Part I

Simon Tavaré: Ancestral Inference in Population Genetics

S. Tavaré and O. Zeitouni: LNM 1837, J. Picard (Ed.), pp. 1–188, 2004.

Ancestral Inference in Population Genetics

Simon Tavaré

Departments of Biological Sciences, Mathematics and Preventive Medicine
University of Southern California.

1 Introduction 6

1.1 Genealogical processes 6
1.2 Organization of the notes 7
1.3 Acknowledgements 8

2 The Wright-Fisher model 9

2.1 Random drift 9
2.2 The genealogy of the Wright-Fisher model 12
2.3 Properties of the ancestral process 19
2.4 Variable population size 23

3 The Ewens Sampling Formula 30

3.1 The effects of mutation 30
3.2 Estimating the mutation rate 32
3.3 Allozyme frequency data 33
3.4 Simulating an infinitely-many alleles sample 34
3.5 A recursion for the ESF 35
3.6 The number of alleles in a sample 37
3.7 Estimating θ 38
3.8 Testing for selective neutrality 41

4 The Coalescent 44

4.1 Who is related to whom? 44
4.2 Genealogical trees 47
4.3 Robustness in the coalescent 47
4.4 Generalizations 52
4.5 Coalescent reviews 53

5 The Infinitely-many-sites Model 54

5.1 Measures of diversity in a sample 56

5.2 Pairwise difference curves 59
5.3 The number of segregating sites 59
5.4 The infinitely-many-sites model and the coalescent 64
5.5 The tree structure of the infinitely-many-sites model 65
5.6 Rooted genealogical trees 67
5.7 Rooted genealogical tree probabilities 68
5.8 Unrooted genealogical trees 71
5.9 Unrooted genealogical tree probabilities 73
5.10 A numerical example 74
5.11 Maximum likelihood estimation 77

6 Estimation in the Infinitely-many-sites Model 79

6.1 Computing likelihoods 79
6.2 Simulating likelihood surfaces 81
6.3 Combining likelihoods 82
6.4 Unrooted tree probabilities 83
6.5 Methods for variable population size models 84
6.6 More on simulating mutation models 86
6.7 Importance sampling 87
6.8 Choosing the weights 90

7 Ancestral Inference in the Infinitely-many-sites Model 94

7.1 Samples of size two 94
7.2 No variability observed in the sample 95
7.3 The rejection method 96
7.4 Conditioning on the number of segregating sites 97
7.5 An importance sampling method 101
7.6 Modeling uncertainty in N and μ 101
7.7 Varying mutation rates 104
7.8 The time to the MRCA of a population given data from a sample . 105
7.9 Using the full data 108

8 The Age of a Unique Event Polymorphism 111

8.1 UEP trees 111
8.2 The distribution of T_Δ 114
8.3 The case $\mu = 0$ 116
8.4 Simulating the age of an allele 118
8.5 Using intra-allelic variability 118

9 Markov Chain Monte Carlo Methods 120

9.1 K-Allele models 121
9.2 A biomolecular sequence model 124
9.3 A recursion for sampling probabilities 125
9.4 Computing probabilities on trees 126
9.5 The MCMC approach 127

9.6 Some alternative updating methods . . . 132
9.7 Variable population size . . . 137
9.8 A Nuu Chah Nulth data set . . . 138
9.9 The age of a UEP . . . 142
9.10 A Yakima data set . . . 145

10 Recombination . . . 151

10.1 The two locus model . . . 151
10.2 The correlation between tree lengths . . . 157
10.3 The continuous recombination model . . . 160
10.4 Mutation in the ARG . . . 163
10.5 Simulating samples . . . 165
10.6 Linkage disequilibrium and haplotype sharing . . . 167

11 ABC: Approximate Bayesian Computation . . . 169

11.1 Rejection methods . . . 169
11.2 Inference in the fossil record . . . 170
11.3 Using summary statistics . . . 175
11.4 MCMC methods . . . 176
11.5 The genealogy of a branching process . . . 177

12 Afterwords . . . 179

12.1 The effects of selection . . . 179
12.2 The combinatorics connection . . . 179
12.3 Bugs and features . . . 180

References . . . 180

1 Introduction

One of the most important challenges facing modern biology is how to make sense of genetic variation. Understanding how genotypic variation translates into phenotypic variation, and how it is structured in populations, is fundamental to our understanding of evolution. Understanding the genetic basis of variation in phenotypes such as disease susceptibility is of great importance to human geneticists. Technological advances in molecular biology are making it possible to survey variation in natural populations on an enormous scale. The most dramatic examples to date are provided by Perlegen Sciences Inc., who resequenced 20 copies of chromosome 21 (Patil *et al.*, 2001) and by Genaissance Pharmaceuticals Inc., who studied haplotype variation and linkage disequilibrium across 313 human genes (Stephens *et al.*, 2001). These are but two of the large number of variation surveys now underway in a number of organisms. The amount of data these studies will generate is staggering, and the development of methods for their analysis and interpretation has become central. In these notes I describe the basics of *coalescent theory*, a useful quantitative tool in this endeavor.

1.1 Genealogical processes

These Saint Flour lectures concern *genealogical processes*, the stochastic models that describe the ancestral relationships among samples of individuals. These individuals might be species, humans or cells – similar methods serve to analyze and understand data on very disparate time scales. The main theme is an account of methods of statistical inference for such processes, based primarily on stochastic computation methods. The notes do not claim to be even-handed or comprehensive; rather, they provide a personal view of some of the theoretical and computational methods that have arisen over the last 20 years. A comprehensive treatment is impossible in a field that is evolving as fast as this one. Nonetheless I think the notes serve as a useful starting point for accessing the extensive literature.

Understanding molecular variation data

The first lecture in the Saint Flour Summer School series reviewed some basic molecular biology and outlined some of the problems faced by computational molecular biologists. This served to place the problems discussed in the remaining lectures into a broader perspective. I have found the books of Hartl and Jones (2001) and Brown (1999) particularly useful.

It is convenient to classify evolutionary problems according to the time scale involved. On long time scales, think about trying to reconstruct the molecular phylogeny of a collection of species using DNA sequence data taken

from a homologous region in each species. Not only is the phylogeny, or branching order, of the species of interest but so too might be estimation of the divergence time between pairs of species, of aspects of the mutation process that gave rise to the observed differences in the sequences, and questions about the nature of the common ancestor of the species. A typical population genetics problem involves the use of patterns of variation observed in a sample of humans to locate disease susceptibility genes. In this example, the time scale is of the order of thousands of years. Another example comes from cancer genetics. In trying to understand the evolution of tumors we might extract a sample of cells, type them for microsatellite variation at a number of loci and then use the observed variability to infer the time since a checkpoint in the tumor's history. The time scale in this example is measured in years.

The common feature that links these examples is the dependence in the data generated by common ancestral history. Understanding the way in which ancestry produces dependence in the sample is the key principle of these notes. Typically the ancestry is never known over the whole time scale involved. To make any progress, the ancestry has to be modelled as a stochastic process. Such processes are the subject of these notes.

Backwards or Forwards?

The theory of population genetics developed in the early years of the last century focused on a *prospective* treatment of genetic variation (see Provine (2001) for example). Given a stochastic or deterministic model for the evolution of gene frequencies that allows for the effects of mutation, random drift, selection, recombination, population subdivision and so on, one can ask questions like 'How long does a new mutant survive in the population?', or 'What is the chance that an allele becomes fixed in the population?'. These questions involve the analysis of the future behavior of a system given initial data. Most of this theory is much easier to think about if the focus is *retrospective*. Rather than ask where the population will go, ask where it has been. This changes the focus to the study of ancestral processes of various sorts. While it might be a truism that genetics is all about ancestral history, this fact has not pervaded the population genetics literature until relatively recently. We shall see that this approach makes most of the underlying methodology easier to derive – essentially all classical prospective results can be derived more simply by this dual approach – and in addition provides methods for analyzing modern genetic data.

1.2 Organization of the notes

The notes begin with forwards and backwards descriptions of the Wright-Fisher model of gene frequency fluctuation in Section 2. The ancestral process that records the number of distinct ancestors of a sample back in time is described, and a number of its basic properties derived. Section 3 introduces

the effects of mutation in the history of a sample, introduces the genealogical approach to simulating samples of genes. The main result is a derivation of the Ewens sampling formula and a discussion of its statistical implications. Section 4 introduces Kingman's coalescent process, and discusses the robustness of this process for different models of reproduction.

Methods more suited to the analysis of DNA sequence data begin in Section 5 with a theoretical discussion of the infinitely-many-sites mutation model. Methods for finding probabilities of the underlying reduced genealogical trees are given. Section 6 describes a computational approach based on importance sampling that can be used for maximum likelihood estimation of population parameters such as mutation rates. Section 7 introduces a number of problems concerning inference about properties of coalescent trees conditional on observed data. The motivating example concerns inference about the time to the most recent common ancestor of a sample. Section 8 develops some theoretical and computational methods for studying the ages of mutations. Section 9 discusses Markov chain Monte Carlo approaches for Bayesian inference based on sequence data. Section 10 introduces Hudson's coalescent process that models the effects of recombination. This section includes a discussion of ancestral recombination graphs and their use in understanding linkage disequilibrium and haplotype sharing.

Section 11 discusses some alternative approaches to inference using approximate Bayesian computation. The examples include two at opposite ends of the evolutionary time scale: inference about the divergence time of primates and inference about the age of a tumor. This section includes a brief introduction to computational methods of inference for samples from a branching process. Section 12 concludes the notes with pointers to some topics discussed in the Saint Flour lectures, but not included in the printed version. This includes models with selection, and the connection between the stochastic structure of certain decomposable combinatorial models and the Ewens sampling formula.

1.3 Acknowledgements

Paul Marjoram, John Molitor, Duncan Thomas, Vincent Plagnol, Darryl Shibata and Oliver Will were involved with aspects of the unpublished research described in Section 11. I thank Lada Markovtsova for permission to use some of the figures from her thesis (Markovtsova (2000)) in Section 9. I thank Magnus Nordborg for numerous discussions about the mysteries of recombination. Above all I thank Warren Ewens and Bob Griffiths, collaborators for over 20 years. Their influence on the statistical development of population genetics has been immense; it is clearly visible in these notes.

Finally I thank Jean Picard for the invitation to speak at the summer school, and the Saint-Flour participants for their comments on the earlier version of the notes.

2 The Wright-Fisher model

This section introduces the Wright-Fisher model for the evolution of gene frequencies in a finite population. It begins with a prospective treatment of a population in which each individual is one of two types, and the effects of mutation, selection, ... are ignored. A genealogical (or retrospective) description follows. A number of properties of the ancestral relationships among a sample of individuals are given, along with a genealogical description in the case of variable population size.

2.1 Random drift

The simplest Wright-Fisher model (Fisher (1922), Wright (1931)) describes the evolution of a two-allele locus in a population of constant size undergoing random mating, ignoring the effects of mutation or selection. This is the so-called 'random drift' model of population genetics, in which the fundamental source of "randomness" is the reproductive mechanism.

A Markov chain model

We assume that the population is of constant size N in each non-overlapping generation n, $n = 0, 1, 2, \ldots$. At the locus in question there are two alleles, denoted by A and B. X_n counts the number of A alleles in generation n. We assume first that there is no mutation between the types. The population at generation $r + 1$ is derived from the population at time r by binomial sampling of N genes from a gene pool in which the fraction of A alleles is its current frequency, namely $\pi_i = i/N$. Hence given $X_r = i$, the probability that $X_{r+1} = j$ is

$$p_{ij} = \binom{N}{j} \pi_i^j (1 - \pi_i)^{N-j}, \quad 0 \le i, j \le N. \tag{2.1.1}$$

The process $\{X_r, r = 0, 1, \ldots\}$ is a time-homogeneous Markov chain. It has transition matrix $P = (p_{ij})$, and state space $\mathcal{S} = \{0, 1, \ldots, N\}$. The states 0 and N are absorbing; if the population contains only one allele in some generation, then it remains so in every subsequent generation. In this case, we say that the population is *fixed* for that allele.

The binomial nature of the transition matrix makes some properties of the process easy to calculate. For example,

$$\mathbb{E}(X_r | X_{r-1}) = N \frac{X_{r-1}}{N} = X_{r-1},$$

so that by averaging over the distribution of X_{r-1} we get $\mathbb{E}(X_r) = \mathbb{E}(X_{r-1})$, and

$$\mathbb{E}(X_r) = \mathbb{E}(X_0), \ r = 1, 2, \ldots. \tag{2.1.2}$$

The result in (2.1.2) can be thought of as the analog of the Hardy-Weinberg law: in an infinitely large random mating population, the relative frequency of the alleles remains constant in every generation. Be warned though that average values in a stochastic process do not tell the whole story! While on average the number of A alleles remains constant, variability must eventually be lost. That is, eventually the population contains all A alleles or all B alleles.

We can calculate the probability a_i that eventually the population contains only A alleles, given that $X_0 = i$. The standard way to find such a probability is to derive a system of equations satisfied by the a_i. To do this, we condition on the value of X_1. Clearly, $a_0 = 0, a_N = 1$, and for $1 \le i \le N-1$, we have

$$a_i = p_{i0} \cdot 0 + p_{iN} \cdot 1 + \sum_{j=1}^{N-1} p_{ij} a_j. \tag{2.1.3}$$

This equation is derived by noting that if $X_1 = j \in \{1, 2, \ldots, N-1\}$, then the probability of reaching N before 0 is a_j. The equation in (2.1.3) can be solved by recalling that $\mathbb{E}(X_1 \mid X_0 = i) = i$, or

$$\sum_{j=0}^{N} p_{ij} j = i.$$

It follows that $a_i = Ci$ for some constant C. Since $a_N = 1$, we have $C = 1/N$, and so $a_i = i/N$. Thus the probability that an allele will fix in the population is just its initial frequency.

