and results in the gene tree shown in Figure 5.1b. In the second case, which is more complex, removal of the mutation on the singleton lineage of the i^{th} haplotype results in a lineage that is non-distinct from the lineages of haplotype j. The resulting gene tree topology is denoted by $\mathcal{T}''_{i-,j+}$. The probability of this configuration change, conditional on a mutation occurring, is the probability that prior to mutation the j^{th} haplotype had multiplicity $n_j + 1$ and was chosen to mutate. In the example data set, removal of mutation 14 leaves a lineage

identical to the lineages subtended by mutation 12, illustrated in Figure 5.1c. The third term of (5.11) relate to a coalescence. The probability of coalescence occurring within the group of haplotype i, conditional on coalescence occurring, is the probability that prior to coalescence the the i^{th} haplotype had multiplicity $n_i - 1$, and that one of these lineages from a possible $n-1$ was a parent in the coalescent event. This is illustrated in Figure 5.1d. Under the assumption of infinitely-many-sites the ancestor of the sample is known, so the boundary condition is

$$p(\mathcal{T} = (0), \mathbf{n} = \mathbf{e}_i) = 1. \tag{5.12}$$

where the tree $((0), n = \mathbf{e}_i)$ is a singleton root.

5.2.3 History States of a Gene Tree Conditional on a Trajectory

Here we present the history states and recursion for the Markov chain on history states when the trajectory is known. They are similar in form to (5.11) but the recursion has now a time dependent component and so the time events in the genealogy occur must be integrated over. A history state $\mathcal{H}_j$ is described by $(\mathcal{T}, \mathbf{n}, \mathbf{m})$, where as before $\mathcal{T}$ is the ancestor's gene tree topology back in time from when $\mathcal{H}_j$ is observed, and $\mathbf{n} = (n_1, \ldots, n_k)$, $\mathbf{m} = (m_1, \ldots, m_l)$ are multiplicities of k types with the selected allele and l types without the selected allele. Mutation 12 is the selected mutation in Figure 5.1a, and so $n = (1, 9)$ and $m = (1, 2, 3, 2, 2)$. Mutations on an edge are exchangeable, for example mutation 12 could be above or below mutation 10. While the full method accounts for the exchangeability by summing over the possible positions of the selected mutation on the edge (see Section 6.3 for an application of this), for the sake of clarity we assume throughout this Section that the lowest mutation on the edge is the selected mutation.

Let $\big(\mathcal{T}(t), \mathbf{n}(t), \mathbf{m}(t)\big)$ denote the ancestral gene tree at time t back from the current time and

$$p(\mathcal{T}, \mathbf{n}, \mathbf{m}, t) = P\big[(\mathcal{T}(t), \mathbf{n}(t), \mathbf{m}(t)) = (\mathcal{T}, \mathbf{n}, \mathbf{m}, t)\big].$$

The generator $\mathcal{G}$ for the Markov process $\big\{(\mathcal{T}(t), \mathbf{n}(t), \mathbf{m}(t)),\ t \geq 0\big\}$ when the trajectory, $\{X(t), t \geq 0\}$, is known satisfies

$$\frac{d}{dt} p(\mathcal{T}, \mathbf{n}, \mathbf{m}, t) = \mathcal{G} p(\mathcal{T}, \mathbf{n}, \mathbf{m}, t). \tag{5.13}$$

The right hand side of (5.13) can be expressed as

$$\mathcal{G} p(\mathcal{T}, \mathbf{n}, \mathbf{m}, t) = \gamma(t, n, m)\Big(\mathcal{A} - I\Big) p(\mathcal{T}, \mathbf{n}, \mathbf{m}, t), \tag{5.14}$$

noting that the total rate of change in the genealogical history at time t is $\gamma(t, n, m)$, and I denotes the identity operator. Let $P\big[(\mathcal{T}, \mathbf{n}, \mathbf{m}) \mid (\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}}, \widetilde{\mathbf{m}}), t\big]$ be the transition probability matrix of a history change forward in time in the tree from a state $(\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}}, \widetilde{\mathbf{m}})$, given an event at t. Then

$$\mathcal{A} p(\mathcal{T}, \mathbf{n}, \mathbf{m}, t) = \sum_{(\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}}, \widetilde{\mathbf{m}})} p\big[(\mathcal{T}, \mathbf{n}, \mathbf{m}) \mid (\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}}, \widetilde{\mathbf{m}}), t\big] p(\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}}, \widetilde{\mathbf{m}}, t + 0). \tag{5.15}$$

Considering u as the time to the next event back in time from a current time t, then

$$p(\mathcal{T}, \mathbf{n}, \mathbf{m}, u) = \int_0^\infty \mathcal{A} p(\mathcal{T}, \mathbf{n}, \mathbf{m}, t) g_{12}(t \mid n, m, u) dt. \tag{5.16}$$

An integral recursion follows from (5.16), similar to equation (14) in Griffiths and Tavaré (1994b). Summing over possible one step history changes at time t which lead

to a current configuration of $(\mathcal{T}, \mathbf{n}, \mathbf{m})$, as in (5.15), gives

$$\begin{aligned}\mathcal{A}p(\mathcal{T}, \mathbf{n}, \mathbf{m}, t) &= \gamma(t, n, m)^{-1}\Big\{ \\ &\frac{\{n,m\}\theta}{2}\sum_i \frac{1}{\{n,m\}} p(\mathcal{T}'_{i-}, \mathbf{n}, \mathbf{m}, t) \\ &+\frac{\{n,m\}\theta}{2}\sum_{i\to j} \frac{(\{n_j, m_j\}+1)}{\{n,m\}} p(\mathcal{T}''_{i-,j+}, \{\mathbf{n},\mathbf{m}\}+\mathbf{e}_j-\mathbf{e}_i, t) \\ &+\binom{n}{2}\frac{1}{X(t)}\sum_{\{i:n_i\geq 2\}} \frac{(n_i-1)}{n-1} p(\mathcal{T}, \mathbf{n}-\mathbf{e}_i, \mathbf{m}, t) \\ &+\binom{m}{2}\frac{1}{1-X(t)}\sum_{\{i:m_i\geq 2\}} \frac{(m_i-1)}{m-1} p(\mathcal{T}, \mathbf{n}, \mathbf{m}-\mathbf{e}_i, t)\Big\}\end{aligned} \tag{5.17}$$

An abbreviated notation $\{n_i, m_i\}$ (and similar other notation) is used to denote either n_i or m_i depending on whether haplotype i belongs to the selected or neutral class of sequences. The probability of a mutation or a coalescence given that an event occurred is found by a relative rate argument as before, see Chapter 4. The boundary condition for the tree $((0), \{\mathbf{n}, \mathbf{m}\} = \mathbf{e}_i)$ is a singleton root, and the probability of this is one as before.

5.3 Importance Sampling Conditional on a Trajectory

The procedure to generate samples from the proposal distribution and the associated importance weights are detailed in this section. It follows a similar scheme to that described in Section 4, with sequential importance sampling weights depending on the time of history changes. The first step is generating a trajectory for the selected allele back in time, conditional on observing n sequences of the selected type, and m neutral sequences. The initial frequency $X(0) = x$ is chosen from the posterior density $f(x \mid n, m, \beta)$, however an importance weight is necessarily generated from this choice of $X(0) = x$. Calculation of the weight is detailed in Section 5.3.3. A Moran model

sample path approximation to a diffusion trajectory $\{X(t), t \geq 0\}$, conditional on absorption at 0, is simulated. This supplies the trajectory as missing data for the gene tree evolution back in time. The distribution of the trajectory is a Markov process with generator $\mathcal{L}$ described in Section 3.1.

A subdivided population coalescent history $\mathcal{H}$ describing a gene tree sample path $\big\{\big(\mathcal{T}(t), \mathbf{n}(t), \mathbf{m}(t)\big),\ t \geq 0\big\}$ is simulated back in time conditional on the data $\mathcal{D}$, and the trajectory $\{X(t), t \geq 0\}$. The joint density of the time and the subpopulation is detailed in column one of Table 5.1. Informally, from a given state with n, m sequences potential times to the next events in each subpopulation are generated, then the next event is chosen to occur in the subpopulation with the minimum time generated. This generates a time t from $g_{12}(t \mid n, m, u)$ as in Chapter 4. By simulating a time from $g_{12}(t \mid n, m, u)$ we take a sample from the distribution being integrated over in equation (5.16) and thus the proposal on a gene tree sample path is correct in respect to the times. An event is chosen to occur in the subpopulation with the minimum generated time to the next event. This is equivalent to choosing a subpopulation 1 or 2 with probability proportional to rate $\gamma_1(t, n)$ or $\gamma_2(t, n)$ respectively at time t.

As before the time of the removal of the selected mutation from the tree is fixed at τ as, under the infinitely-many site assumption, the mutation must have occurred at the time in the history of the sample that it occurred in the population history. A special case occurs when at a time $t < \tau$ there are no possible events in the proposal distribution for either one or both subsamples, until removal of the selected mutation at time τ. This can occur when the most recent common ancestor of the selected subsample has been reached or no further coalescence or mutation events can occur in the nonselected subsample compatible with the gene tree topology. The distribution of the time of the event is then no longer correctly generated by $g_{12}(t \mid u, n, m)$, and a correction importance sampling weight, detailed in Section 5.3.2, is required for this case. The

next event after removal of the selected mutation is generated starting from time τ.

5.3.1 Importance Sampling Within a Subsample

The infinitely-many-sites proposal distribution of Stephens and Donnelly (2000) is used within the chosen subsample to choose an event. The proposal distribution of Stephens and Donnelly (2000) is uniform on the choice of available lineages. An available lineage is one where there is more than one of that type, or a singleton where the removal of the mutation would not violate the gene tree topology. Denote the number of available lineages in the selected and neutral subpopulations by r_1 and r_2. Under the infinitely-many-sites assumption once a lineage has been chosen only one event is possible: if there are others of the same type a coalescence must take place; or if the lineage is a singleton then the mutation is removed. The importance weights associated with sampling are shown in Table 5.1.

After the removal of the selected mutation, m is set to $m+1$, and $n = 0$ as the previously selected lineage becomes neutral. The neutral gene tree ancestral process then behaves as a constant-sized population process, with the coalescence rate while l lineages being $\binom{l}{2}$, and the infinitely-many-sites importance sampling algorithm of Stephens and Donnelly (2000) is used to complete the sample path simulation.

5.3.2 Removal of the Selected Mutation

There may be no possible events in the proposal distribution before the selected mutation is removed at time τ if either $r_1 = 0$ or $r_2 = 0$. Event times can no longer be generated from the distribution $g_{12}(t \mid n, m, u)$ as there are no proposed events in one or both of the subsamples. If $r_i = 0$ then we generate a time from $g_j(t \mid n, u)$ $i \neq j$. When

Table 5.1: The importance sampling when both subtrees still have events to perform (i.e. $r_1 \neq 0$ and $r_2 \neq 0$).

Joint density of t and subsample	$\mathcal{H}_k$	$Q(\mathcal{H}_k \mid \mathcal{H}_{k-1})$	$P(\mathcal{H}_{k-1} \mid \mathcal{H}_k)$	Importance Weight
(A) Events within the selected subsample $\mathbf{n}$				
$\frac{\gamma_1(t,n)}{\gamma_1(t,n)+\gamma_2(t,m)} g_{12}(t \mid n,m,u)$	$\mathcal{T}'_{i-}, \mathbf{n}, \mathbf{m}$	$1/r_1$	$\frac{n\frac{\theta}{2}}{\gamma_1(t,n)+\gamma_2(t,m)} \frac{1}{n} g_{12}(t \mid n,m,u)$	$\frac{r_1\theta}{2\gamma_1(t,n)}$
	$\mathcal{T}''_{i-,j+}, \mathbf{n}-\mathbf{e}_i+\mathbf{e}_j, \mathbf{m}$	$1/r_1$	$\frac{n\frac{\theta}{2}}{\gamma_1(t,n)+\gamma_2(t,m)} \frac{n_j+1}{n} g_{12}(t \mid n,m,u)$	$\frac{r_1(n_j+1)\theta}{2\gamma_1(t,n)}$
	$\mathcal{T}, \mathbf{n}-\mathbf{e}_i, \mathbf{m}$	n_i/r_1	$\frac{\binom{n}{2}\frac{1}{X(t)}}{\gamma_1(t,n)+\gamma_2(t,m)} \frac{n_i-1}{n-1} g_{12}(t \mid n,m,u)$	$\frac{r_1 n(n_i-1)}{2n_i X(t)\gamma_1(t,n)}$
(B) Events within the unselected subsample $\mathbf{m}$				
$\frac{\gamma_2(t,m)}{\gamma_1(t,n)+\gamma_2(t,m)} g_{12}(t \mid n,m,u)$	$\mathcal{T}'_{i-}, \mathbf{n}, \mathbf{m}$	$1/r_2$	$\frac{m\frac{\theta}{2}}{\gamma_1(t,n)+\gamma_2(t,m)} \frac{1}{m} g_{12}(t \mid n,m,u)$	$\frac{r_2\theta}{2\gamma_2(t,m)}$
	$\mathcal{T}''_{i-,j+}, \mathbf{n}, \mathbf{m}-\mathbf{e}_i+\mathbf{e}_j$	$1/r_2$	$\frac{m\frac{\theta}{2}}{\gamma_1(t,n)+\gamma_2(t,m)} \frac{m_j+1}{m} g_{12}(t \mid n,m,u)$	$\frac{r_2(m_j+1)\theta}{2\gamma_2(t,m)}$
	$\mathcal{T}, \mathbf{n}, \mathbf{m}-\mathbf{e}_i$	m_i/r_2	$\frac{\binom{m}{2}\frac{1}{1-X(t)}}{\gamma_1(t,n)+\gamma_2(t,m)} \frac{m_i-1}{m-1} g_{12}(t \mid n,m,u)$	$\frac{r_2 m(m_i-1)}{2m_i(1-X(t))\gamma_2(t,n)}$

Subtable (A) gives details of events within the selected subsample $\mathbf{n}$. Subtable (B) gives details of events in the nonselected subsample $\mathbf{m}$. The first column gives the joint proposal density of time t and the subsample. The second gives the new history state. The proposal probability of the event within the subsample is shown in the third. The forward joint density of the event and the event time t (i.e. the term from (5.17)) are in the fourth. Finally, the fifth column is the importance weight associated with that particular event and time. The time is distributed as $g_{12}(t \mid n,m,u)$ for both the proposal and the true density and thus no term for the time appears in the weight.

an event time is chosen an event is chosen using the Stephens and Donnelly (2000) proposal scheme as before. This sampling of t must be corrected for in the importance weights. For the two cases when either r_1 or r_2 are zero, these are detailed in the first two rows of Table 5.2. If an event in the nonselected subsample is proposed at a time $> \tau$, then the selected mutation must be removed at this time. This case is detailed in the third row. The fourth row has details of the case when there are no proposed events to perform ($r_1 = r_2 = 0$), and the next event is removal of the selected mutation.

For example, if the selected subsample has found its common ancestor at u and the nonselected subsample still has events compatible with the gene tree topology, to be performed, event times will be generated only from $g_2(t \mid m, u)$ and events chosen in that subsample. Every new t generated incurs the importance weight given in the final column of row 1 of Table 5.2. If the the time t is $> \tau$ then the mutation is removed and the the weight given in the fourth row is incurred.

The importance weights described in this section also have a clear probabilistic interpretation. Informally, the weights can viewed as the probability of no events occurring in the subsample history in a subsample that has performed all the events permitted to it.

5.3.3 The Trajectory Starting Frequency

The initial trajectory frequency $X(0)$ is sampled from the density (4.5), conditional on n sample sequences from a sample of $n + m$ sequences containing the selected mutation. However the joint probability distribution of $X(0)$ and a sample with (n, m) such sequences is required. The initial unconditional prior distribution of the frequency of the selected mutation can be taken as the frequency spectrum of the mutation in the

Table 5.2: The importance sampling scheme when one or both subtrees have no more events to perform.

No. Events	$\mathcal{H}_k$,time	time density	$P(\mathcal{H}_{k-1} \mid \mathcal{H}_k)$	Importance Weight
$r_1 \neq 0, r_2 = 0$	$(\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}}, \mathbf{m}), t$	$g_1(t \mid u, n)$	$P(\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}}, \mathbf{m} \mid \mathbf{n}, \mathbf{m}, t)$	$W(\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}}, \mathbf{m} \mid \mathbf{n}, \mathbf{m}, t) \exp\left\{-\int_u^t \gamma_2(s, m) ds\right\}$
$r_1 = 0, r_2 \neq 0$	$(\widetilde{\mathcal{T}}, \mathbf{n}, \widetilde{\mathbf{m}}), t$	$g_2(t \mid u, m)$	$P(\widetilde{\mathcal{T}}, \mathbf{n}, \widetilde{\mathbf{m}} \mid \mathbf{n}, \mathbf{m}, t)$	$W(\widetilde{\mathcal{T}}, \mathbf{n}, \widetilde{\mathbf{m}} \mid n, m, t) \exp\left\{-\frac{\theta(t-u)}{2}\right\}$
$r_1 = 0, r_2 \neq 0$	$(\mathcal{T}''_{i-,j+}, \mathbf{n} - \mathbf{e}_i, \mathbf{m} + \mathbf{e}_j), \tau$	$\exp\left\{-\int_u^\tau \gamma_2(s, m) ds\right\}$	$\frac{\theta}{2} \frac{m_j+1}{m+1} g_{12}(\tau \mid u, 1, m)$	$\frac{\theta}{2} \frac{m_j+1}{m+1} \exp\left\{-\frac{\theta(\tau-u)}{2}\right\}$
$r_1 = 0, r_2 = 0$	$(T''_{i-,j+}, \mathbf{n} - \mathbf{e}_i, \mathbf{m} + \mathbf{e}_j), \tau$	1	$\frac{\theta}{2} \frac{m_j+1}{m+1} g_{12}(\tau \mid u, 1, m)$	$\frac{\theta}{2} \frac{m_j+1}{m+1} \exp\left\{-\int_u^\tau \frac{\theta}{2} + \gamma_2(s, m) ds\right\}$

The different ways either or both subtrees can run out of events at a time u are given in the first column. The history one event back in time is shown in the second column ($\widetilde{\mathcal{T}}, \widetilde{\mathbf{n}}, \mathbf{m}$ and $\widetilde{\mathcal{T}}, \mathbf{n}, \widetilde{\mathbf{m}}$ denote the histories given in parts A and B of Table 5.1, respectively). The time chosen for the next event is also given in the second column ($t < \tau$). The proposal density of the time is shown in the third. The joint forward history probability of the event and the time density is given in the fourth ($P(\cdot \mid \cdot)$ and $W(\cdot \mid \cdot)$ refer to the $P(\mathcal{H}_{k-1} \mid \mathcal{H}_k)$ and the importance weights entries in Table 5.1 respectively). Finally the importance weights associated with the different situations are given in the last column.

population,

$$f(x) \propto m(x)u_0(x).$$

This is an improper prior. The correct joint probability of the sample and the density of the population frequency is then proportional to

$$\binom{n+m}{n} x^n (1-x)^m m(x)u_0(x). \tag{5.18}$$

The importance weight when $X(0) = x$ is then (5.18)/$f_{n,m}(x)$, which is equal to

$$\binom{n+m}{n} \int_0^1 x^n (1-x)^m m(x)u_0(x)dx. \tag{5.19}$$

One interpretation is that the selected site is a random choice of site chosen from many selected segregating sites along the sequences. Another is that the selected site is a specific site, and sampling is chosen at a uniform random time between when the mutation arose and before it became lost or fixed in the population. Then, conditioning on the selected site as a specific site segregating in the sample the frequency spectrum of the selected site in the sample is

$$\frac{\int_0^1 \binom{n+m}{n} x^n (1-x)^m m(x)u_0(x)dx}{\int_0^1 \left(1 - x^{n+m} - (1-x)^{n+m}\right) m(x)u_0(x)dx}, \tag{5.20}$$

and the importance weight is related to this sample frequency spectrum. The general sample frequency spectrum is studied in Griffiths (2003). Under neutrality ($\beta = 0$) $m(x) = \frac{1}{x(1-x)}$ and $u_0(x) = 1 - x$, thus for the neutral drift case

$$\begin{aligned} \frac{P(n, n+m \mid x) f(x \mid \beta = 0)}{f(x \mid n, n+m, \sigma = 0)} &= \binom{n+m}{n} \int_0^1 x^{n-1}(1-x)^m dx & (5.21) \\ &= \binom{n+m}{n} \mathcal{B}(n, m+1) & (5.22) \\ &= \frac{1}{n} & (5.23) \end{aligned}$$

The limit as $\beta \to \infty$ is also of interest,

$$\begin{aligned} m(x)u_0(x) &= \frac{e^{\beta x}}{x(1-x)}\left(\frac{e^{-\beta x}-e^{-\beta}}{1-e^{-\beta}}\right) \qquad (5.24) \\ &= \frac{1}{x(1-x)}\left(\frac{1-e^{-\beta(1-x)}}{1-e^{-\beta}}\right) \\ \lim_{\beta\to\infty} m(x)u_0(x) &= \frac{1}{x(1-x)}. \qquad (5.25) \end{aligned}$$

The limit of (5.19) as $\beta \to \infty$ is

$$\frac{n+m}{nm}$$

5.3.4 Computational Features

A number of features have been incorporated to improve the computational performance of the likelihood algorithm. The independence of the history under the selected site and the rest of the history given the trajectory subdividing the two populations, is used to significantly lower the variance of the estimated likelihood. A number of trees are generated conditional on the same trajectory to improve computational efficiency. The usual importance sampling procedure to find the likelihood (5.9) would be to average the importance weight $W(\cdot \mid \cdot)$ of each history

$$L(\beta) \approx \frac{1}{RM}\sum_{i=1}^{R}\sum_{j=1}^{M} W(\mathcal{H}^{(j)} \mid \{X(t), t \geq 0\}^{(i)}), \qquad (5.26)$$

where M trees are sampled per trajectory and there are R trajectories in total. However using the independence of the two subtrees we instead take the sum

$$\begin{aligned} L(\beta) \approx\ & \frac{1}{RM^2}\sum_{i=1}^{R}\left(\sum_{j=1}^{M} W(\mathcal{H}^{(s,j)} \mid \{X(t), t \geq 0\}^{(i)})\right) \\ & \times\left(\sum_{j=1}^{M} W(\mathcal{H}^{(u,j)} \mid \{X(t), t \geq 0\}^{(i)})\right), \qquad (5.27) \end{aligned}$$

where $\mathcal{H}^{(s,j)}$ is the subtree history under the selected mutation not including the removal of the selected mutation, and $\mathcal{H}^{(u,j)}$ is the complementary nonselected subtree history

which includes removal of the selected mutation. The approach in effect permutes all possible combinations of the histories under the selected mutation and the rest of the history. This results in a reduced sampling variance for $P(\mathcal{D} \mid \{X(t), t \geq 0\})$ allowing (5.27) to converge more quickly than (5.26). This method of variance reduction will be referred to as the two subtree variance reduction method throughout this thesis.

Importance Sampling the Trajectory Away from the Driving β Value

The decomposition of the likelihood into the expectation over trajectories and the histories given the trajectory in (5.9) allows importance sampling of the trajectories at a selection value β_j away from the driving value β_d. The associated importance weight is then simply the ratio of the probabilities of the trajectory, under the two selection values β_j, β_d. Recall the approximating Moran model described in the Section 3.3. If there are r jumps taken in the sample path to reach $Z = 0$ the importance weight is

$$W(\{Z(t), t \geq 0\}) = \frac{\prod_{k=1}^{r} P(Z_k \mid Z_{k-1}, \beta_j) f(T_k - T_{k-1} \mid Z_k, \beta_j)}{\prod_{k=1}^{r} P(Z_k \mid Z_{k-1}, \beta_d) f(T_k - T_{k-1} \mid Z_k, \beta_d)}, \tag{5.28}$$

where T_k is the time of the k^{th} jump, $f(T_k - T_{k-1} \mid Z_k, \beta_j)$ is the density of the next event time, in the Moran model, Z_k is the number of selected genes in the time interval $[T_{k-1}, T_k]$ and $P(Z_k \mid Z_{k-1}, \beta_j)$ is the jump probability of the Moran Model, where $Z_k = Z_{k-1} + 1$ or $Z_k = Z_{k-1} - 1$. After simplification

$$\begin{aligned} &\prod_{k=1}^{r} P(Z_k \mid Z_{k-1}, \beta_j) f(T_k - T_{k-1} \mid Z_k, \beta_j) = \\ &\frac{u_0\left(\frac{1}{N} \mid \beta_j\right)}{u_0\left(\frac{Z(0)}{N} \mid \beta_j\right)} \exp\left(-\frac{N^2}{2} \sum_{k=0}^{r} \sigma^2(x_k)\left(t_{k+1} - t_k\right)\right) \\ &\times \prod_{k=0}^{r} \begin{cases} \frac{N\sigma^2(x_k) + \mu_{\beta_j}(x_k)}{4\sigma^2(x_k)} & \text{if } Z_{k+1} = Z_k + 1 \\ \frac{N\sigma^2(x_k) - \mu_{\beta_j}(x_k)}{4\sigma^2(x_k)} & \text{if } Z_{k+1} = Z_k - 1 \end{cases} \end{aligned} \tag{5.29}$$

Thus the weights can be expressed as

$$
\begin{aligned}
W(\{Z(t), t \geq 0\}) &= \frac{u_0\left(\frac{1}{N} \mid \beta_j\right) u_0\left(\frac{Z(0)}{N} \mid \beta_d\right)}{u_0\left(\frac{Z(0)}{N} \mid \beta_j\right) u_0\left(\frac{1}{N} \mid \beta_d\right)} \\
&\quad \times \prod_{k=0}^{r} \left(\frac{N\sigma^2(x_k) + \mu_{\beta_j}(x_k)}{N\sigma^2(x_k) + \mu_{\beta_d}(x_k)}\right)^{\delta(Z_{k-1}+1, Z_k)} \\
&\quad \times \left(\frac{N\sigma^2(x_k) - \mu_{\beta_j}(x_k)}{N\sigma^2(x_k) - \mu_{\beta_d}(x_k)}\right)^{\delta(Z_{k-1}-1, Z_k)} \qquad (5.30)
\end{aligned}
$$

There is little computational effort involved in recording extra terms for the weight. This importance sampling allows for a section of likelihood surface around β_d to be evaluated so the total surface can be produced with fewer driving values.