The variance of X_r can also be calculated from the fact that

$$\mathrm{Var}(X_r) = \mathbb{E}(\mathrm{Var}(X_r|X_{r-1})) + \mathrm{Var}(\mathbb{E}(X_r|X_{r-1})).$$

After some algebra, this leads to

$$\mathrm{Var}(X_r) = \mathbb{E}(X_0)(N - \mathbb{E}(X_0))(1 - \lambda^r) + \lambda^r \mathrm{Var}(X_0), \tag{2.1.4}$$

where

$$\lambda = 1 - 1/N.$$

We have noted that genetic variability in the population is eventually lost. It is of some interest to assess how fast this loss occurs. A simple calculation shows that

$$\mathbb{E}(X_r(N - X_r)) = \lambda^r \mathbb{E}(X_0(N - X_0)). \tag{2.1.5}$$

Multiplying both sides by $2N^{-2}$ shows that the probability $h(r)$ that two genes chosen at random with replacement in generation r are different is

$$h(r) = \lambda^r h(0). \tag{2.1.6}$$

The quantity $h(r)$ is called the *heterozygosity* of the population in generation r, and it measures the genetic variability surviving in the population. Equation

(2.1.6) shows that the heterozygosity decays geometrically quickly as $r \to \infty$. Since fixation must occur, we have $h(r) \to 0$.

We have seen that variability is lost from the population. How long does this take? First we find an equation satisfied by m_i, the mean time to fixation starting from $X_0 = i$. To do this, notice first that $m_0 = m_N = 0$, and, by conditioning on the first step once more, we see that for $1 \le i \le N-1$

$$\begin{aligned} m_i &= p_{i0} \cdot 1 + p_{iN} \cdot 1 + \sum_{j=1}^{N-1} p_{ij}(1 + m_j) \\ &= 1 + \sum_{j=0}^{N} p_{ij} m_j. \end{aligned} \tag{2.1.7}$$

Finding an explicit expression for m_i is difficult, and we resort instead to an approximation when N is large and time is measured in units of N generations.

Diffusion approximations

This takes us into the world of diffusion theory. It is usual to consider not the total number $X_r \equiv X(r)$ of A alleles but rather the proportion X_r/N. To get a non-degenerate limit we must also rescale time, in units of N generations. This leads us to study the rescaled process

$$Y_N(t) = N^{-1} X(\lfloor Nt \rfloor), \quad t \ge 0, \tag{2.1.8}$$

where $\lfloor x \rfloor$ is the integer part of x. The idea is that as $N \to \infty$, $Y_N(\cdot)$ should converge in distribution to a process $Y(\cdot)$. The fraction $Y(t)$ of A alleles at time t evolves like a continuous-time, continuous state-space process in the interval $\mathcal{S} = [0, 1]$. $Y(\cdot)$ is an example of a diffusion process. Time scalings in units proportional to N generations are typical for population genetics models appearing in these notes.

Diffusion theory is the basic tool of classical population genetics, and there are several good references. Crow and Kimura (1970) has a lot of the 'old style' references to the theory. Ewens (1979) and Kingman (1980) introduce the sampling theory ideas. Diffusions are also discussed by Karlin and Taylor (1980) and Ethier and Kurtz (1986), the latter in the measure-valued setting. A useful modern reference is Neuhauser (2001).

The properties of a one-dimensional diffusion $Y(\cdot)$ are essentially determined by the infinitesimal mean and variance, defined in the time-homogeneous case by

$$\begin{aligned} \mu(y) &= \lim_{h \to 0} h^{-1} \mathbb{E}(Y(t+h) - Y(t) \mid Y(t) = y), \\ \sigma^2(y) &= \lim_{h \to 0} h^{-1} \mathbb{E}((Y(t+h) - Y(t))^2 \mid Y(t) = y). \end{aligned}$$

For the discrete Wright-Fisher model, we know that given $X_r = i$, X_{r+1} is binomially distributed with number of trials N and success probability i/N. Hence

$$\mathbb{E}(X(r+1)/N - X(r)/N \mid X(r)/N = i/N) = 0,$$
$$\mathbb{E}((X(r+1)/N - X(r)/N)^2 \mid X(r)/N = i/N) = \frac{1}{N}\frac{i}{N}\left(1 - \frac{i}{N}\right),$$

so that for the process $Y(\cdot)$ that gives the proportion of allele A in the population at time t, we have

$$\mu(y) = 0, \qquad \sigma^2(y) = y(1-y),\ 0 < y < 1. \tag{2.1.9}$$

Classical diffusion theory shows that the mean time $m(x)$ to fixation, starting from an initial fraction $x \in (0,1)$ of the A allele, satisfies the differential equation

$$\frac{1}{2}x(1-x)m''(x) = -1, \quad m(0) = m(1) = 0. \tag{2.1.10}$$

This equation, the analog of (2.1.7), can be solved using partial fractions, and we find that

$$m(x) = -2(x\log x + (1-x)\log(1-x)), \quad 0 < x < 1. \tag{2.1.11}$$

In terms of the underlying discrete model, the approximation for the expected number m_i of generations to fixation, starting from i A alleles, is $m_i \approx Nm(i/N)$. If $i/N = 1/2$,

$$Nm(1/2) = (-2\log 2)N \approx 1.39N \text{ generations},$$

whereas if the A allele is introduced at frequency $1/N$,

$$Nm(1/N) = 2\log N \text{ generations}.$$

2.2 The genealogy of the Wright-Fisher model

In this section we consider the Wright-Fisher model from a genealogical perspective. In the absence of recombination, the DNA sequence representing the gene of interest is a copy of a sequence in the previous generation, that sequence is itself a copy of a sequence in the generation before that and so on. Thus we can think of the DNA sequence as an 'individual' that has a 'parent' (namely the sequence from which is was copied), and a number of 'offspring' (namely the sequences that originate as a copy of it in the next generation).

To study this process either forwards or backwards in time, it is convenient to label the individuals in a given generation as $1, 2, \ldots, N$, and let ν_i denote the number of offspring born to individual i, $1 \le i \le N$. We suppose that individuals have independent Poisson-distributed numbers of offspring,

subject to the requirement that the total number of offspring is N. It follows that $(\nu_1, \ldots, \nu_N)$ has a symmetric multinomial distribution, with

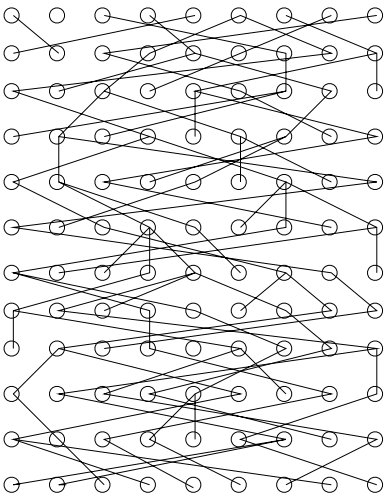
$$\mathbb{P}(\nu_1 = m_1, \ldots, \nu_N = m_N) = \frac{N!}{m_1! \cdots m_N!} \left(\frac{1}{N}\right)^N \tag{2.2.1}$$

provided $m_1 + \cdots + m_N = N$. We assume that offspring numbers are independent from generation to generation, with distribution specified by (2.2.1).

To see the connection with the earlier description of the Wright-Fisher model, imagine that each individual in a given generation carries either an A allele or a B allele, i of the N individuals being labelled A. Since there is no mutation, all offspring of type A individuals are also of type A. The distribution of the number of type A in the offspring therefore has the distribution of $\nu_1 + \cdots + \nu_i$ which (from elementary properties of the multinomial distribution) has the binomial distribution with parameters N and success probability $p = i/N$. Thus the number of A alleles in the population does indeed evolve according to the Wright-Fisher model described in (2.1.1).

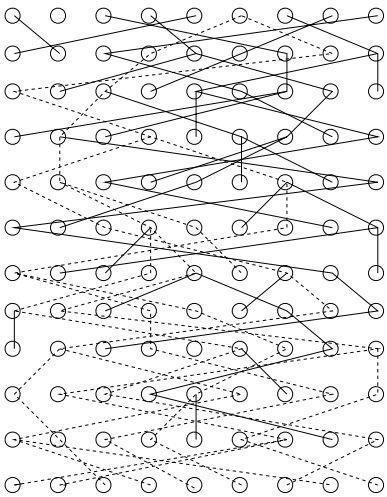
This specification shows how to simulate the offspring process from parents to children to grandchildren and so on. A realization of such a process for $N = 9$ is shown in Figure 2.1. Examination of Figure 2.1 shows that individuals 3 and 4 have their most recent common ancestor (MRCA) 3 generations ago, whereas individuals 2 and 3 have their MRCA 11 generations ago. More

Fig. 2.1. Simulation of a Wright-Fisher model of $N = 9$ individuals. Generations are evolving down the figure. The individuals in the last generation should be labelled 1,2,. . . ,9 from left to right. Lines join individuals in two generations if one is the offspring of the other



generally, for any population size N and sample of size n taken from the present generation, what is the structure of the ancestral relationships linking the members of the sample? The crucial observation is that if we view the process from the present generation back into the past, then individuals choose their parents independently and at random from the individuals in the previous generation, and successive choices are independent from generation to generation. Of course, not all members of the previous generations are ancestors of individuals in the present-day sample. In Figure 2.2 the ancestry of those individuals who are ancestral to the sample is highlighted with broken lines, and in Figure 2.3 those lineages that are not connected to the sample are removed, the resulting figure showing just the successful ancestors. Finally, Figure 2.3 is untangled in Figure 2.4. This last figure shows the tree-like nature of the genealogy of the sample.

Fig. 2.2. Simulation of a Wright-Fisher model of $N = 9$ individuals. Lines indicate ancestors of the sampled individuals. Individuals in the last generation should be labelled 1,2,. . . , 9 from left to right. Dashed lines highlight ancestry of the sample.



Understanding the genealogical process provides a direct way to study gene frequencies in a model with no mutation (Felsenstein (1971)). We content ourselves with a genealogical derivation of (2.1.6). To do this, we ask how long it takes for a sample of two genes to have their first common ancestor. Since individuals choose their parents at random, we see that

$$\mathbb{P}(\text{ 2 individuals have 2 distinct parents}) = \lambda = \left(1 - \frac{1}{N}\right).$$

Fig. 2.3. Simulation of a Wright-Fisher model of $N = 9$ individuals. Individuals in the last generation should be labelled 1,2,. . . , 9 from left to right. Dashed lines highlight ancestry of the sample. Ancestral lineages not ancestral to the sample are removed.

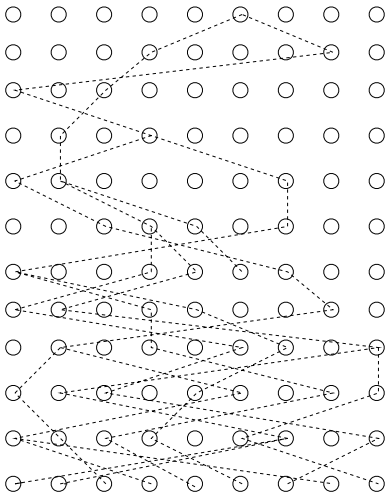
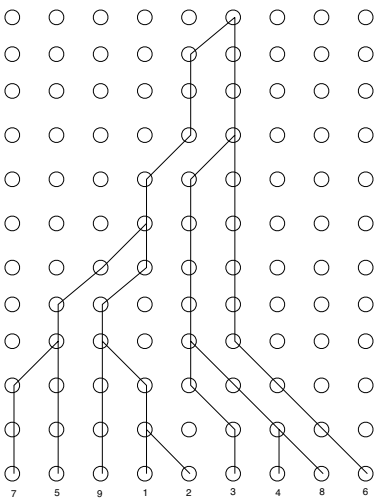


Fig. 2.4. Simulation of a Wright-Fisher model of $N = 9$ individuals. This is an untangled version of Figure 2.3.



Since those parents are themselves a random sample from their generation, we may iterate this argument to see that

$$\mathbb{P}(\text{First common ancestor more than } r \text{ generations ago})$$
$$= \lambda^r = \left(1 - \frac{1}{N}\right)^r. \tag{2.2.2}$$

Now consider the probability $h(r)$ that two individuals chosen with replacement from generation r carry distinct alleles. Clearly if we happen to choose the same individual twice (probability $1/N$) this probability is 0. In the other case, the two individuals are different if and only if their common ancestor is more than r generations ago, and the ancestors at time 0 are distinct. The probability of this latter event is the chance that 2 individuals chosen without replacement at time 0 carry different alleles, and this is just $\mathbb{E}2X_0(N - X_0)/N(N-1)$. Combining these results gives

$$h(r) = \lambda^r \, \frac{(N-1)}{N} \, \frac{\mathbb{E}2X_0(N - X_0)}{N(N-1)} = \lambda^r h(0),$$

just as in (2.1.6).

When the population size is large and time is measured in units of N generations, the distribution of the time to the MRCA of a sample of size 2 has approximately an exponential distribution with mean 1. To see this, rescale time so that $r = Nt$, and let $N \to \infty$ in (2.2.2). We see that this probability is

$$\left(1 - \frac{1}{N}\right)^{Nt} \;\to\; e^{-t}.$$

This time scaling is the same as used to derive the diffusion approximation earlier. This should be expected, as the forward and backward approaches are just alternative views of the same underlying process.