5.4 Convergence

The convergence of any Monte Carlo scheme is an important consideration. For all applications of the method in this thesis multiple runs of the algorithm were performed to check the convergence of the estimated likelihood. Where sufficient computer power is available this is the most desirable diagnostic of the method. Gene trees with an allele segregating in the sample with selection coefficient $\beta = 0$ have the same distribution as the standard neutral gene trees, studied by Griffiths and Tavaré, only the levels of missing data introduced by the importance sampling schemes differ. The Stephens and Donnelly (2000) infinitely-many-sites importance sampler for the standard neutral model provides an estimate of the likelihood for $\beta = 0$ that converges more quickly than the trajectory based scheme described in this chapter. The trajectory represents an unnecessary level of missing data for $\beta = 0$, which is in effect integrated out in the standard neutral model. Thus an implementation of the infinitely-many-sites importance sampler (kindly provided by R. C. Griffiths) provides a good comparison at $\beta = 0$ for the trajectory based importance sampler, which has allowed a number of

discrepancies in the trajectory based method to be found and corrected. It also allows the convergence of the trajectory based method to be checked as the estimates of the likelihood may be compared at $\beta = 0$. Although technically this offers no guarantee for the convergence of the method for $\beta \neq 0$ informally for positive directional selection the trajectories will be most variable for $\beta = 0$, and thus convergence at $\beta = 0$ is reassuring.

A simple gene tree is shown in Figure 5.2, the selected mutation is mutation 2. The likelihood of this gene tree was calculated using naive simulation, as in (5.4). Gene trees were simulated conditional on the selected allele being present in 5 out of the 11 haplotypes in the sample using an implementation of the simulation method described in Chapter 4. Each simulation was weighted by the probability of observing the selected allele in 5 out of the 11 sample haplotypes conditional on the selected allele segregating in the population, i.e. (5.19) as detailed in Section 5.3.3. A weight of $\theta/2$ was also included for the probability of the selected mutation having occurred at the time of loss of the selected allele from the population on the lineage ancestral to the sample. The log likelihood evaluated for a range of θ and β in a model of genic selection is shown in Figure 5.3. Each of the four log likelihood estimates at each of the parameter values was evaluated by 10 million simulations. The naive simulated estimated likelihoods at a parameter value have a small variance. Each of the log likelihood values was also evaluated by 1 million iterations of the importance sampling algorithm. The naive simulation and the importance sampling algorithm estimates of the log likelihood agree well, giving reassurance that the implementation of the importance sampling algorithm is correct.

An example gene tree is shown in Figure 5.4 (this was simulated conditional on $n = 15$ and $m = 15$ using the simulation method described in Chapter 4). In Figure 5.5 six realisations of the likelihood curve for β for the example gene tree are shown. Each

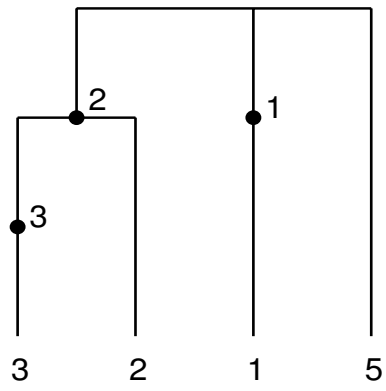

Figure 5.2: A simple gene tree, mutation 2 is the selected mutation.

likelihood point was evaluated using 5 million runs of the algorithm, generating a new trajectory every 500 runs (each point of the surface takes 12.5 minutes on a 2.4 GHz machine). There is good agreement in independent estimates of the likelihood, and so we have reasonable faith in the convergence of the method.

The comparison of the estimated likelihood across multiple runs in encouraging, but for a single long run of the program it is important to have an estimate of the variance of the estimator (and for this variance to be well behaved). Using a notation similar to Section 5.3.4, let R be the number of trajectories sampled and M be the number of histories sampled per trajectory. The variance in the estimated likelihood (at the driving β value) may be written as

$$\frac{1}{R}\mathbb{E}(\mathrm{Var}(W|\{X(t) \mid t \geq 0\})) + \frac{1}{R}\mathrm{Var}(P(\mathcal{D} \mid \{X(t), t \geq 0\})) \tag{5.31}$$

where $\mathrm{Var}(W|\{X(t) \mid t \geq 0\})$ is the variance in the estimate of the likelihood for a trajectory (the W denoting the importance weights) and $\mathrm{Var}(P(\mathcal{D} \mid \{X(t), t \geq 0\}))$ is the variance of the exact likelihood of a trajectory across trajectories. The first term of (5.31) is due to the fact that we can not estimate the likelihood of a trajectory without error. The second term of (5.31) arises because the likelihood of the data depends on the trajectory, and thus a large number of multiple trajectories must be considered. The

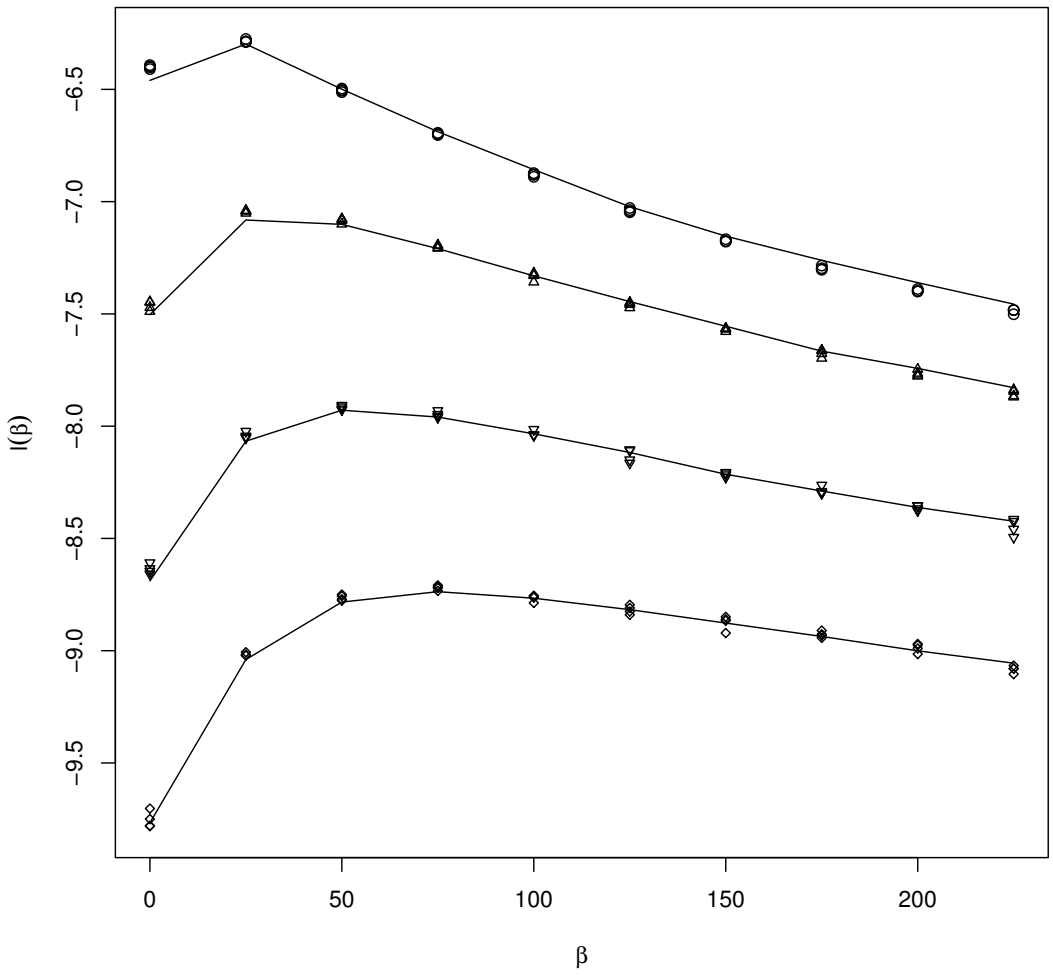

Figure 5.3: A comparison of naive simulation estimates to those of to the importance sampling algorithm. Four log likelihood values, for each value of θ and β, from the naive simulation are shown for $\theta = 1$ as $\bigcirc$, $\theta = 2$ as $\triangle$, $\theta = 3$ as $\bigtriangledown$ and $\theta = 4$ as $\Diamond$. The log likelihood curves evaluated by the importance sampling algorithm for the same parameter values as the naive simulation are shown as solid lines with $\theta = 1, 2, 3, 4$ as respective lines from the top of the figure.

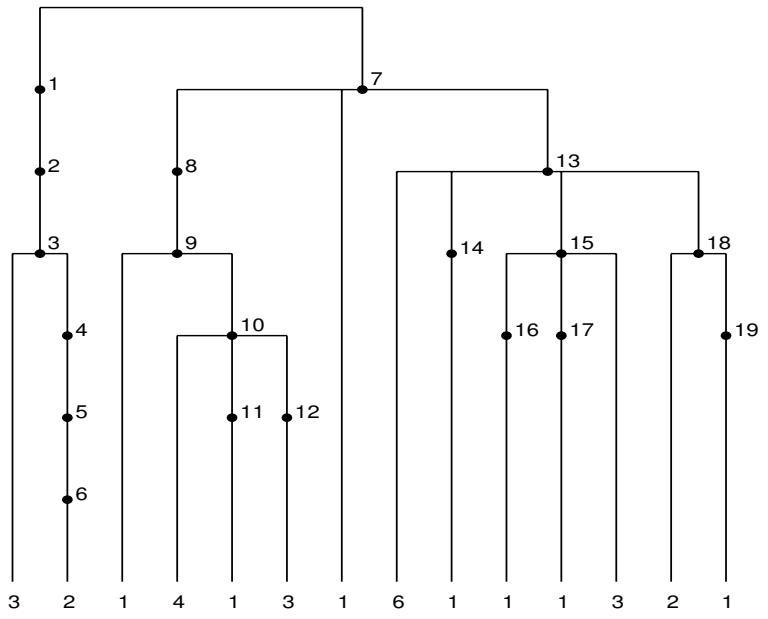

Figure 5.4: An example gene tree, generated with $\beta = 20$ and $\theta = 5$, mutation 13 is the selected mutation.

two subtree variance reduction (TSVR) method described in Section 5.3.4 reduces the variance due to the first term of (5.31). For a given trajectory let $W^{(s)}$ be an importance weight for the subtree history under the selected mutation and $W^{(u)}$ be the importance weight for the complementary nonselected subtree. The variance of the estimated likelihood for a trajectory for the standard averaging (SA) method (5.26) may be written as

$$\frac{1}{M}\mathrm{Var}\left(W^{(u)}\right)\mathrm{Var}\left(W^{(s)}\right) + \frac{1}{M}\mathbb{E}(W^{(s)})^2\mathrm{Var}\left(W^{(u)}\right) + \frac{1}{M}\mathbb{E}(W^{(u)})^2\mathrm{Var}\left(W^{(s)}\right), \tag{5.32}$$

and the variance of the estimate using the TSVR estimator (5.27) as

$$\frac{1}{M^2}\mathrm{Var}\left(W^{(u)}\right)\mathrm{Var}\left(W^{(s)}\right) + \frac{1}{M}\mathbb{E}(W^{(s)})^2\mathrm{Var}\left(W^{(u)}\right) + \frac{1}{M}\mathbb{E}(W^{(u)})^2\mathrm{Var}\left(W^{(s)}\right). \tag{5.33}$$

As can be seen from a comparison of (5.32) to (5.33) the TSVR method reduces the first term of the (5.33) at rate $1/M^2$ compared to the SA method's rate of $1/M$. The remaining two terms of both variances are reduced at the same rate of $1/M$. The variance

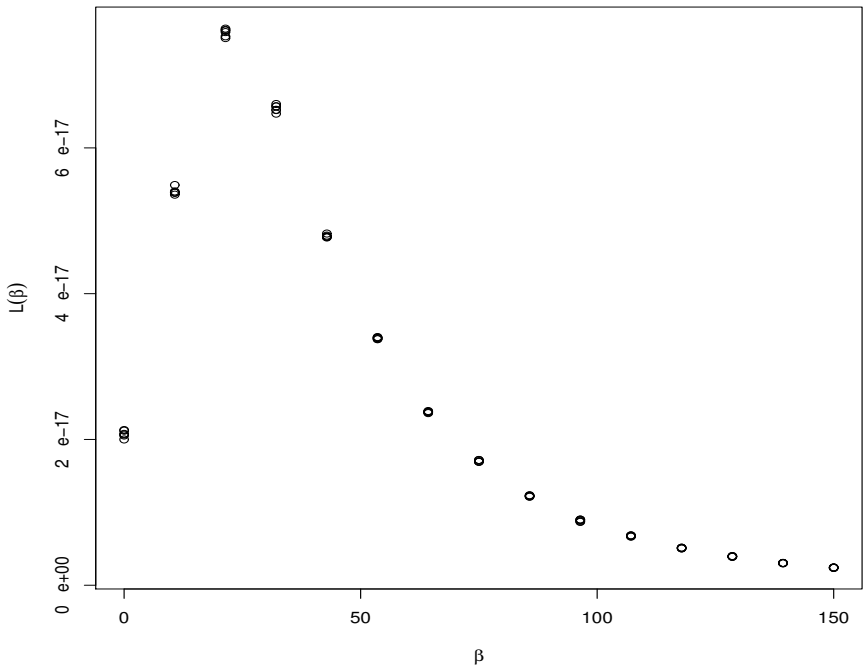

Figure 5.5: Six likelihood surfaces for β for the example gene tree.

within a trajectory, averaged over 1000 trajectories, was calculated using expressions (5.32) and (5.33) for different values of M and β. The results are shown in Table 5.3. (An alternate way of calculating the variance of the SA method's estimate of the likelihood of a trajectory would be to calculate the variance of $W^{(u)}W^{(s)}$ directly. This was found to give similar results to (5.32), results not shown.) The ratio of the average variance calculated by (5.32) and (5.33) is also given in this table to investigate the efficiency of the TSVR method. For large M the ratio of the variances will tend to a constant, as the first term of (5.33) will tends to zero at a higher rate than the other terms. For the two selection coefficients considered in Table 5.3 this happens for 500 runs or more per trajectory. The most efficient number of runs per trajectory, to give the lowest variance for the same amount of computational time, could be found using (5.31) and (5.33). However the variances and expectations in both formulae would need to be known, and these will vary with the data considered and the strength of selection.

Given that we do not know the variance in the exact estimates of the likelihood of the trajectory (i.e. the second term of (5.31)), it is difficult to define an appropriate measure of the variance in the estimator of the overall likelihood $L(\beta)$. We chose to use the variance of the estimated likelihood across trajectories. This differs from the second term of (5.31) in that it is a variance across an estimate of the likelihood of the trajectories not the true likelihoods of trajectories. It is not an ideal estimate of the variance of our estimator. However, it does at least incorporate the variance amongst trajectories and the variance of importance weights for a given trajectory, as it is a variance across estimates not exact likelihoods.

If the importance sampler is performing poorly or insufficient runs have been performed then our measure of the variance will give a poor indication of the true variance. Other indications of the performance of the method are the effective sample size and the ratio of the average weight to the maximum weight (see for example Liu (2001)). However these will also be poor indications of performance if the distribution of the simulated importance weights is not representative of the true distribution of importance weights. If our measure of the variance is a good estimate of the true estimator variance, i.e. if sufficient runs have been performed, it should stay constant as R (the number trajectories) increases. The mean and variance of the estimated likelihood across trajectories for the example gene tree for an increasing number of trajectories are given in Table 5.4. 500 runs were performed on each trajectory and the TSVR method was used. The variance seems to be well behaved with the number of runs and gives a fair representation of the true variation of the importance weights for around 1 million runs upward.

Table 5.3: The average variance of importance weights within a trajectory for different numbers of runs per trajectory for two values of β (averaged over 1000 trajectories). The columns labelled standard are the variances of the SA method, the columns labelled two tree are the variances of the TSVR method. The column labelled ratio is the ratio of the average SA variance to the average TSVR variance.

	$\beta = 0$			$\beta = 100$		
Number per traj.	Standard	Two Tree	Ratio	Standard	Two Tree	Ratio
10	4.32E-32	1.35E-32	3.20	1.97E-32	5.99E-33	3.29
50	3.06E-33	6.44E-34	4.74	1.85E-33	2.40E-34	7.68
100	2.18E-33	3.68E-34	5.92	2.60E-33	1.92E-34	13.5
500	1.33E-33	1.22E-34	11.0	5.64E-34	2.51E-35	22.5
1000	5.11E-34	5.50E-35	9.28	2.87E-34	1.26E-35	22.7
2000	3.12E-34	3.24E-35	9.61	1.56E-34	7.30E-36	21.4

Table 5.4: The mean and variance of the estimate of the likelihood of the trajectory for different numbers of trajectories for two values of β.

	$\beta = 0$		$\beta = 100$	
Number of trajectories	mean	var	mean	var
20	2.40E-17	1.76E-33	8.40E-18	3.33E-35
100	2.48E-17	1.33E-33	7.75E-18	5.75E-35
200	2.36E-17	1.28E-33	8.69E-18	8.45E-35
1000	1.99E-17	1.07E-33	7.66E-18	6.00E-35
2000	2.00E-17	1.00E-33	8.08E-18	7.90E-35
10000	2.10E-17	1.17E-33	8.19E-18	7.19E-35
20000	2.10E-17	1.16E-33	8.01E-18	7.35E-35

Chapter 6

Data Analysis

6.1 G6PD and a Partial Sweep

Alleles at a number of loci in the human genome including the Sickle Cell anaemia, the Duffy blood group and glucose-6-phosphate dehydrogenase (G6PD) genes are known to give resistance to different types of malaria. The evidence from studies of the geographical distribution and sequence diversity around these loci lends support to the theory that natural selection has played a role in the history of these regions.

The G6PD locus on the X chromosome is known to have a role in preventing damage to red blood cells. Low G6PD activity can cause clinical disorders. Despite this apparently important role, alleles that result in deficient enzyme activity occur at high frequencies in many geographic locations. Their presence is often associated with the occurrence of endemic malaria. In particular, the allele $A-$ has risen to an intermediate frequency in sub-Saharan Africa. This allele is at a low frequency in the rest of the world, where the ancestral type (inferred by comparison to chimpanzee sequence), B predominates. Heterozygous females and hemizygous males for the $A-$ allele have been shown by Ruwende et al. (1995) to have an approximately 50% reduced risk of

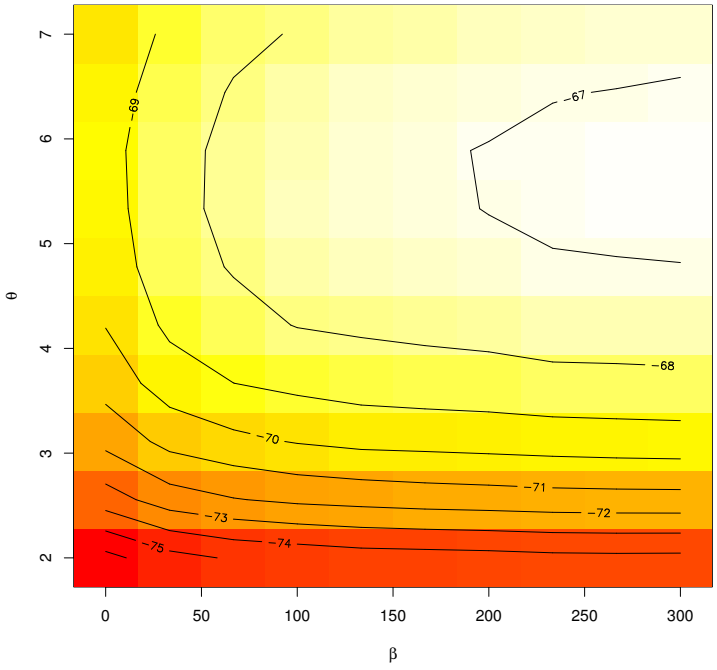

Figure 6.1: Estimated log likelihood surface for the G6PD data set as a function of the selection parameter β and the population scaled mutation rate θ.

infection by malaria (homozygous females for the $A-$ allele are believed to have this reduced risk as well). This reduction is believed to be due to the protection offered to G6PD deficient individuals by their red blood cells offering a more toxic environment to malaria.

Malaria is believed to have become endemic in human populations in the past $10,000$ years. Thus it has been suggested that the $A-$ G6PD allele has arisen recently and is currently sweeping through the sub-Saharan Africa population. (See Verrelli et al. (2002) and references therein for discussion of the points made so far in this section). A number of population genetics studies have focussed on the G6PD region to examine the effect of this putative selective sweep in progress (Tishkoff et al., 2001; Saunders

et al., 2002; Verrelli et al., 2002; Sabeti et al., 2002). Here we focus on the gene tree for the African data presented in Verrelli et al. (2002), reproduced in Figure 6.2. Verrelli et al. (2002) removed a haplotype that appeared twice in the data set to make the data compatible with the infinitely-many-sites and no recombination assumption. The haplotype removed was not from the $A-$ clade nor does it appear to be a recombinant with a sequence from the $A-$ clade. Thus the removal of this haplotype will have little affect on the estimated age of the $A-$ allele. The $A-$ allele is the result of a single amino acid change, mutation 2 in Figure 6.2. For simplicity, we shall ignore any possible negative selection against $A-$ and take the selective advantage of $A-$ to be the same in hemizygous males and both heterozygous and homozygous females. This is the model suggested by Ruwende et al. (1995) of dominant selection in females and genic selection in males. The infinitesimal drift parameters in females and males, $\mu_F(x)$ and $\mu_M(x)$, are respectively

$$\mu_F(x) = \beta x(1-x)^2, \quad \mu_M(x) = \frac{\beta}{2}x(1-x). \tag{6.1}$$

Ignoring any negative selection against $A-$ is reasonable, since, given its probable recent increase in frequency, there will be little effect on the model or information about its effects in the sequence diversity. However, it is worth noting that the general selective scheme possible using this method would allow this to be explored in a more complete analysis. The X chromosome spends on average $2/3$ of its time in the female and $1/3$ in the male, thus a heuristic assumption is made that the total infinitesimal drift coefficient is

$$\mu(x) = \frac{2}{3}\mu_F(x) + \frac{1}{3}\mu_M(x). \tag{6.2}$$

The likelihood of the selection parameter under this selective scheme was evaluated for the Verrelli et al. (2002) gene tree using the importance sampling method described in this thesis. The log likelihood surface for the G6PD data set is shown in Figure 6.1. Each point on the surface was independently evaluated by 5 million iterations of the importance sampling algorithm. Independent runs to evaluate the surface agree

to a high degree, giving confidence in the surface produced (results not shown). The X chromosome has an effective population size of $(3/2)N_e$, compared to a effective population size of $2N_e$ for an autosomal locus, thus $\beta = 3N_e s$. An N_e of $19,800$ was estimated by Verrelli et al. (2002) for the region using divergence data and an estimated of $\theta = 5.35$ (the maximum full likelihood estimate under the assumption of neutrality). The maximum likelihood estimate of $\theta \approx 5.5$ and the shape of the likelihood surface in the θ direction is relatively unaffected by selection due to the low frequency of the selected allele. The surface does not have a maximum in the β direction in the range investigated but instead seems to have a horizontal asymptote. This is unsurprising as no mutations have occurred on the $A-$ allele background and thus, given that the allele segregates in the population, $\beta = \infty$ is possible. The hypothesis that $\beta = 0$ is unlikely to be rejected, but given that the allele $A-$ is at low frequency and thus arose recently there is relatively little information about the selection coefficient of the allele. Under the assumption of 20 years per generation and $N_e = 19,800$, the age in years of the various events in the genealogical history may be estimated. The mean and standard error of the age of the selected mutation, and the TMRCA, are shown in Table 6.1 in coalescent and year time units for $\beta = 0, 50, 100, 200$. The full gene tree with estimated times using a selection parameter $\beta = 200$ is shown in Figure 6.2. The estimate of the age of the $A-$ allele is strongly affected by selection. The TMRCA is relatively unaffected by selection on the $A-$ allele as the low frequency selected allele has little effect on the time in the tree.

6.2 The Factor IX Region and a Model of a Full Sweep

The X chromosome has increasingly become the focus of investigation into patterns of diversity found in human populations (Schaffner, 2004). The X chromosome recombines allowing different regions to have different, albeit correlated, histories. Thus it

Table 6.1: Summary of the age estimates for the African G6PD gene tree, evaluated at $\theta = 5.35$.

Event[a]	β^b	Age (Coalescent units)	Age (Kyr)
$A-$ mutation	0	$0.087^{+}_{-}0.049$	$52^{+}_{-}29$
	50	$0.058^{+}_{-}0.024$	$34^{+}_{-}14$
	100	$0.039^{+}_{-}0.014$	$23^{+}_{-}8.3$
	200	$0.024^{+}_{-}0.0075$	$14^{+}_{-}4.5$
TMRCA of sample	0	$1.1^{+}_{-}0.4$	$640^{+}_{-}220$
	50	$1.1^{+}_{-}0.30$	$630^{+}_{-}180$
	100	$1.0^{+}_{-}0.31$	$620^{+}_{-}180$
	200	$1.0^{+}_{-}0.30$	$610^{+}_{-}180$

a, the event for which the age is calculated. b, the selection parameter that the age is evaluated with.

affords more insight than the Y chromosome or mitochondrial genome into the demographic history of the population. An advantage specific to the X chromosome is that in males the haplotype information is known, simplifying the analysis compared to autosomal chromosomes where phasing the haplotypes can itself be a challenge. Studies of patterns of diversity in regions of the X chromosome have in general found that the data are consistent with a deep genealogical history underlying the sample (Schaffner, 2004). However a few studies have found regions of low diversity where the patterns of variation are indicative of a more 'star-like' genealogy underlying the region, see for example Nachman and Crowell (2000), Harris and Hey (2001) and Nachman et al. (2004). As these studies have included samples from both Africans and Europeans it is unlikely that the standard demographic scenarios explain these patterns. Another explanation for these patterns of diversity is that a number of selective sweeps have occurred on the X chromosome. To investigate this possibility we shall consider in more detail the data presented in Harris and Hey (2001). Harris and Hey (2001) sequenced 36