The ancestral process in a large population

What can be said about the number of ancestors in larger samples? The probability that a sample of size three has distinct parents is

$$\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)$$

and the iterative argument above can be applied once more to see that the sample has three distinct ancestors for more than r generations with probability

$$\left[\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\right]^r = \left(1 - \frac{3}{N} + \frac{2}{N^2}\right)^r.$$

Rescaling time once more in units of N generations, and taking $r = Nt$, shows that for large N this probability is approximately e^{-3t}, so that on the new time scale the time taken to find the first common ancestor in the sample of three genes is exponential with parameter 3. What happens when a common ancestor is found? Note that the chance that three distinct individuals have at most two distinct parents is

$$\frac{3(N-1)}{N^2} + \frac{1}{N^2} = \frac{3N-2}{N^2}.$$

Hence, given that a first common ancestor is found in generation r, the conditional probability that the sample has two distinct ancestors in generation r is

$$\frac{3N-3}{3N-2},$$

which tends to 1 as N increases. Thus in our approximating process the number of distinct ancestors drops by precisely 1 when a common ancestor is found.

We can summarize the discussion so far by noting that in our approximating process a sample of three genes waits an exponential amount of time T_3 with parameter 3 until a common ancestor is found, at which point the sample has two distinct ancestors for a further amount of time T_2 having an exponential distribution with parameter 1. Furthermore, T_3 and T_2 are independent random variables.

More generally, the number of distinct parents of a sample of size k individuals can be thought of as the number of occupied cells after k balls have been dropped (uniformly and independently) into N cells. Thus

$$\begin{aligned} g_{kj} &\equiv \mathbb{P}(k \text{ individuals have } j \text{ distinct parents}) \\ &= N(N-1)\cdots(N-j+1)\mathcal{S}_k^{(j)} N^{-k} \qquad j = 1, 2, \ldots, k \end{aligned} \tag{2.2.3}$$

where $\mathcal{S}_k^{(j)}$ is a Stirling number of the second kind; that is, $\mathcal{S}_k^{(j)}$ is the number of ways of partitioning a set of k elements into j nonempty subsets. The terms in (2.2.3) arise as follows: $N(N-1)\cdots(N-j+1)$ is the number of ways to choose j distinct parents; $\mathcal{S}_k^{(j)}$ is the number of ways assigning k individuals to these j parents; and N^k is the total number of ways of assigning k individuals to their parents.

For fixed values of N, the behavior of this ancestral process is difficult to study analytically, but we shall see that the simple approximation derived above for samples of size two and three can be developed for any sample size n. We first define an ancestral process $\{A_n^N(t) : t = 0, 1, \ldots\}$ where

$$A_n^N(t) \equiv \text{number of distinct ancestors in generation } t \text{ of a sample of size } n \text{ at time } 0.$$

It is evident that $A_n^N(\cdot)$ is a Markov chain with state space $\{1, 2, \ldots, n\}$, and with transition probabilities given by (2.2.3):

$$\mathbb{P}(A_n^N(t+1) = j | A_n^N(t) = k) = g_{kj}.$$

For fixed sample size n, as $N \to \infty$,

$$\begin{aligned} g_{k,k-1} &= \mathcal{S}_k^{(k-1)} \frac{N(N-1)\cdots(N-k+2)}{N^k} \\ &= \binom{k}{2}\frac{1}{N} + O(N^{-2}), \end{aligned}$$

since $\mathcal{S}_k^{(k-1)} = \binom{k}{2}$. For $j < k-1$, we have

$$g_{k,j} = \mathcal{S}_k^{(j)} \frac{N(N-1)\cdots(N-j+1)}{N^k} = O(N^{-2})$$

and

$$\begin{aligned} g_{k,k} &= N^{-k}\, N(N-1)\cdots(N-k+1) \\ &= 1 - \binom{k}{2}\frac{1}{N} + O(N^{-2}). \end{aligned}$$

Writing G_N for the transition matrix with elements $g_{kj}, 1 \le j \le k \le n$. Then

$$G_N = I + N^{-1}Q + O(N^{-2}),$$

where I is the identity matrix, and Q is a lower diagonal matrix with non-zero entries given by

$$q_{kk} = -\binom{k}{2}, \quad q_{k,k-1} = \binom{k}{2}, \ k = n, n-1, \ldots, 2. \tag{2.2.4}$$

Hence with time rescaled for units of N generations, we see that

$$G_N^{Nt} = \left(I + N^{-1}Q + O(N^{-2})\right)^{Nt} \to e^{Qt}$$

as $N \to \infty$. Thus the number of distinct ancestors in generation Nt is approximated by a Markov chain $A_n(t)$ whose behavior is determined by the matrix Q in (2.2.4). $A_n(\cdot)$ is a pure death process that starts from $A_n(0) = n$, and decreases by jumps of size one only. The waiting time T_k in state k is exponential with parameter $\binom{k}{2}$, the T_k being independent for different k.

Remark. We call the process $A_n(t), t \ge 0$ the *ancestral process* for a sample of size n.

Remark. The ancestral process of the Wright-Fisher model has been studied in several papers, including Karlin and McGregor (1972), Cannings (1974), Watterson (1975), Griffiths (1980), Kingman (1980) and Tavaré (1984).

2.3 Properties of the ancestral process

Calculation of the distribution of $A_n(t)$ is an elementary exercise in Markov chains. One way to do this is to diagonalize the matrix Q by writing $Q = RDL$, where D is the diagonal matrix of eigenvalues $\lambda_k = -\binom{k}{2}$ of Q, and R and L are matrices of right and left eigenvalues of Q, normalized so that $RL = LR = I$. From this approach we get, for $j = 1, 2, \ldots, n$,

$$\begin{aligned} g_{nj}(t) &\equiv \mathbb{P}(A_n(t) = j) \\ &= \sum_{k=j}^{n} e^{-k(k-1)t/2} \frac{(2k-1)(-1)^{k-j} j_{(k-1)} n_{[k]}}{j!(k-j)! n_{(k)}} \end{aligned} \tag{2.3.1}$$

where

$$\begin{aligned} a_{(n)} &= a(a+1)\cdots(a+n-1) \\ a_{[n]} &= a(a-1)\cdots(a-n+1) \\ a_{(0)} &= a_{[0]} = 1. \end{aligned}$$

The mean number of ancestors at time t is given by

$$\mathbb{E}A_n(t) = \sum_{k=1}^{n} e^{-k(k-1)t/2} \frac{(2k-1) n_{[k]}}{n_{(k)}}, \tag{2.3.2}$$

and the falling factorial moments are given by

$$\mathbb{E}(A_n(t))_{[r]} = \sum_{k=r}^{n} \frac{n_{[k]}}{n_{(k)}} e^{-k(k-1)t/2} (2k-1) \frac{(r+k-2)!}{(r-1)!(k-r)!},$$

for $r = 2, \ldots, n$. In Figure 2.5 $\mathbb{E}A_n(t)$ is plotted as a function of t for $n = 5, 10, 20, 50$.

The process $A_n(\cdot)$ is eventually absorbed at 1, when the sample is traced back to its most recent common ancestor (MRCA). The time it takes the sample to reach its MRCA is of some interest to population geneticists. We study this time in the following section.

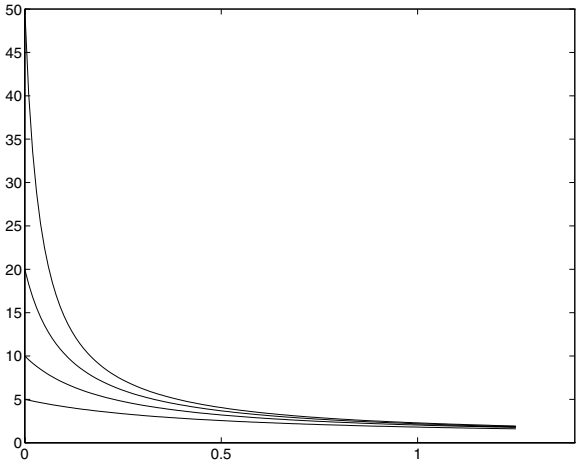
The time to the most recent common ancestor

Many quantities of genetic interest depend on the time W_n taken to trace a sample of size n back to its MRCA. Remember that time here is measured in units of N generations, and that

$$W_n = T_n + T_{n-1} + \cdots + T_2 \tag{2.3.3}$$

where T_k are independent exponential random variables with parameter $\binom{k}{2}$. It follows that

Fig. 2.5. The mean number of ancestors at time t (x axis) for samples of size $n = 5, 10, 20, 50$, from (2.3.2).



$$\mathbb{E}W_n = \sum_{k=2}^{n} \mathbb{E}T_k = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k} \right) = 2\left(1 - \frac{1}{n}\right).$$

Therefore

$$1 = \mathbb{E}W_2 \leq \mathbb{E}W_n \leq \mathbb{E}W_N < 2,$$

where W_N is thought of as the time until the whole population has a single common ancestor. Note that $\mathbb{E}W_n$ is close to 2 even for moderate n. Also

$$\mathbb{E}(W_N - W_n) = 2\left(\frac{1}{n} - \frac{1}{N}\right) < \frac{2}{n}$$

so the mean difference between the time for a sample to reach its MRCA, and the time for the whole population to reach its MRCA, is small.

Note that T_2 makes a substantial contribution to the sum (2.3.3) defining W_n. For example, on average for over half the time since its MRCA, the sample will have exactly two ancestors. Further, using the independence of the T_k,

$$\begin{aligned} \mathrm{Var}W_n &= \sum_{k=2}^{n} \mathrm{Var}T_k = \sum_{k=2}^{n} \binom{k}{2}^{-2} \\ &= 8\sum_{k=1}^{n-1} \frac{1}{k^2} - 4\left(1 - \frac{1}{n}\right)\left(3 + \frac{1}{n}\right) \end{aligned}$$

It follows that

$$1 = \mathrm{Var}W_2 \leq \mathrm{Var}W_n \leq \lim_{n\to\infty} \mathrm{Var}W_n = 8\frac{\pi^2}{6} - 12 \approx 1.16.$$

We see that T_2 also contributes most to the variance.

The distribution of W_n can be obtained from (2.3.1):

$$\mathbb{P}(W_n \leq t) = \mathbb{P}(A_n(t) = 1) = \sum_{k=1}^{n} e^{-k(k-1)t/2} \frac{(2k-1)(-1)^{k-1} n_{[k]}}{n_{(k)}}. \quad (2.3.4)$$

From this it follows that

$$\mathbb{P}(W_n > t) = 3\frac{n-1}{n+1}e^{-t} + O(e^{-3t}) \text{ as } t \to \infty.$$

Now focus on two particular individuals in the sample and observe that if these two individuals do not have a common ancestor at t, the whole sample cannot have a common ancestor. Since the two individuals are themselves a random sample of size two from the population, we see that

$$\mathbb{P}(W_n > t) \geq \mathbb{P}(W_2 > t) = e^{-t},$$

an inequality that also follows from (2.3.3). A simple Markov chain argument shows that

$$\mathbb{P}(W_n > t) \leq \frac{3(n-1)e^{-t}}{n+1},$$

so that

$$e^{-t} \leq \mathbb{P}(W_n > t) \leq 3e^{-t}$$

for all n and t (see Kingman (1980), (1982c)).

The density function of W_n follows immediately from (2.3.4) by differentiating with respect to t:

$$f_{W_n}(t) = \sum_{k=2}^{n} (-1)^k e^{-k(k-1)t/2} \frac{(2k-1)k(k-1)n_{[k]}}{2n_{(k)}}. \quad (2.3.5)$$

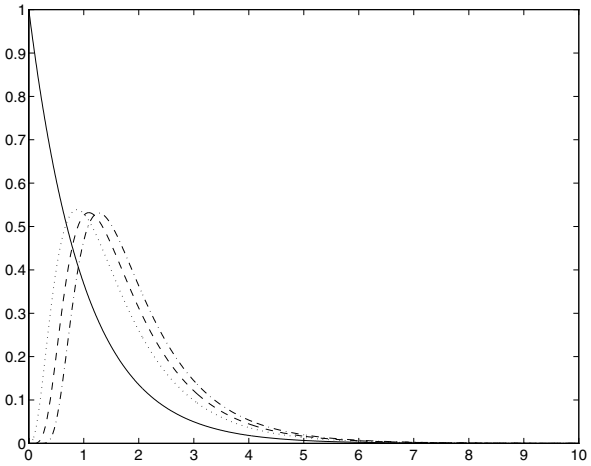
In Figure 2.6, this density is plotted for values of $n = 2, 10, 100, 500$. The shape of the densities reflects the fact that most of the contribution to the density comes from T_2.

The tree length

In contrast to the distribution of W_n, the distribution of the total length $L_n = 2T_2 + \cdots + nT_n$ is easy to find. As we will see, L_n is the total length of the branches in the genealogical tree linking the individuals in the sample. First of all,

$$\mathbb{E}L_n = 2\sum_{j=1}^{n-1} \frac{1}{j} \sim 2\log n,$$

Fig. 2.6. Density functions for the time W_n to most recent common ancestor of a sample of n individuals, from (2.3.5). – $n = 2$; $\cdots\cdots$ $n = 10$; $- - - -$ $n = 100$; $- \cdot - \cdot$ $n = 500$.



and

$$\mathrm{Var}L_n = 4\sum_{j=1}^{n-1}\frac{1}{j^2} \sim 2\pi^2/3.$$

To find the distribution of L_n, let $E(\lambda)$ denote an exponential random variable with mean $1/\lambda$, all occurrences being independent of each other, and write $=_d$ for equality in distribution. Then

$$\begin{aligned} L_n &= \sum_{j=2}^{n} jT_j =_d \sum_{j=2}^{n} E((j-1)/2) \\ &=_d \sum_{j=1}^{n-1} \min_{1\le k\le j} E_{jk}(1/2) \\ &=_d \max_{1\le j\le n-1} E_j(1/2), \end{aligned}$$

the last step following by a coupling argument (this is one of many proofs of Feller's representation of the distribution of the maximum of independent and identically distributed exponential random variables as a sum of independent random variables). Thus

$$\mathbb{P}(L_n \le t) = \left(1 - e^{-t/2}\right)^{n-1}, \ t \ge 0.$$

It follows directly that $L_n - 2\log n$ has a limiting extreme value distribution with distribution function $\exp(-\exp(-t/2)), -\infty < t < \infty$.