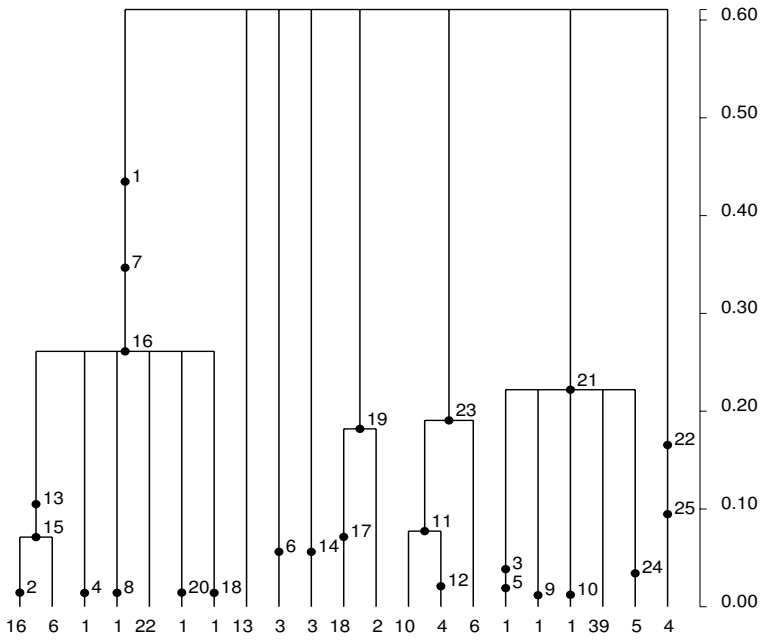

Figure 6.2: The full African G6PD rooted gene tree with estimated times (in millions of years) for $\beta = 200$, $\theta = 5.35$, $N_e = 19,800$ and a generation length of 20 years. The $A-$ allele is defined by mutation 2.

males from Old World populations (Asian, European and African) in a 3700 base pair region of the Factor IX (FIX) gene on the X chromosome. The FIX locus was found to have one of the lowest levels of diversity of any locus yet investigated on the X chromosome. By using a chimpanzee outgroup, Harris and Hey (2001) constructed a rooted gene tree for the FIX region and this is reproduced in Figure 6.3. The gene tree shows little genealogical structure, the few polymorphisms present appear at low frequency. In contrast Harris and Hey (2001) find that the gene tree of the PDHA1 region, constructed from sequences in a closely similar set of individuals to those sequenced at the FIX locus, reveals a very deep subdivision between the Africans and Europeans. This technique, also utilised by Nachman and Crowell (2000), of examining two closely similar samples of individuals is intended to reduce the effect of differing sampling

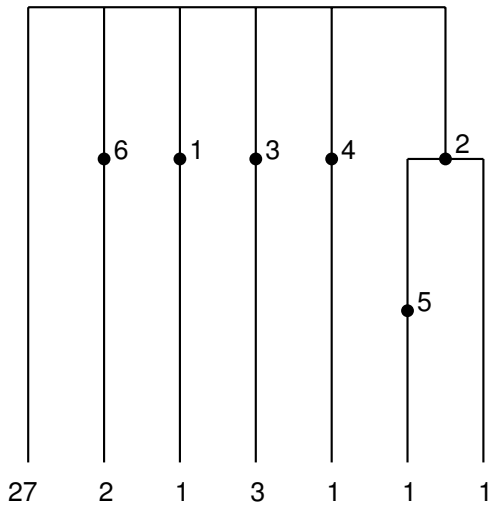

Figure 6.3: The rooted gene tree for the FIX locus for a sample of 36 males from Old World populations.

strategies leading to different signals from different regions (see for discussion Ptak and Przeworski (2002)). Both the FIX and PDHA1 loci show similar levels of divergence between human and chimpanzee, and a tree rooted with an orangutan outgroup shows that the human and chimpanzee lineages had approximately the same rate of divergence for the FIX gene. Thus neither mutation rate variation nor different sampling strategy seem to account for the low level of diversity and lack of visible genealogical structure observed in the FIX region.

To investigate whether a recent sweep through the entire population by a positively selected allele could account for the patterns of variation observed, a model of a full sweep was implemented. In this model an allele experiencing genic selection, with a population scaled selection coefficient of β, has arisen and moved through the population fixing at a scaled time T in the past, see Figure 6.4. To model the trajectory of the allele backwards in time let $\{Z(t) = N, t < T\}$ and then simulate $\{Z(t), t \geq T, X(T) = N - 1\}$ using the reversed Moran model conditional on the allele being lost from the population. (Forward in time this is a model where an allele is initially at

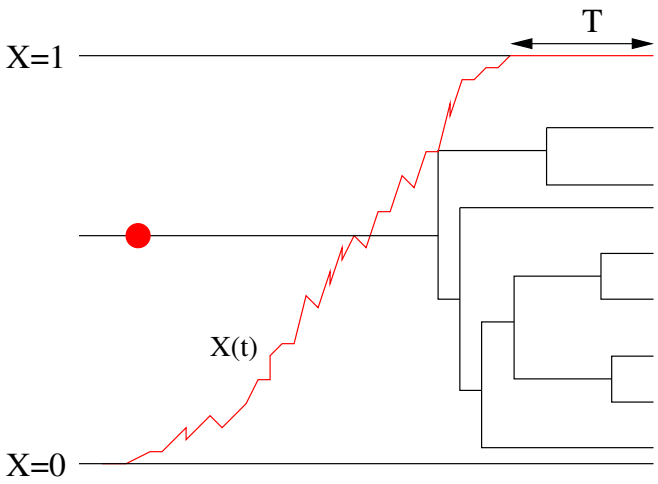

Figure 6.4: Caricature of the genealogy and the population trajectory of the selected allele. The red line is the trajectory $\{X(t) \mid t \geq 0\}$, and the red dot is the mutation on the lineage ancestral to the sample.

$Z = 1$ in the population then the path is simulated conditional on fixation at $Z = N$. The sample is taken a time T after the trajectory of the selection allele reached fixation). The likelihood of the data is calculated conditional on the allele having fixed in the population and so does not include the probability of the allele fixing in the population. If this term were to be included the unconditional likelihood relative to $\beta = 0, T = 0$, would be

$$L_0(\beta, T, \theta) = \lim_{p \to 0} \frac{L(\beta, T, \theta) u_1(p \mid \beta)}{L(\beta = 0, T = 0, \theta) u_1(p \mid \beta = 0)}, \tag{6.3}$$

where $u_1(p \mid \beta)$ is the probability from the diffusion process of an allele with selective coefficient β hitting 1 before 0, given an initial frequency p. We choose to use the conditional likelihood as it gives a better idea of the information present in the data. The pattern of diversity left by a sweeping allele with selective coefficient $-\beta$ conditional on it fixing is the same as that left by a sweeping allele with coefficient β, see the discussion below (3.30). This means that the conditional surface will be symmetric around $\beta = 0$, so we chose to only investigate the half of the surface where $\beta > 0$.

In Figure 6.5 a series of joint log likelihood surfaces for β and T for different values of θ are shown. For all values of θ the main feature of the surface is a ridge of maximum likelihood for β and T. As there are no mutations deep in the gene tree the most likely genealogies are those whose deeper branches are short relative to their expectation under the standard neutral model. This shortening is most likely when the frequency of the allele decreases rapidly backward in time after the mutations in the lower parts of the tree have occurred. For low θ more time is needed lower in the tree, thus the maximum likelihood ridge for β and T is at larger T.

The joint maximum likelihood estimate of the three parameters is found in the parameter region $\theta \geq 6$, $\beta > 150$ and $T \approx 0$. There are two null hypotheses of interest, the first is the standard neutral model, with one parameter θ, the second is that a neutral allele recently fixed in the population. The likelihood ratio test of the hypothesis that the pattern of diversity at the locus was generated by a fixation event is

$$\psi_1 = 2 \log \left(\frac{max_{\beta \geq 0, T, \theta} \{ L(\beta, T, \theta) \}}{max_{\theta} \{ L(\theta) \}} \right), \tag{6.4}$$

where $L(\theta)$ is the likelihood under the standard neutral model. $L(\theta)$ is identical to $L(\beta, T = \infty, \theta)$, the likelihood when the conditioned upon sweep is in the distant past and so has no effect on the pattern of diversity. If an allele was known to have fixed at some recent time, the hypothesis of interest might be that the allele fixing was selected rather than neutral. This could be tested with the likelihood ratio test

$$\psi_2 = 2 \log \left(\frac{max_{\beta \geq 0, T, \theta} \{ L(\beta, T, \theta) \}}{max_{T, \theta} \{ L(\beta = 0, T, \theta) \}} \right) \tag{6.5}$$

As can been seen in Figure 6.5, the (β, T, θ) likelihood surface reaches a maximum at $(\beta = 150, T \approx 0, \theta = 6)$ at the beginning of a ridge of maximum likelihood running in the direction of increasing θ, increasing β. The likelihood for the θ model (i.e. no sweep) is maximum at $\theta = 2$, and under the $(\beta = 0, T, \theta)$ model the maximum is at $T \approx 0, \theta = 2$. The test of any sweep at all compared to no sweep has $\psi_1 = 6.5$, and

$\theta = 1$ $\theta = 2$ $\theta = 3$

$\theta = 4$ $\theta = 5$ $\theta = 6$

Figure 6.5: Joint log likelihood surfaces for the strength of selection and time the sweep ended for different values of θ for the FIX gene tree. The lighter colours indicate regions of higher likelihood.

the test that the allele fixing in the population is selected has $\psi_2 = 4.6$. The rejection region is unknown for either hypothesis, as the genealogy underlying the data means that the data are nonindependent draws. The first and second tests have two and one degrees of freedom respectively. Both tests are on the boundary of the parameter space, as $\beta, T \geq 0$. The rejection region could be calculated by the empirical distribution of the test statistic, conditional on the number of sites, by the following steps:

1. simulate a genealogy under the null model in question
2. place 6 polymorphisms at random on the genealogy (note that this step is only approximate as the true distribution of genealogies given the number of mutations is dependent on θ see Markovtsova et al. (2001) and Wall and Hudson (2001) for discussion)
3. calculate the likelihood surface for θ, β, T
4. perform either test (6.4) or (6.5)
5. repeat $(1-4)$ a large number of times to form an empirical distribution of the likelihood ratio statistic, and find the λ cutoff for the desired level of significance.

The fact that the parameter θ is unknown considerably reduces the power of the tests. The tests do not take into account the lack of diversity in the region. As the likelihood is maximised over θ separately for the null and alternative hypotheses, the only information lies in the positions of the mutations in the gene tree. If θ was known then a better estimate of the strength and age of the sweep could be obtained, and a more powerful test could be performed. To find an approximate estimate of N_e we assume it to be an average of the N_es found for six other X chromosome regions, compiled in Table 2 of Harris and Hey (2001), giving an $N_e \sim 16000$. For the X chromosome $\theta = 3N_e u$ where u is the mutation rate per generation. Taking a generation length of 20 years (as in Harris and Hey (2001)) and an estimated u of 10^{-9} mutations per base per year (Harris and Hey, 2001), gives $\theta \approx 3.5$. The log likelihood surface for $\theta = 3.5$

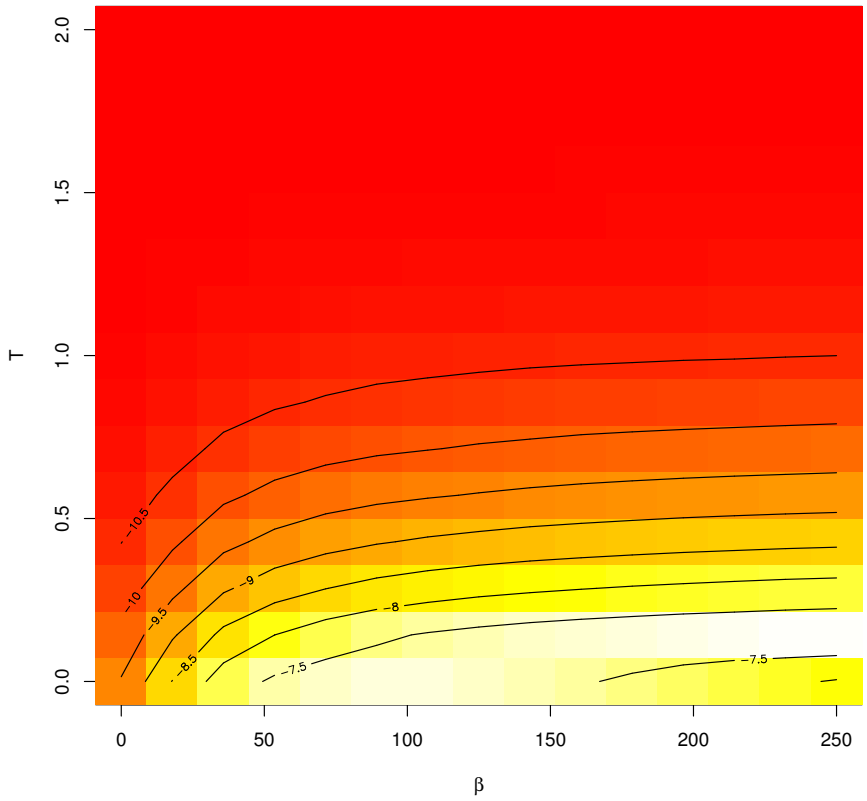

Figure 6.6: The joint log likelihood surface for the strength of selection (β) and the time the sweep finished (T) for $\theta = 3.5$ for the FIX gene tree. The lighter colours indicate regions of higher likelihood.

is shown in Figure 6.6. The test, (6.4), for presence of a putative sweep in the region conditional on $\theta = 3.5$, has a $\psi_1 = 9.9$. The test, (6.5), of a sweep being that of a selected allele conditional on $\theta = 3.5$, has a $\psi_2 = 4.8$. It is interesting to note that a model of the patterns of diversity left by the fixation of a neutral allele can offer quite an improvement over the standard neutral model, and that rejection of the standard neutral model does not necessarily imply rejection of the neutral model. See Tajima (1990) for a discussion of the effect of a neutral mutation fixing in the population. A compromise between the approach that assumes θ is unknown and the approach using a particular

value of θ would be to perform Bayesian inference. A prior on N_e could be constructed informed by the inferred N_es of other X chromosome regions. This prior, along with a prior on u the mutation rate, could then be used to integrate out θ from the likelihood to give the marginal likelihood of the data given only the selection coefficient.

6.3 DCP1 and Balancing Selection

The angiotensin converting enzyme (ACE) is involved in the pathway for controlling blood pressure and fluid electrolyte balance. ACE is encoded for by the gene DCP1. A number of association studies have found that the presence of an ALU insertion or deletion in DCP1 is associated with cardiovascular disease pathology, elite athletic performance and response to physical training (Rieder et al. (1999) and references therein). A resequencing study of the sequence variation in 24kb of the DCP1 region was performed by Rieder et al. (1999) in six European Americans and five African Americans. By comparison to chimpanzee and human specific ALU elements, the absence of the ALU element was found to be ancestral state of the human lineage. Of the 78 variable sites found in the region, 17 were in complete linkage disequilibrium with the ALU element. The authors suggest a number of explanations for this high level of linkage disequilibrium including a change in mutation rate, demographic factors or natural selection acting to maintain the ALU element or linked variation within the population. Rieder et al. (1999) construct a gene tree for the region, excluding 5 sites that were incompatible with the assumption of the infinitely-many-sites model and no recombination. The robustness of the method applied to data with departures from the modelling assumptions is not known, thus the results must be treated with caution. As the individuals were genotyped, singleton mutations were not assigned to a particular haplotype thus the gene tree excludes the 26 singleton mutations present in the data. The rooted gene tree for the region (personal communication A. G. Clark) is shown in

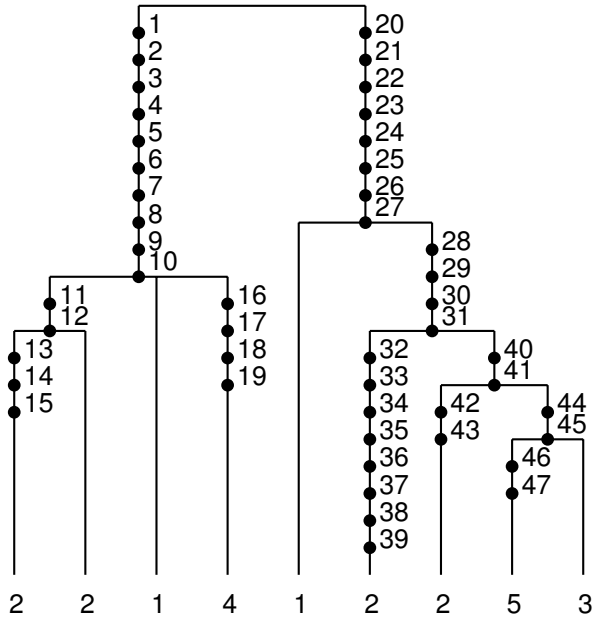

Figure 6.7: The rooted gene tree for the DCP1 locus for a sample of 11 males from African and European populations

Figure 6.7.

To investigate the hypothesis that balancing selection has acted to maintain the ALU element or linked variation, the likelihood of the gene tree in Figure 6.7 under a model of complete overdominant (heterozygote advantage) selection at a biallelic locus was calculated. The balanced polymorphism was assumed to be one of the mutations on the edge subtended by the 9 individuals with the ALU polymorphism, i.e. the selected mutation is any one of the mutations labelled 1 to 10 in the gene tree. As the gene tree is only partially ordered up to the ordering of mutations on an edge, the selected mutation could be in any position on the edge. The possible positions of the selected mutation on the edge must be summed over

$$L(\beta,\theta) = \sum_{k=1}^{M} L(\beta,\theta, \text{selected mutation is the } k^{th} \text{ on the edge}) \qquad (6.6)$$

where M is the total number of mutations on the same edge as the selected mutation.

Let the time until the selected subsample finds its common ancestor be denoted by s and the age of the mutation be denoted by τ. The likelihood of $k-1$ mutations occurring in the time $\tau - s$ on the edge below the selected mutation is Poisson with parameter $\theta(\tau - s)/2$, i.e. the term appearing in the likelihood is

$$\frac{(\theta(\tau - s)/2)^{k-1} e^{-\theta(\tau - s)/2}}{(k-1)!}. \tag{6.7}$$

This is equivalent to, and replaces, the $\exp(-\theta(\tau - u)/2)$ term that appears in the $r_1 = 0$ entries of Table 5.2. The joint surface for θ and β is shown in Figure 6.8. A large value of β will maintain the selected polymorphism for a long time in the population and so will lead to a long time in the root edges of the tree compared to neutrality. Hence low values of θ have higher relative likelihoods for $\beta > 0$ than $\beta = 0$ compared to high θ where less time in the root branches is necessary to produce the observed pattern of diversity. The likelihood surface peaks close to $\beta = 0$ for the maximum likelihood estimate of θ. The likelihood curve over β for a particular θ should be relatively unaffected by the exclusion of singletons from the tree as balancing selection, at a biallelic locus, has only a small effect on the time in singleton edges (Spencer and McVean, 2004). More problematic is the effect of the exclusion of singletons on the estimate of θ, as the mutation rate obviously plays a key role in the inference of the amount of time in the root edges and hence inference about the strength of balancing selection. The maximum likelihood estimate of θ from the data set including singletons will almost surely be higher than that of the data set excluding them. Given this, it is unlikely that the hypothesis of $\beta > 0$ will be accepted. One simple way that the singletons could be used to inform the analysis is by noting that the expected number of singleton mutations is θ and so the number of singleton mutations may be used as a moment estimate of θ (Fu and Li, 1993). This yields an estimate of $\theta = 26$. For this value of θ the maximum likelihood estimate of β is close to 0 and the hypothesis of $\beta > 0$ is even less likely to be accepted.

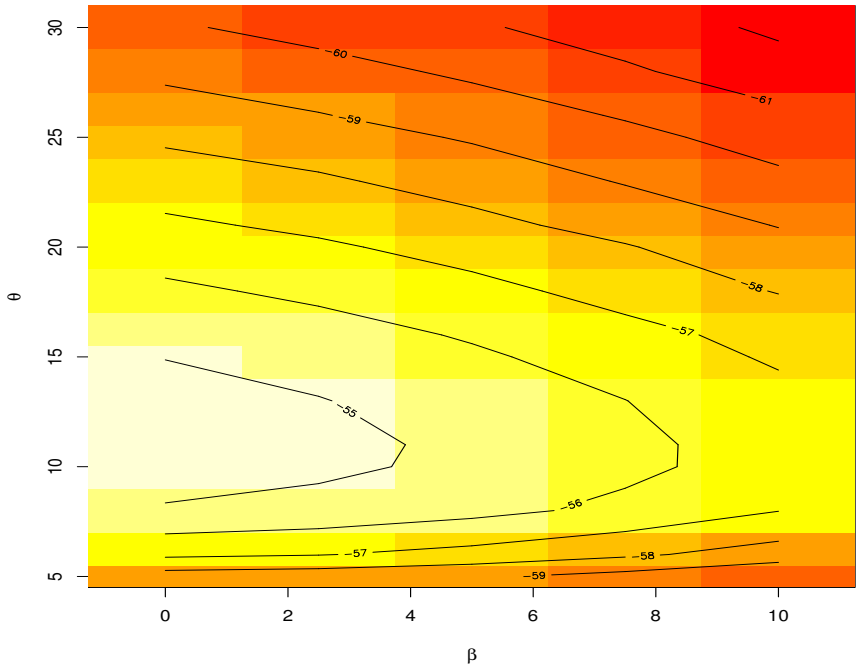

Figure 6.8: The joint log likelihood surface for θ and β for the gene tree of the DCP1 region. The lighter colours indicate regions of higher likelihood.

Chapter 7

Conclusion

A full likelihood method for inference on the selection coefficient of a single mutation in a gene tree has been developed (Coop and Griffiths, 2004). This method utilises both theoretical results from Diffusion processes and Markov Chain theory and computation techniques such as importance sampling. Throughout the thesis the idea of reversibility of the trajectory of the allele has been central. A review of some of the different views on this reversibility and its development through time by successive authors is provided in Chapter 3. The author's personal opinion is that a Green function approach where the average occupancy of a state is considered as an unnormalised prior or stationary distribution on the frequency is particularly appealing. Its appeal is mainly due to simplicity, as other approaches call for more complicated recurrent processes in order to achieve a true stationary distribution. However a full understanding of the reversibility is perhaps only gained in the light of the recurrent processes.

A full stochastic approach to simulate a genealogy where a single site experiences natural selection has been developed. This incorporates all the sources of uncertainty present in the model including the frequency of the selected allele in the current day

and the trajectory of the selected allele through the population. The present population frequency is drawn from the distribution of the population frequency given the sample configuration of the selected allele and its selection coefficient, see Chapter 3 and Griffiths (2003). Given the current population frequency the trajectory may be modelled, utilising the reversibility of the diffusion, by a forward in time diffusion process conditional on the allele being lost from the population (Nagasawa and Maruyama, 1979). We chose to approximate this by first approximating the forward in time unconditional diffusion with a Moran model with birth and death rates determined by the infinitesimal mean and variance of the diffusion. This Moran model is then simulated conditional on the allele being lost from the population. Given the trajectory, the genealogical process is identical to that of a two deme subdivided population with the frequency of the selected allele giving the size of the demes through time (Kaplan et al., 1988). The extension of the simulation method to models with repeat mutation at the selected site and recombination is discussed and the latter has been implemented (Spencer and Coop, 2004).

A gene tree, under the assumptions of the infinitely-many-sites model and no recombination, may uniquely represent a sample of sequence data. The likelihood of the gene tree may be written as an integral over the missing data of the trajectory of the selected allele and the genealogy relating the sample members, given this trajectory. The probability of a genealogy can be expressed as a Markov chain on history states embedded in the genealogy. The history states of a gene tree where the trajectory is known were discussed in Section 5.2.3. The process on history states is similar to those of Griffiths and Tavaré (1994b) and Bahlo and Griffiths (2000) who study the probability of a gene tree in a varying sized and subdivided population, respectively. To explore the space of genealogies consistent with the data a sequential importance sampling scheme is employed, the initial state of which is the history state in the current day, i.e. the

data. This importance sampling scheme draws from the exact distribution for the time between events and the subpopulation the event occurs in. The choice of event given the subsample is taken using the Stephens and Donnelly (2000) importance sampler. Trajectories generated at a driving selection coefficient may be used to evaluate the likelihood at other selection coefficients away from this driving value using importance sampling. This technique is not specific to the gene tree method and could be used to improve the performance of any inference method that makes use of stochastic trajectories. The likelihood curves from different β driving values could be combined using the technique of bridge sampling (Meng and Wong, 1996; Fearnhead and Donnelly, 2001). It was noted, given the trajectory, that the subtree below the selected site and the remaining tree are independent. This observation leads to an estimate of the likelihood with reduced variance by averaging over runs independently, for a given trajectory, for the two subtrees before combining the estimates. A possible improvement to the method, that has not been implemented, is to use the 'prune and enrichment' scheme, see for example Liu (2001). This has been used in related problems to greatly improve the efficiency of importance sampling.

To illustrate possible applications of the method it was applied to three data sets where selection was hypothesised to have occurred. The first was from a study of the G6PD region by Verrelli et al. (2002). In this case a particular mutation is known to have a favourable phenotypic effect (Ruwende et al., 1995) and the allele is believed to have recently increased in frequency in Africa. However the gene tree for the region gave little evidence of this partial sweep, perhaps because of the low frequency of the allele in the sample. Other authors have found evidence for the sweep in different data sets. Tishkoff et al. (2001) uses microsatellites, the high mutation rate of which will allow a better estimate of the age of the allele. Saunders et al. (2002) and Sabeti et al. (2002) investigate the extent of haplotype sharing away from the selected site in a surround-

ing large region. They find that the extent of long range haplotype sharing provides good evidence of the young age of the allele. A relatively short resequenced region does not have a high enough overall mutation rate to distinguish between a short genealogy beneath a low frequency neutral allele and a very short genealogy beneath a low frequency selected allele. Only marker types where events occurs at a high rate (e.g. microsatellite mutations or recombination between distant SNPs) are fast enough to notice the difference between the two events.

In the second analysis the evidence for a putative full sweep in the recent history of the FIX region, sequenced by Harris and Hey (2001), was examined. The likelihood surface indicates that it is likely that the sweep has occurred in the recent past of the region leading to the patterns of diversity seen. A point of interest is the effect that a neutral mutation that has recently reached fixation could have on patterns of diversity. A neutral allele is of course far less likely to have fixed in the population, but if the hypothesis is asked conditional on the fixation, it may be hard to exclude the fixing neutral allele hypothesis. The lack of information about the population scaled mutation rate considerably reduces the strength of the inferences made about the selection coefficient and timing of the sweep. The possibility of using information from other regions to improve our knowledge of θ and hence the inference is suggested. The first suggested method described conditions on an estimate of θ constructed from other regions. The second suggested method incorporates the uncertainty in the estimate of θ by adopting a Bayesian approach to the problem by placing a prior on θ informed by other regions. θ may then be integrated out of the joint likelihood of (θ, β, T), where T is the fixation time, with respect to the prior on θ to give the marginal surface for β and T.