2.4 Variable population size

In this section we discuss the behavior of the ancestral process in the case of deterministic fluctuations in population size. For convenience, suppose the model evolves in discrete generations and label the current generation as 0. Denote by $N(j)$ the number of sequences in the population j generations before the present. We assume that the variation in population size is due to either external constraints *e.g.* changes in the environment, or random variation which depends only on the total population size *e.g.* if the population grows as a branching process. This excludes so-called density dependent cases in which the variation depends on the genetic composition of the population, but covers many other settings. We continue to assume neutrality and random mating.

Here we develop the theory for a particular class of population growth models in which, roughly speaking, all the population sizes are large. Time will be scaled in units of $N \equiv N(0)$ generations. To this end, define the relative size function $f_N(x)$ by

$$\begin{aligned} f_N(x) &= \frac{N(\lceil Nx \rceil)}{N} \\ &= \frac{N(j)}{N}, \quad \frac{j-1}{N} < x \leq \frac{j}{N},\ j = 1, 2, \ldots \end{aligned} \tag{2.4.1}$$

We are interested in the behavior of the process when the size of each generation is large, so we suppose that

$$\lim_{N\to\infty} f_N(x) = f(x) \tag{2.4.2}$$

exists and is strictly positive for all $x \geq 0$.

Many demographic scenarios can be modelled in this way. For an example of geometric population growth, suppose that for some constant $\rho > 0$

$$N(j) = \lfloor N\,(1 - \rho/N)^j \rfloor.$$

Then

$$\lim_{N\to\infty} f_N(x) = e^{-\rho x} \equiv f(x),\ x > 0.$$

A commonly used model is one in which the population has constant size prior to generation V, and geometric growth from then to the present time. Thus for some $\alpha \in (0, 1)$

$$N(j) = \begin{cases} \lfloor N\alpha \rfloor, & j \geq V \\ \lfloor N\alpha^{j/V} \rfloor, & j = 0, \ldots, V \end{cases}$$

If we suppose that $V = \lfloor Nv \rfloor$ for some $v > 0$, so that the expansion started v time units ago, then

$$f_N(x) \to f(x) = \alpha^{\min(x/v,1)}.$$

The ancestral process

In a Wright-Fisher model of reproduction, note that the probability that two individuals chosen at time 0 have distinct ancestors s generations ago is

$$\mathbb{P}(T_2(N) > s) = \prod_{j=1}^{s} \left(1 - \frac{1}{N(j)}\right),$$

where $T_2(N)$ denotes the time to the common ancestor of the two individuals. Recalling the inequality

$$x \le -\log(1-x) \le \frac{x}{1-x}, \ x < 1,$$

we see that

$$\sum_{j=1}^{s} \frac{1}{N(j)} \le -\sum_{j=1}^{s} \log\left(1 - \frac{1}{N(j)}\right) \le \sum_{j=1}^{s} \frac{1}{N(j)-1}.$$

It follows that

$$\lim_{N\to\infty} -\sum_{j=1}^{\lfloor Nt \rfloor} \log\left(1 - \frac{1}{N(j)}\right) = \lim_{N\to\infty} \sum_{j=1}^{\lfloor Nt \rfloor} \frac{1}{N(j)}.$$

Since

$$\sum_{j=1}^{s} \frac{1}{N(j)} = \int_0^{s/N} \frac{dx}{f_N(x)},$$

we can use (2.4.2) to see that for $t > 0$, with time rescaled in units of N generations,

$$\lim_{N\to\infty} \mathbb{P}(T_2(N) > \lfloor Nt \rfloor) = \exp\left(-\int_0^t \lambda(u)du\right),$$

where $\lambda(\cdot)$ is the intensity function defined by

$$\lambda(u) = \frac{1}{f(u)}, \quad u \ge 0. \tag{2.4.3}$$

If we define

$$\Lambda(t) = \int_0^t \lambda(u)du,$$

the integrated intensity function, then (2.4.2) shows that as $N \to \infty$

$$N^{-1}T_2(N) \Rightarrow T_2,$$

where

$$\mathbb{P}(T_2 > t) = \exp(-\Lambda(t)),\ t \geq 0. \tag{2.4.4}$$

We expect the two individuals to have a common ancestor with probability one, this corresponding to the requirement that

$$\lim_{t\to\infty} \Lambda(t) = \infty,$$

which we assume from now on. When the population size is constant, $\Lambda(t) = t$ and the time to the MRCA has an exponential distribution with mean 1. From (2.4.4) we see that

$$\mathbb{E}T_2 = \int_0^\infty \mathbb{P}(T_2 > t)dt = \int_0^\infty e^{-\Lambda(t)}dt.$$

If the population has been expanding, so that $f(t) \leq 1$ for all t, then $\Lambda(t) \geq t$, and therefore

$$\mathbb{P}(T_2 > t) \leq \mathbb{P}(T_2^c > t),\quad t \geq 0,$$

where T_2^c denotes the corresponding time in the constant population size case. We say that T_2^c is *stochastically larger* than T_2, so that in particular $\mathbb{E}T_2 \leq \mathbb{E}T_2^c = 1$. This corresponds to the fact that if the population size has been shrinking into the past, it should be possible to find the MRCA sooner than if the population size had been constant.

In the varying environment setting, the ancestral process satisfies

$$\begin{aligned}\mathbb{P}(A_2(t+s) = 1|A_2(t) = 2) &= \mathbb{P}(T_2 \leq t+s|T_2 > t)\\ &= \mathbb{P}(t < T_2 \leq t+s)/\mathbb{P}(T_2 > t)\\ &= 1 - \exp(-(\Lambda(t+s) - \Lambda(t))),\end{aligned}$$

so that

$$\mathbb{P}(A_2(t+h) = 1|A_2(t) = 2) = \lambda(t)h + o(h),\quad h \downarrow 0.$$

We see that $A_2(\cdot)$ is a non-homogeneous Markov chain. What is the structure of $A_n(\cdot)$?

Define $T_k(N)$ to be the number of generations for which the sample has k distinct ancestors. In the event that the sample never has exactly k distinct ancestors, define $T_k(N) = \infty$. We calculate first the joint distribution of $T_3(N)$ and $T_2(N)$. The probability that $T_3(N) = k, T_2(N) = l$ is the probability that the sample of size 3 has 3 distinct ancestors in generations 1, 2, ..., $k-1$, 2 distinct ancestors in generations $k, \ldots, k+l-1$, and 1 in generation $l+k$. The probability that a sample of three individuals taken in generation

$j-1$ has three distinct parents is $N(j)(N(j)-1)(N(j)-2)/N(j)^3$, and the probability that three individuals in generation $k-1$ have two distinct parents is $3N(k)(N(k)-1)/N(k)^3$. Hence

$$\begin{aligned}&\mathbb{P}(T_3(N)=k, T_2(N)=l)\\ &=\left\{\prod_{j=1}^{k-1}\frac{(N(j)-1)(N(j)-2)}{N(j)^3}\right\}\frac{3(N(k)-1)}{N(k)^2}\left\{\prod_{j=k+1}^{k+l-1}\frac{N(j)-1}{N(j)}\right\}\frac{1}{N(k+l)}.\end{aligned}$$

For the size fluctuations we are considering, the first term in brackets is

$$\prod_{j=1}^{k-1}\left(1-\frac{3}{N(j)}+\frac{2}{N(j)^2}\right)\sim\exp\left(-3\int_0^{k/N}\frac{dx}{f_N(x)}\right),$$

while the second term in brackets is

$$\prod_{j=k+1}^{k+l-1}\left(1-\frac{1}{N(j)}\right)\sim\exp\left(-\int_{k/N}^{(k+l)/N}\frac{dx}{f_N(x)}\right).$$

For $k\sim Nt_3, l\sim Nt_2$ with $t_3>0, t_2>0$, we see via (2.4.2) that $N^2\mathbb{P}(T_3(N)=k,T_2(N)=l)$ converges to

$$f(t_3,t_2):=e^{-3\Lambda(t_3)}3\lambda(t_3)e^{-(\Lambda(t_2+t_3)-\Lambda(t_3))}\lambda(t_3+t_2) \tag{2.4.5}$$

as $N\to\infty$. It follows that

$$N^{-1}(T_3(N),T_2(N))\Rightarrow(T_3,T_2),$$

where (T_3,T_2) have joint probability density $f(t_3,t_2)$ given in (2.4.5).

This gives the joint law of the times spent with different numbers of ancestors, and shows that in the limit the number of ancestors decreases by one at each jump. Just as in the constant population-size case, the ancestral process for the Wright-Fisher model is itself a Markov chain, since the distribution of the number of distinct ancestors in generation r is determined just by the number in generation $r-1$. The Markov property is inherited in the limit, and we conclude that $\{A_3(t), t\geq 0\}$ is a Markov chain on the set $\{3,2,1\}$. Its transition intensities can be calculated as a limit from the Wright-Fisher model. We see that

$$\mathbb{P}(A_3(t+h)=j|A_3(t)=i)=\begin{cases}\binom{i}{2}\lambda(t)h+o(h), & j=i-1\\ 1-\binom{i}{2}\lambda(t)h+o(h), & j=i\\ 0, & \text{otherwise}\end{cases}$$

We can now establish the general case in a similar way. The random variables $T_n(N),\ldots,T_2(N)$ have a joint limit law when rescaled:

$$N^{-1}(T_n(N),\ldots,T_2(N))\Rightarrow(T_n,\ldots,T_2)$$

for each fixed n as $N \to \infty$, and the joint density $f(t_n, \ldots, t_2)$ of $T_n, \ldots, T_2$ is given by

$$f(t_n, \ldots, t_2) = \prod_{j=2}^{n} \binom{j}{2} \lambda(s_j) \exp\left\{ -\binom{j}{2} (\Lambda(s_j) - \Lambda(s_{j+1})) \right\}, \qquad (2.4.6)$$

for $0 \leq t_n, \ldots, t_2 < \infty$, where $s_{n+1} = 0, s_n = t_n, s_j = t_j + \cdots + t_n$, $j = 2, \ldots, n-1$.

Remark. The joint density in (2.4.6) should really be denoted by $f_n(t_n, \ldots, t_2)$, and the limiting random variables $T_{nn}, \ldots, T_{n2}$, but we keep the simpler notation. This should not cause any confusion.

From this it is elementary to show that if $S_j \equiv T_n + \cdots + T_j$, then the joint density of $(S_n, \ldots, S_2)$ is given by

$$g(s_n, \ldots, s_2) = \prod_{j=2}^{n} \binom{j}{2} \lambda(s_j) \exp\left(-\binom{j}{2} (\Lambda(s_j) - \Lambda(s_{j+1})) \right),$$

for $0 \leq s_n < s_{n-1} < \cdots < s_2$. This parlays immediately into the distribution of the time the sample spends with j distinct ancestors, given that $S_{j+1} = s$:

$$\mathbb{P}(T_j > t | S_{j+1} = s) = \exp\left(-\binom{j}{2} (\Lambda(s+t) - \Lambda(s)) \right).$$

Note that the sequence $S_{n+1} := 0, S_n, S_{n-1}, \ldots, S_2$ is a Markov chain. The approximating ancestral process $\{A_n(t), t \geq 0\}$ is a non-homogeneous pure death process on $[n]$ with $A_n(0) = n$ whose transition rates are determined by

$$\mathbb{P}(A_n(t+h) = j | A_n(t) = i) = \begin{cases} \binom{i}{2}\lambda(t)h + o(h), & j = i-1 \\ 1 - \binom{i}{2}\lambda(t)h + o(h), & j = i \\ 0, & \text{otherwise} \end{cases} \qquad (2.4.7)$$

The time change representation

Denote the process that counts the number of ancestors at time t of a sample of size n taken at time 0 by $\{A_n^v(t), t \geq 0\}$, the superscript ^{v} denoting variable population size. We have seen that $A_n^v(\cdot)$ is now a time-inhomogeneous Markov process. Given that $A_n^v(t) = j$, it jumps to $j-1$ at rate $j(j-1)\lambda(t)/2$. A useful way to think of the process $A_n^v(\cdot)$ is to notice that a realization may be constructed via

$$A_n^v(t) = A_n(\Lambda(t)), \ t \geq 0, \qquad (2.4.8)$$

where $A_n(\cdot)$ is the corresponding ancestral process for the constant population size case. This may be verified immediately from (2.4.7). We see that the variable population size model is just a deterministic time change of the constant

population size model. Some of the properties of $A_n^v(\cdot)$ follow immediately from this representation. For example,

$$\mathbb{P}(A_n^v(t) = j) = g_{nj}(\Lambda(t)),\ j = 1, \ldots, n$$

where $g_{nj}(t)$ is given in (2.3.1), and so

$$\mathbb{E}A_n^v(t) = \sum_{j=1}^{n} e^{-j(j-1)\Lambda(t)/2} \frac{(2l-1)n_{[j]}}{n_{(j)}}, t \geq 0.$$

It follows from (2.4.8) that $A_n(s) = A_n^v(\Lambda^{-1}(s)),\ s > 0$. Hence if $A_n(\cdot)$ has a jump at time s, then $A_n^v(\cdot)$ has one at time $\Lambda^{-1}(s)$. Since $A_n(\cdot)$ has jumps at $S_n = T_n, S_{n-1} = T_n + T_{n-1}, \ldots, S_2 = T_n + \cdots + T_2$, it follows that the jumps of $A_n^v(\cdot)$ occur at $\Lambda^{-1}(S_n), \ldots, \Lambda^{-1}(S_2)$. Thus, writing T_j^v for the time the sample from a variable-size population spends with j ancestors, we see that

$$\begin{aligned} T_n^v &= \Lambda^{-1}(S_n) \\ T_j^v &= \Lambda^{-1}(S_j) - \Lambda^{-1}(S_{j+1}),\ j = n-1, \ldots, 2. \end{aligned} \tag{2.4.9}$$

This result provides a simple way to simulate the times $T_n^v, T_{n-1}^v, \ldots, T_2^v$. Let $U_n, \ldots, U_2$ be independent and identically distributed random variables having the uniform distribution on (0,1).