The final data set examined was that of Rieder et al. (1999) where the patterns of diver-

sity had led the authors to suggest that balancing selection may have played a role in the evolution of the locus. There were a number of possible positions of the selected mutation in the gene tree, as there are multiple mutations on the same edge as the putative selected mutation, and these possible locations were summed over. The joint likelihood surface for the population scaled mutation rate and the strength of selection suggest that balancing selection is unlikely to be the cause of the pattern of diversity found at the locus.

Another possible application of the importance sampling algorithm is a data set where a selective sweep is believed to be in progress but the position of the selected mutation on the edges in the gene tree is not known. In this case the likelihood surface for β and θ could be produced for each of the candidate selected mutations in the gene tree, and the likelihood ratio test performed. The mutation with the highest likelihood ratio could be said to be the best candidate for the selected site.

It is inappropriate to use the gene tree inference method if recombination or repeat mutation is thought to have played a major role in the evolution of the region of interest. Recent research has suggested that the human genome is made up of regions where the recombination rate is relatively low interspersed with short stretches with high recombination, termed recombination hotspots, see for example McVean et al. (2004) and Kauppi et al. (2004). Hence, an inference method that assumes a tightly linked region may be appropriate for the analysis of the patterns of diversity found in some regions. Bahlo and Griffiths (2000) study the distribution of gene trees in a subdivided population model with recurrent migration. If recombination were to be included in the inference it would be incorporated as migration between the two selective classes as discussed in Section 4.3, and developed in Hudson and Kaplan (1988) and Barton and Etheridge (2004). Full likelihood inference for models of recombination is already

a considerable challenge, thus the development of an effective full likelihood inference method for selection and recombination seems unlikely at present. A more appealing method for inference about selection in a recombining region is offered by summary statistics inference methods, see for example Przeworski (2003), as they are computationally more practical and easier to implement than full likelihood methods. Work has been done by others to improve the computational efficiency of summary statistic inference, see for example Beaumont et al. (2002) and Marjoram et al. (2003), and so more complex summaries of the data may be used allowing summary statistic inference to become more powerful. Another promising avenue of research is the development of approximate coalescent methods for inference, e.g. about recombination rates (Li and Stephens, 2003). An approximate coalescent method for inference about selection could be constructed in a similar way, either by using ideas from the ancestral selection graph or trajectories.

The full likelihood method uses all of the information present in the data and so is the most powerful method. A few heuristic observations on the power and robustness of the method will now be discussed. When a selected allele is at low frequency in the population there will be little to distinguish it from a neutral allele as both will be young. In the case of a selective sweep, if θ is known then the full likelihood approach will be most powerful when the allele has just reached a high frequency in the population. This allows the difference in age between the selected allele and a neutral allele to be apparent. However, when θ is unknown, more time in the overall genealogy better informs our inference on θ. Better knowledge of θ helps to improve our estimation of times in the tree and hence aids estimation of β. For a partial sweep this problem is naturally solved as the nonselected subtree provides information on θ. As long as there are a reasonable number of nonselected sample members, a strongly selected allele at medium to high frequencies should be detectable. A recent full sweep, leaves a shallow

tree which offers little information on θ. This reduces the strength of inference about β, as was seen in the FIX locus example. On the other hand, the time in the deep ancestral lineages is highly variable so if the sweep occurs too far back in the history there will be little information about β. A more powerful situation is one where a strong sweep occurs a short time into the history allowing mutations to occur lower in the tree. This, under neutrality, will seem at odds with the paucity of mutations deeper in the tree, and a model of a full sweep will be more plausible.

Wiuf (2003) studies the information about demographic parameters available if the full tree, including times, is known. He finds that there is relatively little information about the exact demographic parameters underlying the evolution of the tree. Even if the genealogy is that of the whole population, the estimate of the parameters can behave poorly and is statistically inconsistent. Given the analogies between the coalescent in a varying sized population and the coalescent when the trajectory is known, it is likely that there is relatively little information about the shape of the trajectory even if the genealogy is known. Watterson (1979) finds that the full information about the trajectory is insufficient for a consistent estimate of β. This taken with the lack of information in the genealogy about the trajectory suggests that only the crudest indication of the selection coefficient is in general possible.

Different demographies can often leave similar patterns of diversity to the signal of selection, see Kreitman (2000) and Charlesworth et al. (2003) for reviews. A brief reduction in population size, a bottleneck, can lead to the rapid coalescence of some lineages while other lineages coalesce after the bottleneck. This results in an excess of intermediate and low frequency alleles, which can resemble the pattern of diversity under a model of a partial sweep. However the gene tree approach is perhaps less susceptible to this than some simple summary statistic approaches. For the gene tree to

be likely under model of a partial sweep, the coalescences before the bottleneck must mainly subtend a particular lineage surviving the bottleneck. Under a bottleneck model, lineages coalescing during the bottleneck subtend at random lineages surviving the bottleneck. Thus, a large clustering of coalesces under a particular lineage is reasonably unlikely. A more convincing signal of a partial sweep from a gene tree might perhaps be given by a gene tree generated under a two deme subdivided population model, where one deme is large and the other is relatively small. A mutation on the lineage ancestral to sample members that coalesce rapidly in the small deme will appear to be selected if the subdivision is ignored.

Rapid population expansion will leave a similar pattern of variation to that of a full sweep of a selected allele. A deep population subdivision will increase the time in the basal branches and this will leave an identical pattern to balancing selection. It is clear that an apparent signal of selection must be considered carefully. Selection will only affect loci that are in tight linkage with the selected locus. Demography, unlike selection, affects all loci equally. A putatively selected locus should be compared to other loci, sequenced in a closely similar sample, to see if the locus of interest has a substantively different pattern of variation. However, there can still be a great deal of variation in the pattern of diversity at different loci even if the same demographic model applies to all loci. Thus caution must be exercised in any conclusions drawn about the action of selection. A quantitative method to examine the effect of demography over loci and selection at a particular locus, would be to estimate the demographic parameters from other the loci and then build, and perform inference under, a model of selection and demography using this information at the locus of interest.

A general framework for inference on parameters in a subdivided random environment coalescent process has been developed here. This approach allows an intuitive under-

standing of the incorporation of natural selection into the coalescent. The generality of the diffusion process and the coalescent would allow a wide variety of random background models, where the generator of the diffusion is known, to be simulated from and inference to be performed. For example, the method offers an inference framework into which recombination and recurrent mutation could be incorporated at a later date.

The importance of fluctuations in frequency away from the deterministic approximation to a trajectory of a selected allele is not well understood and will often vary depending on the type of selection considered (Barton and Etheridge, 2004). The computationally efficient fully stochastic treatment of selection and the coalescent described in this thesis allows these fluctuations to be incorporated. The method is valid for a large number of selective schemes and strengths of selection, and thus the method offers a viable alternative to the ancestral selection graph and deterministic trajectories.

Bibliography

Akey, J. M., Zhang, G., Zhang, K., Jin, L. and Shriver, M. D. (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.*, **12**, 1805–1814.

Bahlo, M. and Griffiths, R. C. (2000) Inference from gene trees in a subdivided population. *Theor. Popul. Biol.*, **57**, 79–95.

Bamshad, M. and Wooding, S. P. (2003) Signatures of natural selection in the human genome. *Nat. Rev. Genet.*, **4**, 99–111.

Barbour, A. B., Ethier, S. N. and Griffiths, R. C. (2000) A transition function expansion for a diffusion model with selection. *Ann. Appl. Probab.*, **10**, 123–162.

Barton, N. H. and Etheridge, A. M. (2004) The effect of selection on genealogies. *Genetics*, **166**, 1115–1131.

Barton, N. H., Etheridge, A. M. and Strum, A. K. (2004) Coalescence in a random background. *Ann. Appl. Probab.*, **14**, 754–785.

Beaumont, M. A. and Balding, D. J. (2004) Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.*, **13**, 969–980.

Beaumont, M. A., Zhang, W. and Balding, D. J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Begun, D. J., Betancourt, A. J., Langley, C. H. and Stephan, W. (1999) Is the fast/slow allozyme variation at the Adh locus of *Drosophila melanogaster* an ancient balanced polymorphism? *Mol. Biol. Evol.*, **16**, 1816–1819.

Bustamante, C. D., Nielsen, R. and Hartl, D. L. (2003) Maximum likelihood and bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theor. Popul. Biol.*, **63**, 91–103.

Bustamante, C. D., Nielsen, R., Sawyer, S. A., Olsen, K. M., Purugganan, M. D. and Hartl, D. L. (2002) The cost of inbreeding in *Arabidopsis*. *Nature*, **416**, 531–534.

Charlesworth, B., Charlesworth, D. and Barton, N. H. (2003) The effects of genetic and geographic structure on neutral variation. *Annu. Rev. Eco. Evol. S.*, **34**, 99–125.

Charlesworth, D., Charlesworth, B. and Morgan, M. T. (1995) The pattern of neutral molecular variation under the background selection model. *Genetics*, **141**, 1619–1632.

Coop, G. and Griffiths, R. C. (2004) Ancestral inference on gene trees under selection. *Theor. Popul. Biol.* To Appear.

Crow, J. F. and Kimura, M. (1970) *An introduction to population genetics theory*. Minneapolis: Alpha Editions.

Darwin, C. (1859) *The origin of species*. Harmondsworth: Penguin, 1, reprint edn.

De Iorio, M. and Griffiths, R. C. (2004a) Importance sampling on coalescent histories, I. *Adv. Appl. Probab.*, **36**, 417–433.

— (2004b) Importance sampling on coalescent histories, II. Subdivided population models. *Adv. Appl. Probab.*, **36**, 434–454.

De Iorio, M., Griffiths, R. C., Leblois, R. and Rousset, F. (2004) Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theor. Popul. Biol.* Submitted.

Diamond, J. (2003) The double puzzle of diabetes. *Nature*, **423**, 599–602.

Donnelly, P. (1984) The transient behaviour of the Moran model in population genetics. *Math. Proc. Camb. Philos. Soc.*, **95**, 349–358.

Donnelly, P. and Kurtz, T. G. (1999) Genealogical processes for Fleming-Viot models with selection and recombination. *Ann. Appl. Probab.*, **9**, 1091–1148.

Donnelly, P., Nordberg, M. and Joyce, P. (2001) Likelihoods and simulation methods for a class of nonneutral population genetics models. *Genetics*, **159**, 853–867.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. and Solomon, W. (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307–1320.

Enard, W., Przeworski, M., Fisher, S. E., Lai, C. S., Wiebe, V., Kitano, T. Monaco, A. P. and Pääbo, S. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, **418**, 869–872.

Ewens, W. J. (1963) The diffusion equation and a pseudo-distribution in genetics. *J. R. Stat. Soc. B*, **25**, 405–412.

— (1964) The pseudo-transient distribution and its uses in genetics. *J. Appl. Prob.*, **1**, 141–156.

— (1972) The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, **3**, 87–112.

— (1973) Conditional diffusion processes in population genetics. *Theor. Popul. Biol.*, **4**, 21–30.

— (1979) *Mathematical population genetics*, vol. 9 of *Biomathematics*. New York: Springer-Verlag.

Fearnhead, P. (2001) Perfect simulation of population genetic models with selection. *Theor. Popul. Biol.*, **59**, 263–279.

— (2002) The common ancestor at a non-neutral locus. *J. Appl. Prob.*, **39**, 38–54.

— (2003) Ancestral processes for non-neutral models of complex diseases. *Theor. Popul. Biol.*, **63**, 115–130.

Fearnhead, P. and Donnelly, P. (2001) Estimating recombination rates from population genetic data. *Genetics*, **159**, 1299–1318.

Feller, W. (1968) *An introduction to probability theory and its applications*, vol. 2. New York: Wiley.

Felsenstein, J., Kuhner, M. K., Yamato, J. and Beerli, P. (1999) Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In *Statistics in Molecular Biology and Genetics* (ed. F. Seillier-Moiseiwitsch), vol. 33 of *IMS Lect. Notes Monogr. Ser.*, 163–185.

Fisher, R. A. (1930) *The genetical theory of natural selection: a complete variorum edition*. Oxford: Oxford University Press.

Fu, Y. X. and Li, W. H. (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Fullerton, S. M., Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Stengard, J. H., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E. and Sing, C. F. (2000) Apolipoprotein E variation at the sequence haplotype level: Implications for the origin and maintenance of a major human polymorphism. *Am. J. Hum. Genet.*, **67**, 881–900.

Griffiths, R. C. (1980) Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theor. Popul. Biol.*, **17**, 37–50.

— (1989) Genealogical-tree probabilities in the infinitely-many-sites model. *J. Math. Biol.*, **27**, 667–680.

— (2002) Ancestral inference from gene trees. In *Modern Developments in Theoretical Population Genetics* (eds. M. Veuille and M. Slatkin), 94–117. Oxford: Oxford University Press.

— (2003) The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theor. Popul. Biol.*, **64**, 241–251.

Griffiths, R. C. and Majoram, P. (1996) Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, **3**, 479–502.

Griffiths, R. C. and Tavaré, S. (1994a) Simulating probability distributions in the coalescent. *Theor. Popul. Biol.*, **46**, 131–159.

— (1994b) Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. Lond. B*, **344**, 403–310.

— (1998) The age of a mutation in a general coalescent tree. *Stochastic Models*, **14**, 273–295.

— (2003) The genealogy of a neutral mutation. In *Highly structured stochastic systems* (eds. P. J. Green, N. L. Hjort and S. Richardson), 94–117. Oxford: Oxford University Press.

Gusfield, D. (1991) Efficient algorithms for inferring evolutionary trees. *Networks*, **21**, 19–28.

Hamblin, M. T. and Di Rienzo, A. (2000) Detection of the signature of natural selection in humans: Evidence from the Duffy blood group. *Am. J. Hum. Genet.*, **66**, 1669–1679.

Harris, E. E. and Hey, J. (2001) Human populations show a reduced DNA sequence variation at the Factor IX locus. *Curr. Biol.*, **11**, 774–778.

Hoppe, F. M. (1984) Pólya-like urns and the Ewens' sampling formula. *J. Math. Biol.*, **20**, 91–94.

Hudson, R. R., Bailey, K., Skarecky, D., Kwiatowski, J. and Ayala, F. (1994) Evidence for positive selection in the Superoxide Dismutase (Sod) region of *Drosophila melanogaster*. *Genetics*, **136**, 1329–1340.

Hudson, R. R. and Kaplan, N. L. (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, **111**, 147–164.

— (1988) The coalescent process in models with selection and recombination. *Genetics*, **120**, 831–840.

Hudson, R. R., Kreitman, M. and Aguade, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.

Kaplan, N. L., Darden, T. and Hudson, R. R. (1988) The coalescent process in models with selection. *Genetics*, **120**, 819–829.

Karlin, S. and Taylor, H. (1975) *A first course in stochastic processes*. London: Academic Press.

— (1981) *A second course in stochastic processes*. London: Academic Press.

Kauppi, L., Jeffreys, A. J. and Keeney, S. (2004) Where the crossovers are: recombination distributions in mammals. *Nat. Gen. Rev.*, **5**, 413–424.

Keilson, J. (1965) A review of transient behaviour in regular diffusion and birth and death processes II. *J. Appl. Prob.*, **2**, 405–428.

Kelly, F. P. (1977) Exact results for the Moran neutral allele model. *J. Appl. Prob.*, **9**, 197–201.

Kim, Y. and Stephan, W. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, **160**, 765–777.

Kimura, M. (1964) *Diffusion models in population genetics*, vol. 2 of *Methuen's review series in applied probability*, 42–47. London: Methuen.

Kimura, M. and Ohta, T. (1969) The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, **61**, 763–771.

— (1973) The age of a neutral mutation persisting in a finite population. *Genetics*, **75**, 199–212.

Kingman, J. F. C. (1982) The coalescent. *Stochastic Processes Appl.*, **13**, 235–248.

Kreitman, M. (2000) Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics. Hum. Genet.*, **1**, 539–559.

Krone, S. M. and Neuhauser, C. (1997) Ancestral processes with selection. *Theor. Popul. Biol.*, **51**, 210–237.

Levikson, B. (1977) The age distribution of a Markov chain. *J. Appl. Prob.*, **14**, 492–506.

Lewontin, R. C. and Krakauer, J. (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Li, N. and Stephens, M. (2003) Modelling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165**, 2213–2233.

Liu, J. S. (2001) *Monte Carlo strategies in scientific computing*. New York: Springer Verlag.

Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003) Markov Chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci.*, **23**, 15324–15328.

Markovtsova, L., Marjoram, P. and Tavare, S. (2001) On a test of Depaulis and Veuille. *Mol. Biol. Evol.*, **18**, 1132 – 1133.

Maruyama, T. (1972) The average number and variance of generations at a particular gene frequency in the course of fixation of a mutant gene in a finite population. *Genetical Research*, **19**, 109–114.

— (1974) The age of an allele in a finite population. *Genetical Research*, **23**, 137–143.

— (1977) *Stochastic Problems in Population Genetics*. Lecture Notes in Biomathematics. New York: Springer-Verlag.

Maruyama, T. and Fuerst, P. (1983) Analyses of the age of genes and the first arrival times in a finite population. *Genetics*, **105**, 1041–1059.

Maruyama, T. and Kimura, M. (1971) Some methods for treating continuous stochastic processes in population genetics. *Japanese Journal of Genetics*, **46**, 407–410.

— (1974) A note on the speed of gene frequency changes in reverse directions in a finite population. *Evolution*, **28**, 161–163.

— (1975) Moments for the sum of an arbitrary function of gene frequency along a stochastic path of gene frequency change. *Proc. Nat. Acad. Sci.*, **72**, 1602–1604.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. and Donnelly, P. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–584.

Meng, X. and Wong, W. H. (1996) Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat. Sinica*, **6**, 831–860.

Moran, P. A. P. (1958) Random processes in genetics. *Proc. Camb. Phil. Soc.*, **54**, 60–71.

Nachman, M. W. and Crowell, S. L. (2000) Contrasting evolutionary histories of two introns of the duchenne muscular dystrophy gene, Dmd, in humans. *Genetics*, **155**, 1855–1864.

Nachman, M. W., D'Agostino, S. L., Tillquist, C. R., Mobasher, Z. and Hammer, M. F. (2004) Nucleotide variation at Msn and Alas2, two genes flanking the centromere of the X chromosome in humans. *Genetics*, **167**, 423–437.

Nagasawa, M. and Maruyama, T. (1979) An application of time reversal of Markov processes to a problem of population genetics. *Adv. Appl. Prob.*, **11**, 457–478.

Neuhauser, C. (1999) The ancestral selection graph and gene genealogy under frequency-dependant selection. *Theor. Popul. Biol.*, **56**, 203–214.

Neuhauser, C. and Krone, S. M. (1997) The genealogy of samples in models with selection. *Genetics*, **145**, 519–534.

Nielsen, R. (1999) Changes in d_s/d_n in the HIV-1 env gene. *Mol. Biol. Evol.*, **16**, 711–714.

— (2001a) Mutations as missing data: Inferences on the ages and distributions of non-synonymous and synonymous mutations. *Genetics*, **159**, 401–411.

— (2001b) Statistical tests of neutrality in the age of genomics. *Heredity*, **86**, 641–647.

Nielsen, R. and Weinreich, D. (1999) The age of nonsynonymous and synonymous mutations in animal mtDNA and implications for the mildly deleterious theory. *Genetics*, **153**, 497–506.

Nordborg, M. (2001) Coalescent theory. In *Handbook of Statistical Genetics* (eds. D. J. Balding, M. Bishop and C. Cannings), 179–212. Chichester: Wiley.

Ohta, T. and Gillespie, J. H. (1996) Development of neutral and nearly neutral theories. *Theor. Popul. Biol.*, **49**, 128–142.

Pakes, A. G. (1979) The age of a Markov process. *Stochastic Process. Appl.*, **8**, 277–303.

Pakes, A. G. and Tavaré, S. (1981) Comments on the age distribution of Markov processes. *Adv. Appl. Probab.*, **13**.

Patterson, N. J. (2004) How old is the most recent ancestor of two copies of an allele? Submitted.

Payseur, B. A., Cutter, A. D. and Nachman, M. W. (2002) Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol. Biol. Evol.*, **19**, 1143–1153.

Perlitz, M. and Stephan, W. (1997) The mean and variance of the number of segregating sites since the last hitchhiking event. *J. Math. Biol.*, **36**, 1–23.

Provine, W. B. (1971) *The origins of theoretical population genetics*. Chicago: University of Chicago Press.

Przeworski, M. (2003) Estimating the time since the fixation of a beneficial allele. *Genetics*, **164**, 1667–1676.

Ptak, S. E. and Przeworski, M. (2002) Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Genet.*, **18**, 559–563.

Rieder, M. J., Taylor, S., Clark, A. G. and Nickerson, D. A. (1999) Sequence variation in the angiotensin converting enzyme. *Nature Genet.*, **22**, 59–62.

Römpler, H., Schulz, A., Pitra, C., Coop, G., Przeworski, M., Pääbo, S. and Schöneberg, T. (2004) The rise and fall of the chemoattractant receptor GPR33. In preparation.

Rosenberg, N. A. and Hirsh, A. E. (2003) On the use of star-shaped genealogies in inference of coalescence times. *Genetics*, **164**, 1677–1682.

Ruwende, C., Khoo, S., Snow, R., Yates, S., Kwiatkowski, D., Gupta, S., Warn, P., Allsopp, C., Gilbert, S., Peschu, N., Newbold, C., Greenwood, B., Marsh, K. and Hill, A. (1995) Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature*, **376**, 246–249.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. and Lander, E. S. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Saunders, M. A., Hammer, M. F. and Nachman, M. W. (2002) Nucleotide variability at G6PD and the signature of malarial selection in humans. *Genetics*, **162**, 1849–1861.

Sawyer, S. (1977) On the past history of an allele now known to have frequency p. *J. Appl. Prob.*, **14**, 439–450.

Sawyer, S. and Hartl, D. (1992) Population genetics of polymorphism and divergence. *Genetics*, **132**, 1161–1176.

Schaffner, S. F. (2004) The X chromosome in population genetics. *Nat. Rev. Gen.*, **5**, 43–51.

Slade, P. F. (2000a) Most recent common ancestor probability distributions in gene genealogies under selection. *Theor. Popul. Biol.*, **58**, 291–305.

— (2000b) Simulation of selected genealogies. *Theor. Popul. Biol.*, **57**, 35–49.

Slatkin, M. (2001) Simulating genealogies of selected alleles in a population of variable size. *Genet. Res.*, **78**, 49–57.

— (2002) The age of alleles. In *Modern Developments in Theoretical Population Genetics* (eds. M. Veuille and M. Slatkin), 233–260. Oxford: Oxford University Press.

Slatkin, M. and Hudson, R. R. (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, **129**, 555–562.

Spencer, C. and Coop, G. (2004) Selsim: A program to simulate population genetic data with selection and recombination. *Bioinformatics*. To appear.

Spencer, C. and McVean, G. A. T. (2004) The effect of selection on population genetic estimates of recombination. In preparation.

Stephens, M. (2000) Times on trees and the age of an allele. *Theor. Popul. Biol.*, **57**, 109–119.

— (2001) Inference under the coalescent. In *Handbook of Statistical Genetics* (eds. D. Balding, M. Bishop and C. Cannings). Chichester: Wiley.

Stephens, M. and Donnelly, P. (2000) Inference in molecular population genetics. *J. R. Stat. Soc. B*, **62**, 605–655.

— (2003) Ancestral inference in population genetics models with selection. *Aust. N. Z. J. Stat.*, **45**, 395–430.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

— (1990) Relationship between DNA polymorphism and fixation time. *Genetics*, **125**, 447–454.

Takahata, N., Satta, Y. and Klein, J. (1992) Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics*, **130**, 925–938.

Tavaré, S. (1979) Dual diffusions, killed diffusions, and the age distribution problem in population genetics. *Theor. Popul. Biol.*, **16**, 253–266.

— (1980) Time reversal and age distributions. I. Discrete-time Markov chains. *J. Appl. Probab.*, **17**, 33–46.

— (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.*, **26**, 119–164.

Tishkoff, S. A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J., Piro, A., Stoneking, M. F., Tagarelli, A., Tagarelli, G., Touma, E. H., Williams, S. M. and Clark, A. G. (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science*, **293**, 455–462.

van Herwaarden, O. A. and van der Wal, N. J. (2002) Extinction time and age of an allele in a large finite population. *Theor. Popul. Biol.*, **61**, 311–318.

Verrelli, B. C., McDonald, J. H., Argyropoulos, G., Destro-Bisol, G., Froment, A., Drousiotou, A., Lefranc, G., Helal, A., Loiselet, J. and Tishkoff, S. A. (2002) Evidence for balancing selection from nucleotide sequence analyses of human G6PD. *Am. J. Hum. Genet.*, **71**, 1112–1128.

Wall, J. D. and Hudson, R. R. (2001) Coalescent simulations and statistical tests of neutrality. *Mol. Biol. Evol.*, **18**, 1134 – 1135.

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

— (1976) Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. *Theor. Popul. Biol.*, **10**, 239–253.

— (1977) Reversibility and the age of an allele. II. Two allele models, with selection and mutation. *Theor. Popul. Biol.*, **12**, 179–196.

— (1979) Estimating and testing selection: The two-alleles, genic selection diffusion model. *Adv. Appl. Prob.*, **11**, 14–30.

— (1996) Motoo Kimura's use of diffusion theory in population genetics. *Theor. Popul. Biol.*, **12**, 154–188.

Watterson, G. A. and Guess, H. A. (1977) Is the most frequent allele the oldest? *Theor. Popul. Biol.*, **11**, 141–160.

Wiehe, T. H. E. (1998) The effect of selective sweeps on the variance of the allele distribution of a linked multiallele locus: hitchhiking of microsatellites. *Theor. Popul. Biol.*, **53**, 272–83.

Wilson, I. J., Weale, M. E. and Balding, D. J. (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. R. Stat. Soc. A*, **166**, 155–187.

Wiuf, C. (2001) Rare alleles and selection. *Theor. Popul. Biol.*, **58**, 287–296.

— (2003) Inferring population history from genealogical trees. *J. Math. Biol.*, **46**, 241–264.

Wiuf, C. and Donnelly, P. (1999) Conditional genealogies and the age of a neutral mutant. *Theor. Popul. Biol.*, **56**, 183–201.

Wright, S. (1938) The distribution of gene frequencies under irreversible mutation. *Proceeding of the National Academy of Sciences*, **24**, 253–259.

— (1968) *Theory of gene frequencies*, vol. 2 of *evolution and the genetics of populations*. Chicago: University of Chicago Press.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A. M. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.

Zollner, S., Wen, X., Hanchard, N. A., Herbert, M. A., Ober, C. and Pritchard, J. K. (2004) Evidence for extensive transmission distortion in the human genome. *Am. J. Hum. Genet.*, **74**, 62–72.