Algorithm 2.1 Algorithm to generate $T_n^v, \ldots, T_2^v$ for a variable size process with intensity function Λ:

1. Generate $t_j = -\frac{2\log(U_j)}{j(j-1)}$, $j = 2, 3, \ldots, n$
2. Form $s_n = t_n, s_j = t_j + \cdots + t_n,\ j = 2, \ldots, n-1$
3. Compute $t_n^v = \Lambda^{-1}(s_n)$, $t_j^v = \Lambda^{-1}(s_j) - \Lambda^{-1}(s_{j+1}),\ j = n-1, \ldots, 2.$
4. Return $T_j^v = t_j^v, j = 2, \ldots, n.$

There is also a sequential version of the algorithm, essentially a restatement of the last one:

Algorithm 2.2 Step-by-step version of Algorithm 2.1.

1. Set $t = 0, j = n$
2. Generate $t_j = -\frac{2\log(U_j)}{j(j-1)}$
3. Solve for s the equation

$$\Lambda(t+s) - \Lambda(t) = t_j \tag{2.4.10}$$

4. Set $t_j^v = s, t = t + s, j = j - 1$. If $j \geq 2$, go to 2. Else return $T_n^v = t_n^v, \ldots, T_2^v = t_2^v$.

Note that t_j generated in step 2 above has an exponential distribution with parameter $j(j-1)/2$. If the population size is constant then $\Lambda(t) = t$, and so $t_j^v = t_j$, as it should.

Example For an exponentially growing population $f(x) = e^{-\rho x}$, so that $\Lambda(t) = (e^{\rho t} - 1)/\rho$. It follows that $\Lambda^{-1}(y) = \rho^{-1}\log(1+\rho y)$, and

$$T_n^v = \rho^{-1}\log(1+\rho T_n),\ T_j^v = \frac{1}{\rho}\left(\frac{1+\rho S_j}{1+\rho S_{j+1}}\right), j = 2,\ldots,n-1. \quad (2.4.11)$$

In an exponentially growing population, most of the coalescence events occur near the root of the tree, and the resulting genealogy is then star-like; it is harder to find common ancestors when the population size is large. See Section 4.2 for further illustrations.

3 The Ewens Sampling Formula

In this section we bring mutation into the picture, and show how the genealogical approach can be used to derive the classical Ewens sampling formula. This serves as an introduction to statistical inference for molecular data based obtained from samples.

3.1 The effects of mutation

In Section 2.1 we looked briefly at the process of random drift, the mechanism by which genetic variability is lost through the effects of random sampling. In this section, we study the effect of mutation on the evolution of gene frequencies at a locus with two alleles.

Now we suppose there is a probability $\mu_A > 0$ that an A allele mutates to a B allele in a single generation, and a probability $\mu_B > 0$ that a B allele mutates to an A. The stochastic model for the frequency X_n of the A allele in generation n is described by the transition matrix in (2.1.1), but where

$$\pi_i = \frac{i}{N}(1-\mu_A) + \left(1 - \frac{i}{N}\right)\mu_B. \tag{3.1.1}$$

The frequency π_i reflects the effects of mutation in the gene pool. In this model, it can be seen that $p_{ij} > 0$ for all $i, j \in \mathcal{S}$. It follows that the Markov chain $\{X_n\}$ is irreducible; it is possible to get from any state to any other state. An irreducible finite Markov chain has a limit distribution $\rho = (\rho_0, \rho_1, \ldots, \rho_N)$:

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \rho_k > 0,$$

for any initial distribution for X_0. The limit distribution ρ is also invariant (or stationary), in that if X_0 has distribution ρ then X_n has distribution ρ for every n. The distribution ρ satisfies the balance equations

$$\rho = \rho P,$$

where $\rho_0 + \cdots + \rho_N = 1$.

Once more, the binomial conditional distributions make some aspects of the process simple to calculate. For example,

$$\mathbb{E}(X_n) = \mathbb{E}\mathbb{E}(X_n|X_{n-1}) = N\mu_B + (1-\mu_A-\mu_B)\mathbb{E}(X_{n-1}).$$

At stationarity, $\mathbb{E}(X_n) = \mathbb{E}(X_{n-1}) \equiv \mathbb{E}(X)$, so

$$\mathbb{E}(X) = \frac{N\mu_B}{\mu_A + \mu_B}. \tag{3.1.2}$$

This is also the limiting value of $\mathbb{E}(X_n)$ as $n \to \infty$.

Now we investigate the stationary distribution ρ when N is large. To get a non-degenerate limit, we assume that the mutation probabilities μ_A and μ_B satisfy

$$\lim_{N\to\infty} 2N\mu_A = \theta_A > 0, \lim_{N\to\infty} 2N\mu_B = \theta_B > 0, \tag{3.1.3}$$

so that mutation rates are of the order of the reciprocal of the population size. We define the total mutation rate θ by

$$\theta = \theta_A + \theta_B.$$

Given $X_n = i$, X_{n+1} is binomially distributed with parameters N and π_i given by (3.1.1). Exploiting simple properties of the binomial distribution shows that the diffusion approximation for the fraction of allele A in the population has

$$\mu(x) = -x\theta_A/2 + (1-x)\theta_B/2, \qquad \sigma^2(x) = x(1-x),\ 0 < x < 1. \tag{3.1.4}$$

The stationary density $\pi(y)$ of $Y(\cdot)$ satisfies the ordinary differential equation

$$-\mu(y)\pi(y) + \frac{1}{2}\frac{d\{\sigma^2(y)\pi(y)\}}{dy} = 0,$$

and it follows readily that

$$\pi(y) \propto \frac{1}{\sigma^2(y)} \exp\left(\int^y 2\frac{\mu(u)}{\sigma^2(u)}\, du\right).$$

Hence $\pi(y) \propto y^{\theta_B - 1}(1-y)^{\theta_A - 1}$ and we see that at stationarity the fraction of A alleles has the beta distribution with parameters θ_B and θ_A. The density π is given by

$$\pi(y) = \frac{\Gamma(\theta)}{\Gamma(\theta_A)\Gamma(\theta_B)} y^{\theta_B - 1}(1-y)^{\theta_A - 1},\ 0 < y < 1.$$

In particular,

$$\mathbb{E}(Y) = \frac{\theta_B}{\theta}, \quad \mathrm{Var}(Y) = \frac{\theta_A\theta_B}{\theta^2(\theta+1)}. \tag{3.1.5}$$

Remark. An alternative description of the mutation model in this case is as follows. Mutations occur at rate $\theta/2$, and when a mutation occurs the resulting allele is A with probability π_A and B with probability π_B. This model can be identified with the earlier one with $\theta_A = \theta\pi_A, \theta_B = \theta\pi_B$.

Remark. In the case of the K-allele model with mutation rate $\theta/2$ and mutations resulting in allele A_i with probability $\pi_i > 0, i = 1, 2, \ldots, K$, the stationary density of the (now $(K-1)$-dimensional) diffusion is

$$\pi(y_1, \ldots, y_K) = \frac{\Gamma(\theta)}{\Gamma(\theta\pi_1)\cdots\Gamma(\theta\pi_K)} y_1^{\theta\pi_1 - 1} \cdots y_K^{\theta\pi_K - 1},$$

for $y_i > 0, i = 1, \ldots, K,\ y_1 + \cdots + y_K = 1$.

3.2 Estimating the mutation rate

Modern molecular techniques have made it possible to sample genomic variability in natural populations. As a result, we need to develop the appropriate sampling theory to describe the statistical properties of such samples. For the models described in this section, this is easy to do. If a sample of n chromosomes is drawn with replacement from a stationary population, it is straightforward to calculate the distribution of the number N_A of A alleles in the sample. This distribution follows from the fact that given the population frequency Y of the A allele, the sample is distributed like a binomial random variable with parameters n and Y. Thus

$$\mathbb{P}(N_A = k) = \mathbb{E}\left(\binom{n}{k} Y^k (1-Y)^{n-k}\right).$$

Since Y has the Beta(θ_B, θ_A) density, we see that N_A has the Beta-Binomial distribution:

$$\mathbb{P}(N_A = k) = \binom{n}{k} \frac{\Gamma(\theta)\Gamma(k+\theta_B)\Gamma(n-k+\theta_A)}{\Gamma(\theta_B)\Gamma(\theta_A)\Gamma(n+\theta)}, k = 0, 1, \ldots, n. \quad (3.2.1)$$

It follows from this that

$$\mathbb{E}(N_A) = n\frac{\theta_B}{\theta}, \quad \mathrm{Var}(N_A) = \frac{n(n+\theta)\theta_A\theta_B}{\theta^2(\theta+1)}. \quad (3.2.2)$$

The probability that a sample of size one is an A allele is just $p \equiv \theta_B/\theta$. Had we ignored the dependence in the sample, we might have assumed that the genes in the sample were independently labelled A with probability p. The number N_A of As in the sample then has a binomial distribution with parameters n and p. If we wanted to estimate the parameter p, the natural estimator is $\hat{p} = N_A/n$, and

$$\mathrm{Var}(\hat{p}) = p(1-p)/n.$$

As $n \to \infty$, this variance tends to 0, so that $\hat{p}$ is a (weakly) consistent estimator of p. Of course, the sampled genes are *not* independent, and the true variance of N_A/n is, from (3.2.2),

$$\mathrm{Var}(N_A/n) = \left(1 + \frac{\theta}{n}\right) \frac{\theta_A\theta_B}{\theta^2(1+\theta)}.$$

It follows that $\mathrm{Var}(N_A/n)$ tends to the positive limit $\mathrm{Var}(Y)$ as $n \to \infty$. Indeed, N_A/n is not a consistent estimator of $p = \theta_A/\theta$, because (by the strong law of large numbers) $N_A/n \to Y$, the population frequency of the A allele. This simple example shows how strong the dependence in the sample can be, and shows why consistent estimators of parameters in this subject are

the exception rather than the rule. Consistency typically has to be generated, at least in principle, by sampling variability at many independent loci.

The example in this section is our first glimpse of the difficulties caused by the relatedness of sequences in the sample. This relatedness has led to a number of interesting approaches to estimation and inference for population genetics data. In the next sections we describe the Ewens sampling formula (Ewens (1972)), the first systematic treatment of the statistical properties of estimators of the compound mutation parameter θ.

3.3 Allozyme frequency data

By the late 1960s, it was possible to sample, albeit indirectly, the molecular variation in the DNA of a population. These data came in the form of allozyme frequencies. A sample of size n resulted in a set of genes in which differences between genes could be observed, but the precise nature of the differences was irrelevant. Two *Drosophila* allozyme frequency data sets, each having 7 distinct alleles, are given below:

- *D. tropicalis* Esterase-2 locus [$n = 298$]
  234, 52, 4, 4, 2, 1, 1
- *D. simulans* Esterase-C locus [$n = 308$]
  91, 76, 70, 57, 12, 1, 1

It is clear that these data come from different distributions. Of the first set, Sewall Wright (1978, p303) argued that

> *... the observations do not agree at all with the equal frequencies expected for neutral alleles in enormously large populations.*

This raises the question of what shape these distributions should have under a neutral model. The answer to this was given by Ewens (1972). Because the labels are irrelevant, a sample of genes can be broken down into a set of alleles that occurred just once in the sample, another collection that occurred twice, and so on. We denote by $C_j(n)$ the number of alleles represented j times in the sample of size n. Because the sample has size n, we must have

$$C_1(n) + 2C_2(n) + \cdots + nC_n(n) = n.$$

In this section we derive the distribution of $(C_1(n), \ldots, C_n(n))$, known as the Ewens Sampling Formula (henceforth abbreviated to ESF). To do this, we need to study the effects of mutations in the history of a sample.

Mutations on a genealogy

In Section 3 we will give a detailed description of the ancestral relationships among a sample of individuals. For now, we recall from the last section that in a large population, the number of distinct ancestors at times t in the past

is described by the ancestral process $A_n(t)$. It is clear by symmetry that when the ancestral process moves from k to $k-1$, the two ancestors chosen to join are randomly chosen from the k possibilities. Thus the ancestral relationships among a sample of individuals can be represented as a random rooted bifurcating tree that starts with n leaves (or tips), and joins random pairs of ancestors together at times T_n, $T_n + T_{n-1}$, ..., $W_n = T_n + \cdots + T_2$. All the individuals in the sample are traced back to their most recent common ancestor at time W_n.

Next we examine the effects of mutation in the coalescent tree of a sample. Suppose that a mutation occurs with probability u per gene per generation. The expected number of mutations along a lineage of g generations is therefore gu. With time measured in units of N generations, this is of the form tNu which is finite if u is of order $1/N$. Just as in (3.1.3), we take

$$\theta = 2Nu$$

to be fixed as $N \to \infty$. In the discrete process, mutations arise in the ancestral lines independently on different branches of the genealogical tree. In the limit, it is clear that they arise at the points of independent Poisson processes of rate $\theta/2$ on each branch.

We can now superimpose mutations on the genealogical tree of the sample. For allozyme frequency data, we suppose that every mutation produces a type that has not been seen in the population before. One concrete way to achieve this is to label types by uniform random variables; whenever a mutation occurs, the resulting individual has a type that is uniformly distributed on (0,1), independently of other labels. This model is an example of an *infinitely-many alleles model*.

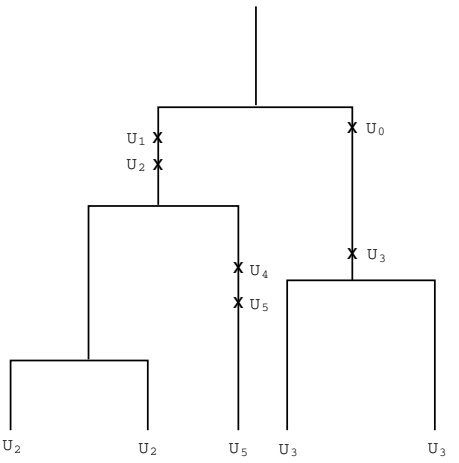
3.4 Simulating an infinitely-many alleles sample

As we will see, the reason that genealogical approaches have become so useful lies first in the fact that they provide a simple way to simulate samples from complex genetics models, and so to compare models with data. To simulate a sample, one need not simulate the whole population first and then sample from that – this makes these methods extremely appealing. Later in these notes we will see the same ideas applied in discrete settings as well, particularly for branching process models. This top down, or 'goodness-of-fit', approach has been used extensively since the introduction of the coalescent by Kingman (1982), Tajima (1983) and Hudson (1983) to simulate the behavior of test statistics which are intractable by analytical means.

To simulate samples of data following the infinitely-many-alleles model is, in principle, elementary. First simulate the genealogical tree of the sample by simulating observations from the waiting times $T_n, T_{n-1}, \ldots, T_2$ and choosing pairs of nodes to join at random. Then we superimpose mutations according to a Poisson process of rate $\theta/2$, independently on each branch.

The effects of each mutation are determined by the mutation process. In the present case, the result of a mutation on a branch replaces the current label with an independently generated uniform random variable. An example is given in Figure 3.1, and the types represented in the sample are labelled U_5, U_2, U_2, U_3, U_3 respectively.

Fig. 3.1. A coalescent tree for $n = 5$ with mutations



3.5 A recursion for the ESF

To derive the ESF, we use a coalescent argument to find a recursion satisfied by the joint distribution of the sample configuration in an infinitely-many-alleles model. Under the infinitely-many-alleles mutation scheme, a sample of size n may be represented as a configuration $\boldsymbol{c} = (c_1, \ldots, c_n)$, where

$$c_i = \text{ number of alleles represented } i \text{ times}$$

and $|\boldsymbol{c}| \equiv c_1 + 2c_2 + \cdots + nc_n = n$. It is convenient to think of the configuration $\boldsymbol{b}$ of samples of size $j < n$ as being an n-vector with coordinates $(b_1, b_2, \ldots, b_j, 0, \ldots, 0)$, and we assume this in the remainder of this section. We define $\boldsymbol{e}_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$, the ith unit vector.

We derive an equation satisfied by the sampling probabilities $q(\boldsymbol{c}), n = |\boldsymbol{c}| > 1$ defined by

$$q(\boldsymbol{c}) = \mathbb{P}(\text{sample of size } |\boldsymbol{c}| \text{ taken at stationarity has configuration } \boldsymbol{c}), \tag{3.5.1}$$

with $q(\boldsymbol{e}_1) = 1$. Suppose then that the configuration is $\boldsymbol{c}$. Looking at the history of the sample, we will either find a mutation or we will be able to

trace two individuals back to a common ancestor. The first event occurs with probability

$$\frac{n\theta/2}{n\theta/2 + n(n-1)/2} = \frac{\theta}{\theta + n - 1},$$

and results in the configuration $\boldsymbol{c}$ if the configuration just before the mutation was $\boldsymbol{b}$, where

(i) $\boldsymbol{b} = \boldsymbol{c}$, and mutation occurred to one of the c_1 singleton lines (probability c_1/n);
(ii) $\boldsymbol{b} = \boldsymbol{c} - 2\boldsymbol{e}_1 + \boldsymbol{e}_2$, and a mutation occurred to an individual in the 2-class (probability $2(c_2 + 1)/n$);
(iii) $\boldsymbol{b} = \boldsymbol{c} - \boldsymbol{e}_1 - \boldsymbol{e}_{j-1} + \boldsymbol{e}_j$ and the mutation occurred to an individual in a j-class, producing a singleton mutant and a new $(j-1)$-class (probability $j(c_j + 1)/n$).

On the other hand, the ancestral join occurred with probability $(n-1)/(\theta + n - 1)$, and in that case the configuration $\boldsymbol{b} = \boldsymbol{c} + \boldsymbol{e}_j - \boldsymbol{e}_{j+1}$, and an individual in one of $c_j + 1$ allelic classes of size j had an offspring, reducing the number of j-classes to c_j, and increasing the number of $(j+1)$-classes to c_{j+1}. This event has probability $j(c_j + 1)/(n-1), j = 1, \ldots, n-1$. Combining these possibilities, we get

$$q(\boldsymbol{c}) = \frac{\theta}{\theta + n - 1}\left[\frac{c_1}{n}q(\boldsymbol{c}) + \sum_{j=2}^{n}\frac{j(c_j+1)}{n}q(\boldsymbol{c} - \boldsymbol{e}_1 - \boldsymbol{e}_{j-1} + \boldsymbol{e}_j)\right] + \frac{n-1}{\theta + n - 1}\left[\sum_{j=1}^{n-1}\frac{j(c_j+1)}{n-1}q(\boldsymbol{c} + \boldsymbol{e}_j - \boldsymbol{e}_{j+1})\right], \tag{3.5.2}$$

where we use the convention that $q(\boldsymbol{c}) = 0$ if any $c_i < 0$. Ewens (1972) established the following result.

Theorem 3.1 *In a stationary sample of size n, the probability of sample configuration $\boldsymbol{c}$ is*

$$\begin{aligned} q(\boldsymbol{c}) &= \mathbb{P}(C_1(n) = c_1, \ldots, C_n(n) = c_n) \\ &= \mathbb{1}(|\boldsymbol{c}| = n)\frac{n!}{\theta_{(n)}}\prod_{j=1}^{n}\left(\frac{\theta}{j}\right)^{c_j}\frac{1}{c_j!}, \end{aligned} \tag{3.5.3}$$

where (as earlier) we have written $x_{(j)} = x(x+1)\cdots(x+j-1), j = 1, 2, \ldots,$ and $|\boldsymbol{c}| = c_1 + 2c_2 + \cdots + nc_n$.

Proof. This can be verified by induction on $n = |\boldsymbol{c}|$ and $k = ||\boldsymbol{c}|| := c_1 + \cdots + c_n$ in the equation (3.5.2) by noting that the right-hand side of the equation has terms with $|\boldsymbol{b}| = n - 1$ and $||\boldsymbol{b}|| \le k$, or with $|\boldsymbol{b}| = n$ and $||\boldsymbol{b}|| < k$.

Remark. Watterson (1974) noted that if $Z_1, Z_2, \ldots$ are independent Poisson random variables with $\mathbb{E}Z_j = \theta/j$, then

$$\mathcal{L}(C_1(n), C_2(n), \ldots, C_n(n)) = \mathcal{L}\left(Z_1, Z_2, \ldots, Z_n | \sum_{i=1}^{n} iZ_i = n\right), \qquad (3.5.4)$$

where $\mathcal{L}(X)$ means 'the distribution of X.'

The ESF typically has a very skewed distribution, assigning most mass to configurations with several alleles represented a few times. In particular, the distribution is far from 'flat'; recall Wright's observation cited in the introduction of this section. In the remainder of the section, we will explore some of the properties of the ESF.

Remark. The ESF arises in many other settings. See Tavaré and Ewens (1997) and Ewens and Tavaré (1998) for a flavor of this.

3.6 The number of alleles in a sample

The random variable $K_n = C_1(n) + \cdots + C_n(n)$ is the number of distinct alleles observed in a sample. Its distribution can be found directly from (3.5.3):

$$\mathbb{P}(K_n = k) = \sum_{\boldsymbol{c}:||\boldsymbol{c}||=k} q(\boldsymbol{c}) = \frac{\theta^k}{\theta_{(n)}} n! \sum_{\boldsymbol{c}:||\boldsymbol{c}||=k} \left(\frac{1}{j}\right)^{c_j} \frac{1}{c_j!} = \frac{\theta^k |S_n^k|}{\theta_{(n)}}, \qquad (3.6.1)$$

where $|S_n^k|$ is the Stirling number of the first kind,

$$|S_n^k| = \text{coefficient of } x^k \text{ in } x(x+1)\cdots(x+n-1),$$

and the last equality follows from Cauchy's formula for the number of permutations of n symbols having k distinct cycles.

Another representation of the distribution of K_n can be found by noting that

$$\begin{aligned}\mathbb{E}s^{K_n} &= \sum_{l=1}^{n} s^l \frac{\theta^l |S_n^l|}{\theta_{(n)}} = \frac{(\theta s)_{(n)}}{\theta_{(n)}} = \frac{\theta s(\theta s+1)\cdots(\theta s+n-1)}{\theta(\theta+1)\cdots(\theta+n-1)} \\ &= s\left(\frac{1}{\theta+1} + \frac{\theta}{\theta+1}s\right)\cdots\left(\frac{n-1}{\theta+n-1} + \frac{\theta}{\theta+n-1}s\right) = \prod_{j=1}^{n} \mathbb{E}s^{\xi_j}\end{aligned}$$

where the ξ_j are independent Bernoulli random variables satisfying

$$\mathbb{P}(\xi_j = 1) = 1 - \mathbb{P}(\xi_j = 0) = \frac{\theta}{\theta + j - 1}, \ j = 1, \ldots, n. \qquad (3.6.2)$$

It follows that we can write

$$K_n = \xi_1 + \cdots + \xi_n, \tag{3.6.3}$$

a sum of independent, but not identically distributed, Bernoulli random variables. Therefore

$$\mathbb{E}(K_n) = \sum_{j=1}^{n} \mathbb{E}\xi_j = \sum_{j=0}^{n-1} \frac{\theta}{\theta + j}, \tag{3.6.4}$$

and

$$\mathrm{Var}(K_n) = \sum_{j=1}^{n} \mathrm{Var}(\xi_j) = \sum_{j=0}^{n-1} \frac{\theta}{\theta + j} - \sum_{j=0}^{n-1} \frac{\theta^2}{(\theta + j)^2} = \sum_{j=0}^{n-1} \frac{\theta j}{(\theta + j)^2}. \tag{3.6.5}$$

For large n, we see that $\mathbb{E}K_n \sim \theta \log n$ and $\mathrm{Var}(K_n) \sim \theta \log n$. It can be shown (cf. Barbour, Holst and Janson (1992)) that the total variation distance between a sum $W = \xi_1 + \cdots + \xi_n$ of independent Bernoulli random variables ξ_i with means p_i, and a Poisson random variable P with mean $p_1 + \cdots + p_n$ satisfies

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(P)) \leq \frac{p_1^2 + \cdots + p_n^2}{p_1 + \cdots + p_n}.$$

It follows from the representation (3.6.3) that there is a constant c such that

$$d_{TV}(\mathcal{L}(K_n), \mathcal{L}(P_n)) \leq \frac{c}{\log n}, \tag{3.6.6}$$

where P_n is a Poisson random variable with mean $\mathbb{E}K_n$. As a consequence,

$$\frac{K_n - \mathbb{E}K_n}{\sqrt{\mathrm{Var}K_n}} \Rightarrow \mathrm{N}(0, 1), \tag{3.6.7}$$

and the same result holds if the mean and variance of K_n are replaced by $\theta \log n$.

3.7 Estimating θ

In this section, we return to the question of inference about θ from the sample. We begin with an approach used by population geneticists prior to the advent of the ESF.

The sample homozygosity

It is a simple consequence of the ESF (with $n = 2$) that

$$\mathbb{P}(\text{two randomly chosen genes are identical}) = \frac{1}{1 + \theta}.$$

In a sample of size n, define for $i \neq j$

$$\delta_{ij} = \begin{cases} 1 & \text{if genes } i \text{ and } j \text{ are identical} \\ 0 & \text{otherwise} \end{cases}$$

and set

$$F_n^* = \frac{2}{n(n-1)} \sum_{i<j} \delta_{ij}.$$

We call F_n^* the *homozygosity* of the sample; it is the probability that two randomly chosen distinct members of the sample of size n have identical types. It is elementary to show that

$$\mathbb{E}(F_n^*) = \frac{1}{1+\theta}. \tag{3.7.1}$$

The variance of F_n^* is more difficult to calculate, but it can be shown that

$$\mathbb{E}(F_n^*)^2 = \frac{1}{n(n-1)} \left(\frac{2}{1+\theta} + \frac{8(n-2)}{(1+\theta)(2+\theta)} + \frac{(n-2)(n-3)(6+\theta)}{(1+\theta)(2+\theta)(3+\theta)} \right). \tag{3.7.2}$$

The results in (3.7.1) and (3.7.2) can be combined to calculate $\mathrm{Var}(F_n^*)$. We see that as $n \to \infty$,

$$\mathrm{Var}(F_n^*) \to \frac{2\theta}{(1+\theta)^2(2+\theta)(3+\theta)}, \tag{3.7.3}$$

as found by Stewart (1976). It turns out that F_n^* converges in distribution as $n \to \infty$ to a limiting random variable F^* having variance given in (3.7.3).

If there are l types in the sample, with μ_j of type $j, j = 1, \ldots, l$, then

$$F_n^* = \sum_{j=1}^{l} \frac{\mu_j(\mu_j - 1)}{n(n-1)}. \tag{3.7.4}$$

We note that the homozygosity is often calculated as

$$F_n = \sum_{j=1}^{l} \left(\frac{\mu_j}{n} \right)^2. \tag{3.7.5}$$

The difference between F_n and F_n^* is of order n^{-1}: F_n is the probability that two genes taken *with* replacement are identical, F_n^* the probability that two genes sampled *without* replacement are identical.

We have seen that $\mathbb{E}(F_n^*) = 1/(1+\theta)$. This suggests a method of moments estimator for θ obtained by equating the observed sample homozygosity to its expectation:

$$\tilde{\theta} = \frac{1}{F_n^*} - 1$$

The right hand side of (3.7.4) shows that $\tilde{\theta}$ depends largely on the partition of the data into types, and not on the number of types. We will see that the

latter is sufficient for θ, so standard statistical theory suggests that $\tilde{\theta}$ might not be a good estimator – it is based largely on those parts of the data which are uninformative for θ. To examine this issue further, we used a coalescent simulation to generate 10,000 samples of size 100 from the infinitely-many-alleles process for different values of the target θ, and computed the estimator $\tilde{\theta}$ for each of them. In Table 1 below are some summary statistics from these simulations.

Table 1. Simulated properties of $\tilde{\theta}$ in samples of size $n = 100$

	$\theta = 0.1$	$\theta = 1.0$	$\theta = 5.0$	$\theta = 10.0$
mean	0.15	1.38	6.00	11.38
std. dev.	0.32	1.03	2.60	4.15
RMSE†	0.32	1.10	2.79	4.37
median	0.00	1.19	5.73	11.01
5th %ile	0.00	0.09	2.21	5.25
95th %ile	0.94	3.36	10.73	18.80

†RMSE: root mean square error. 10,000 replicates used.

It can be seen that the estimator $\tilde{\theta}$ is biased upwards. This might be anticipated, because

$$\mathbb{E}(\tilde{\theta}) = \mathbb{E}(1/F_n^* - 1) \geq 1/\mathbb{E}(F_n^*) - 1 = \theta,$$

the inequality following from an application of Jensen's Inequality. We note that the estimator $\tilde{\theta}$ has a non-degenerate limit as $n \to \infty$, precisely because F_n^* does. Thus $\tilde{\theta}$ is *not* a consistent estimator of θ. However, a consistent estimator can be derived by using the number of types observed in the sample, as we now show.

Estimation using the number of types in the sample

Notice from (3.5.3) and (3.6.1) that the conditional distribution of $\boldsymbol{c}$, given that $K_n = k$, does not depend on θ:

$$\begin{aligned}
\mathbb{P}(\boldsymbol{c}|K_n = k) &= q(\boldsymbol{c}) \,/\mathbb{P}(K_n = k) \\
&= \frac{n!\theta^k}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{1}{j}\right)^{c_j} \frac{1}{c_j!} \Big/ \frac{\theta^k |S_n^k|}{\theta_{(n)}} \\
&= \frac{n!}{|S_n^k|} \prod_{j=1}^{n} \left(\frac{1}{j}\right)^{c_j} \frac{1}{c_j!}. \qquad (3.7.6)
\end{aligned}$$

It follows that K_n is a sufficient statistic for θ; it contains all the information useful for estimating θ. The maximum likelihood estimator of θ may be found from (3.6.1). If k alleles are observed in the sample, then the log-likelihood is

$$\log L(\theta) = \log(|S_n^k|) + k \log \theta - \sum_{j=0}^{n-1} \log(\theta + j).$$

Differentiating with respect to θ shows that the maximum likelihood estimator $\hat{\theta}$ of θ may be found by solving the equation

$$k = \sum_{j=0}^{n-1} \frac{\theta}{\theta + j}. \tag{3.7.7}$$

As can be seen from (3.6.4), this is also the moment estimator of θ. The Fisher information may be calculated readily from the log-likelihood, and we find that the asymptotic variance of $\hat{\theta}$ is

$$\mathrm{Var}(\hat{\theta}) \approx \theta^2 / \mathrm{Var}(K_n). \tag{3.7.8}$$

Therefore $\hat{\theta}$ is consistent for θ. Indeed, asymptotically $\hat{\theta}$ has a Normal distribution with mean θ and variance $\theta / \log n$. We used the simulated data described above to assess the properties of the estimator $\hat{\theta}$. Some results are given in Table 2. It can be seen that the distribution of $\hat{\theta}$ is much more concentrated than that of $\tilde{\theta}$, and $\hat{\theta}$ seems to be somewhat less biased than $\tilde{\theta}$. Histograms comparing the two estimators appear in Figure 3.2.

It is worth relating these two approaches to estimating θ. If we were given the values of each $\delta_{ij}, 1 \le i < j \le n$, then we would be able to calculate the value of K_n, and each of the allele frequencies. We can see that summarizing the δ_{ij} in the form of F_n^* throws away a lot of information – for example, the summary statistic results in an inconsistent estimator of θ. We shall meet this phenomenon again when we investigate estimation in the infinitely-many-sites model.

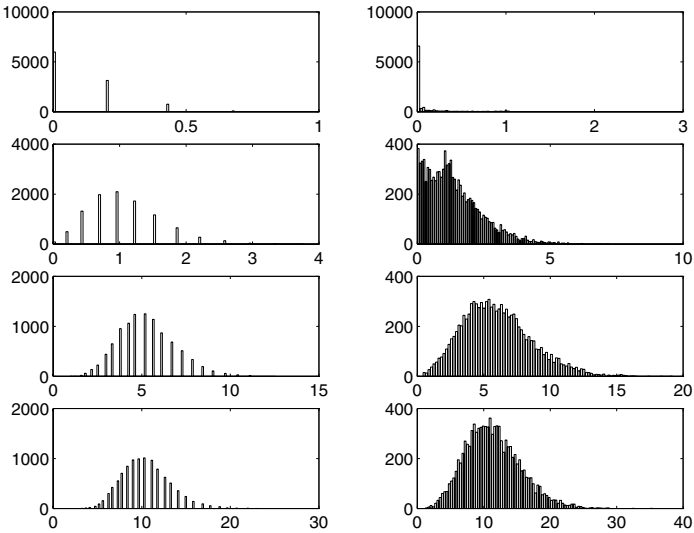
3.8 Testing for selective neutrality

One might try to perform a "goodness of fit" test on genetic data to see whether the Ewens sampling formula is appropriate. If the fit is rejected, it may be evidence of selection (or of geographical structure, variation in population sizes, other mutation mechanisms or other unnamed departures from the model). Watterson (1977) suggested using the sample homozygosity F_n defined in (3.7.5) as a test statistic. Under neutrality, the conditional distribution of the counts is given by (3.7.6), from which the null distribution of F_n follows. F_n will tend to have larger values when the allele frequencies are skewed, and smaller values when the allele frequencies are more equal. When testing for heterosis, small values of the test statistic lead to rejection

Table 2. Simulated properties of $\hat{\theta}$ in samples of size $n = 100$.

	$\theta = 0.1$	$\theta = 1.0$	$\theta = 5.0$	$\theta = 10.0$
mean	0.11	1.03	5.12	10.17
std. dev.	0.15	0.54	1.57	2.70
RMSE	0.15	1.03	1.57	2.71
median	0.00	0.95	5.14	9.70
5th %ile	0.00	0.20	2.95	6.15
95th %ile	0.43	1.87	7.82	15.00

Fig. 3.2. Histograms of 10,000 replicates of estimators of θ based on samples of size $n = 100$. Left hand column is MLE $\hat{\theta}$, right hand column is $\tilde{\theta}$. First row corresponds to $\theta = 0.1$, second to $\theta = 1.0$, third to $\theta = 5.0$, and fourth to $\theta = 10.0$.



of neutrality. For the *D. tropicalis* data in the introduction we have $F_{298} = 0.6475$, while for the *D. simulans* data we have $F_{308} = 0.2356$. Significance points of the distribution under neutrality were given in Watterson (1978), but they can be simulated rapidly. One approach, with ties to combinatorics, is outlined in the complements. Using this method, the P-value for the first set is 0.87, while for the second set it is 0.03. Thus, in contrast to Wright's expectation, the *D. simulans* do not fit neutral expectations. We will not focus further on tests of neutrality in these notes. An up-to-date discussion about detecting neutrality is given in Kreitman (2000).

4 The Coalescent

In the last two sections we studied the behavior of the genealogy of a sample from a Wright-Fisher model when the population size N is large. We introduced the ancestral process $A_n(t)$ that records the number of distinct ancestors of a sample of size n a time t earlier, and we studied some of its properties. In this section we describe in more detail the structure of Kingman's coalescent, a continuous time process whose state space is the set of equivalence relations on the set $[n] \equiv \{1, 2, \ldots, n\}$. We also give an alternative representation as a bifurcating tree, and we discuss the robustness of these approximations to different models of reproduction.

4.1 Who is related to whom?

We record information not only about the number of ancestors at various times in the past, but also information about which individuals are descended from which ancestors. For some fixed time t, one way of doing this is by labelling the individuals in the sample from the set $\{1, \ldots, n\}$ and defining a (random) equivalence relation $\sim$ on $[n]$ by

$i \sim j$ if and only if individuals i and j share a common ancestor at time t.

It is often easiest to describe the equivalence relation by listing the equivalence classes. Note that each equivalence class corresponds to a particular ancestor of the sample at time t, and that the individuals in the equivalence class are exactly those who are descended from the ancestor of the class.

More formally, we could label the individuals in the sample from the set $[n]$. If at time t there are $A_n(t) = k$ ancestors of the sample, we could list the members of the sample descended from each particular ancestor. This would give us an unordered collection $E_1 \equiv \{i_{11}, \cdots, i_{1l_1}\}, E_2 \equiv \{i_{21}, \cdots, i_{2l_2}\}, \ldots, E_k \equiv \{i_{k1}, \cdots, i_{kl_k}\}$ of sets which would partition $[n]$, *i.e.* $E_i \cap E_j = \emptyset$ $i \neq j$ and $E_1 \cup \cdots \cup E_k = [n]$. We often refer to the sets $E_1, \ldots, E_k$ as *classes*, or *equivalence classes.*

Denote by $C(t)$ the (random) partition (or equivalently, equivalence relation) which is obtained from the genealogy in this way. What are the dynamics of the process $\{C(t) : \ t \geq 0\}$? Suppose that $C(t) = \alpha$ for some partition α with k classes (we write $|\alpha| = k$). As t increases and we go further into the past, the process will remain constant until the first occasion that two of the k individuals who are the ancestors of the classes are involved in a coalescence. When this happens, those two individuals and hence all their descendants in the two equivalence classes will share a common ancestor. The effect is to merge or coalesce the two classes corresponding to these two individuals. The rate at which this happens to a particular pair of individuals (and hence to a particular pair of classes) is 1. Note that this argument and the fact that population events happen at the points of a Poisson process ensures that the process $C(\cdot)$ is Markovian.

In summary, denote by $\mathcal{E}_n$ the set of equivalence relations on $[n]$. The process $\{C(t) : t \geq 0\}$ is a continuous-time Markov chain on $\mathcal{E}_n$ with

$$\begin{aligned} C(0) = \Delta &\equiv \{(i,i), \quad i = 1, 2, \ldots, n\} \\ &\equiv \{\{1\}\{2\}\ldots\{n\}\}, \end{aligned}$$

the state in which "nobody is related to anyone else", and transition rates $\{q_{\alpha\beta},\ \alpha, \beta \in \mathcal{E}_n\}$ given by

$$q_{\alpha\beta} = \begin{cases} -\binom{k}{2} & \text{if } \alpha = \beta,\ |\alpha| = k \\ 1 & \text{if } \alpha \prec \beta \\ 0 & \text{otherwise} \end{cases} \tag{4.1.1}$$

where the notation $\alpha \prec \beta$ means that the partition β may be obtained from α by merging two of the classes in α. The observation that the sample may eventually be traced back to a single common ancestor means that almost surely

$$\begin{aligned} \lim_{t\to\infty} C(t) = \Theta &\equiv \{(i,j),\ i,j = 1, 2, \ldots, n\} \\ &= \{\{1, 2, \ldots, n\}\} \end{aligned}$$

so that everybody is related to everybody else and there is just one class.

The process $\{C(t), t \geq 0\}$ is known as the $n-$*coalescent*, or *coalescent*. To calculate its distribution, it is convenient to study the discrete time (embedded) jump chain $\{\mathcal{C}_k;\ k = n, n-1, \ldots, 1\}$ obtained by watching the continuous-time process $C(\cdot)$ only at those times when it changes state. This chain starts from $\mathcal{C}_n = \Delta$ and has transition probabilities

$$\mathbb{P}(\mathcal{C}_{k-1} = \beta | \mathcal{C}_k = \alpha) = \begin{cases} \binom{k}{2}^{-1} & \text{if } \alpha \prec \beta,\ |\alpha| = k \\ 0 & \text{otherwise.} \end{cases}$$

Thus $C(\cdot)$ moves through a sequence $\Delta = \mathcal{C}_n \prec \mathcal{C}_{n-1} \prec \cdots \prec \mathcal{C}_1 = \Theta$, spending (independent) exponential amounts of time in each state $\mathcal{C}_k \in \mathcal{E}_n$ with respective parameters $\binom{k}{2},\ k = n, n-1, \ldots, 2$, before being absorbed in state Θ.

Notice that in $C(\cdot)$ transition rates from a state α (and hence the time spent in α) depend on α only through $|\alpha|$, and that

$$|C(t)| = A_n(t),$$

since classes in $C(t)$ correspond to ancestors of the sample. Thus the joint distributions of $\{A_n(t); t \geq 0\}$ conditional on the sequence $\mathcal{C}_n, \ldots, \mathcal{C}_1$ are just the same as its unconditional distributions. Hence $\{\mathcal{C}_k\}$ and $\{A_n(t)\}$ are independent processes. Thus

$$C(t) = \mathcal{C}_{A_n(t)},\ t \geq 0$$

and

$$\begin{aligned}\mathbb{P}(C(t)=\alpha) &= \sum_{j=1}^{n}\mathbb{P}(C(t)=\alpha|A_n(t)=j)\mathbb{P}(A_n(t)=j)\\ &= \sum_{j=1}^{n}\mathbb{P}(\mathcal{C}_j=\alpha)\mathbb{P}(A_n(t)=j)\\ &= \mathbb{P}(A_n(t)=|\alpha|)\mathbb{P}(\mathcal{C}_{|\alpha|}=\alpha).\end{aligned}$$

The distribution of $A_n(t)$ has been given earlier. That of $\mathcal{C}_j$ is given in the following theorem of Kingman (1982a).

Theorem 4.1 *For the jump chain of the n-coalescent,*

$$\mathbb{P}(\mathcal{C}_j=\alpha)=\frac{(n-j)!j!(j-1)!}{n!(n-1)!}\lambda_1!\cdots\lambda_j!$$

where $|\alpha|=j$ *and* $\lambda_1,\ldots,\lambda_j$ *are the sizes of the equivalence classes of* α.

Proof. Use backward induction. The result is clearly true when $j=n$. Then

$$\begin{aligned}\mathbb{P}(\mathcal{C}_{j-1}=\beta)\equiv p_{j-1}(\beta) &= \sum_{\alpha\in\mathcal{E}_n}p_j(\alpha)\mathbb{P}(\mathcal{C}_{j-1}=\beta|\mathcal{C}_j=\alpha)\\ &= \sum_{\alpha\prec\beta}p_j(\alpha)\frac{2}{j(j-1)}\end{aligned}$$

Write $\lambda_1,\ldots,\lambda_{j-1}$ for the sizes of the equivalence classes of β. Then those of α are $\lambda_1,\ldots,\lambda_{l-1},m,\lambda_l-m,\lambda_{l+1},\ldots,\lambda_{j-1}$ for some $l,\ l=1,\ldots,j-1$ and some $m,\ m=1,2,\ldots,\lambda_l-1$. Using the inductive hypothesis, we have

$$\begin{aligned}p_{j-1}(\beta) &= \sum_{l=1}^{j-1}\sum_{m=1}^{\lambda_l-1}\frac{2}{j(j-1)}\frac{(n-j)!j!(j-1)!}{n!(n-1)!}\\ &\qquad\times\lambda_1!\cdots\lambda_{l-1}!m!(\lambda_l-m)!\lambda_{l+1}!\cdots\lambda_{j-1}!\frac{1}{2}\binom{\lambda_l}{m}\\ &= \frac{(n-j)!(j-1)!(j-2)!}{n!(n-1)!}\lambda_1!\cdots\lambda_{j-1}!\sum_{l=1}^{j-1}\sum_{m=1}^{\lambda_l-1}1\\ &= \frac{(n-j+1)(n-j)!(j-1)!(j-2)!}{n!(n-1)!}\lambda_1!\cdots\lambda_{j-1}!\end{aligned}$$

as required. □

Note that the distribution of $\mathcal{C}$ and hence $C(\cdot)$ depends only on the sizes of the equivalence classes rather than on which individuals are in these classes.

4.2 Genealogical trees

Knowledge of a sample path of the n-coalescent, the value of $C(t)$ for all $t \geq 0$, specifies the time for which there are n distinct ancestors of the sample, which two individuals share an ancestor when the number of ancestors drops by 1, the time for which there are $n-1$ distinct ancestors, which two ancestors share an ancestor when the number drops from $n-1$ to $n-2$, and so on. Eventually we have information about the times between coalescences and knowledge of which ancestors coalesce. Another, perhaps more natural, way of representing this information is as a genealogical tree. The lengths of the various branches are proportional to the times between the various events.

It is convenient to think of the n-coalescent as a random, rooted, binary tree, with lengths attached to the edges, instead of its original form as a stochastic process where values are partitions of $[n]$. The structure of the genealogical process translates easily to the random tree: the leaves of the tree represent the n sequences in the sample. The first join in the tree occurs at time T_n, and results in the joining of two randomly chosen sequences. There are now $n-1$ nodes in the tree, and the next coalescence event occurs a time T_{n-1} later, and results in the joining of two nodes chosen at random from the $n-1$. This structure is continued until the final two nodes are joined at the most common ancestor, at time W_n.

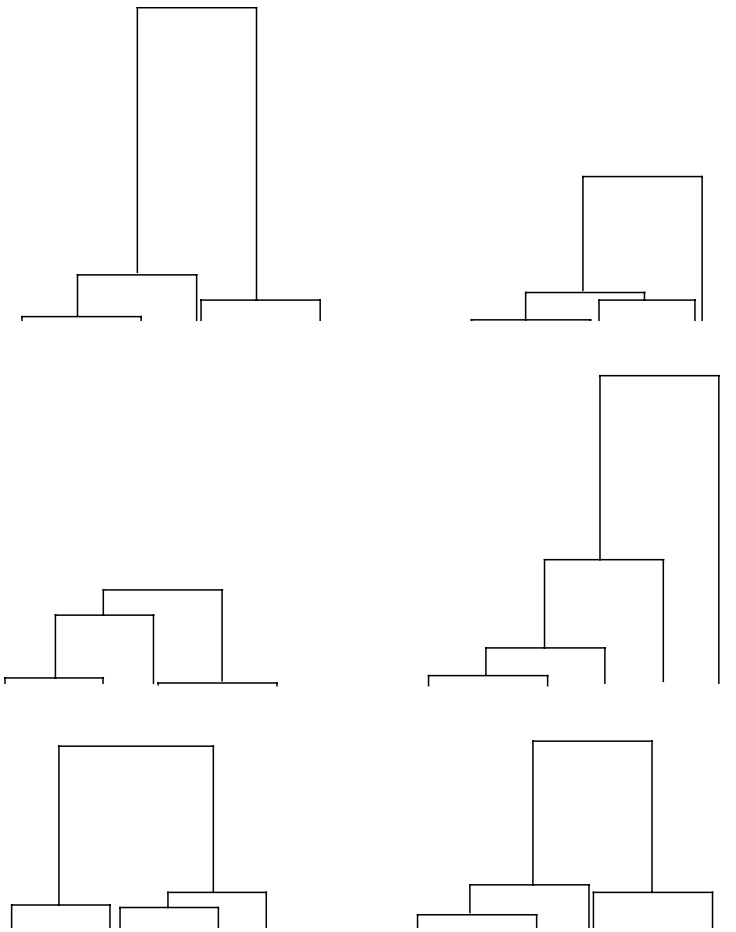
Some simulated genealogical trees for a sample of size 5 from a constant population are shown in Figure 4.1. It is instructive derive the values of the coalescent process from such a tree.

In Figure 4.2 coalescent trees for samples of size 6 and 32 from a constant size population are shown, and in Figure 4.3 trees for samples of size 6 in both constant and exponentially growing populations are shown. One of the most striking qualitative properties, which is evident in Figure 4.2, is the extent to which the tree is dominated by the last few branches. The mean time for which the tree has two branches is 1. The mean time for which the tree has more than two branches, namely $(1 - 2/n)$, is smaller: for much of the time since its common ancestor, the sample has only two ancestors. Further, for any sample size n, the variability in the time T_2 for which there are two branches in the tree accounts for most of the variability in the depth of the whole tree. These observations reinforce the theoretical results given earlier in Section 2.3. The simulated tree with exponential growth in Figure 4.3 clearly displays the star-like nature of the tree alluded to in Section 2.4.

4.3 Robustness in the coalescent

We have seen that the genealogy of the Wright-Fisher model can be described by the coalescent when the population size is large. In this section, we outline how the coalescent arises as an approximation for a wide variety of other reproduction models having constant population size.

Fig. 4.1. Six realizations, drawn on the same scale, of coalescent trees for a sample of $n = 5$. (In each tree the labels 1,2,3,4,5 should be assigned at random to the leaves.)



We noted earlier that in the Wright-Fisher model individuals have independent Poisson-distributed numbers of offspring, conditioned on the requirement that the total population size be fixed at N. Let ν_i be the number of offspring born to individual i, $i = 1, 2, \ldots, N$. We saw in (2.2.1) that $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$ has a multinomial distribution:

$$\mathbb{P}(\nu_1 = m_1, \ldots, \nu_N = m_N) = \frac{N!}{m_1! \cdots m_N!} \left(\frac{1}{N}\right)^N$$

provided $m_1 + \cdots + m_N = N$. In particular the ν_i are identically distributed (but not of course independent), and

Fig. 4.2. Coalescent trees for samples of size 6 and 32 from a population of constant size

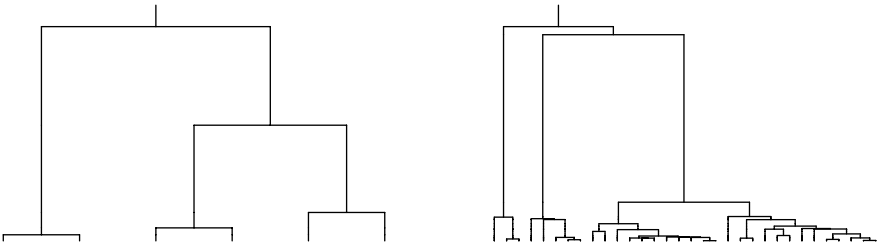
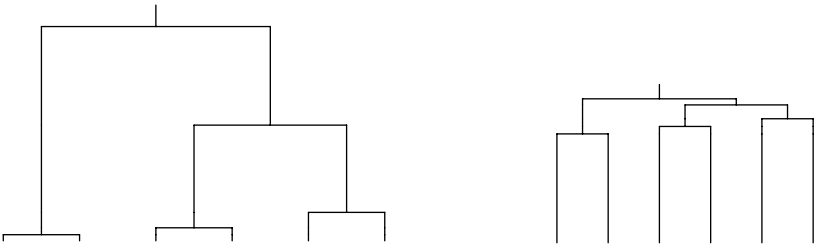


Fig. 4.3. The coalescent tree of a sample of size 6 (constant population size in left panel, exponentially growing population in right panel)



$$\mathbb{E}(\nu_1) = 1,\ \sigma_N^2 \equiv \mathrm{Var}(\nu_1) = 1 - \frac{1}{N}. \tag{4.3.1}$$

Next we consider two other reproduction models that capture some of the features of the Wright-Fisher case. Suppose first that $\boldsymbol{\nu} \equiv (1, 1, \ldots, 1)$, so that each individual has precisely one offspring. For this model,

$$\mathbb{E}(\nu_1) = 1,\ \sigma_N^2 = 0.$$

Now consider the opposite extreme, in which precisely one individual has all the offspring. Then $\boldsymbol{\nu} = N\boldsymbol{e}_i = N(0, \ldots, 1, 0, \ldots, 0)$ for some $i = 1, \ldots, N$. For this case,

$$\mathbb{E}(\nu_1) = 1,\ \sigma_N^2 = N - 1. \tag{4.3.2}$$

Our interest focuses on the asymptotic behavior of the genealogy as $N \to \infty$. In the second model the individuals in the sample never share common ancestors, and in the third the sample can be traced back to a single individual in one generation. Clearly neither of these models has an interesting genealogy! We shall see that the way to distinguish the three models can be based on the behavior of σ_N^2: for the Wright-Fisher model, $\sigma_N^2 \to 1$, for the second model $\sigma_N^2 = 0$, and for the third model $\sigma_N^2 \to \infty$. If time is to be rescaled in units proportional to N, then we get a non-degenerate genealogy if $\sigma_N^2 \to \sigma^2 \in (0, \infty)$.

General reproduction models with reproductive symmetry, introduced by Cannings (1974), can be formulated as follows.

(i) Constant population size requires that $\nu_1 + \cdots + \nu_N = N$.
(ii) The collection of random variables $\nu_1, \ldots, \nu_N$ is exchangeable. That is, the distribution of offspring numbers does not depend on the way in which the individuals are labelled.
(iii) The distribution of $(\nu_1, \ldots, \nu_N)$ is the same in each generation. This is time stationarity.
(iv) The joint distribution of $(\nu_1, \ldots, \nu_N)$ is independent of family sizes in other generations. This is neutrality: offspring numbers for particular individuals do not depend on ancestral offspring numbers.

Some properties of this general model are elementary to obtain. For example, since

$$\nu_1 + \cdots + \nu_N = N \tag{4.3.3}$$

and the ν_i have identical distributions it follows that

$$\mathbb{E}(\nu_1) = 1.$$

Squaring (4.3.3) and taking expectations shows that

$$\mathrm{Cov}(\nu_1, \nu_2) = \frac{-\sigma_N^2}{N-1}. \tag{4.3.4}$$

Any particular distribution for $(\nu_1, \ldots, \nu_N)$ which satisfies the conditions above specifies a model for the reproduction of the population. The main result is that under minor additional conditions, the n-coalescent provides a good approximation of the genealogy of such a model when the population size is large, and time is measured in units proportional to N generations.

We begin by studying the ancestral process in a sample of size n from a population model of size N. The analog of (2.2.3) is given in the next lemma; cf. Cannings (1974) and Gladstien (1978).

Lemma 4.2 *For $1 \le k \le n$, we have*

$$g_{kj} = \binom{N}{k}^{-1} \binom{N}{j} \sum_{\boldsymbol{b} \in \Delta_j^k} \mathbb{E}\binom{\nu_1}{b_1} \cdots \binom{\nu_j}{b_j} \tag{4.3.5}$$

where

$$\Delta_j^k = \{(l_1, \ldots, l_j) : \ l_i \in \mathbb{N}, \ i = 1, \ldots, j; \ l_1 + \cdots + l_j = k\}.$$

Proof. Conditional on the offspring numbers $\nu = (\nu_1, \ldots, \nu_N)$ we have

$$\mathbb{P}(k \text{ have } j \text{ distinct parents}|\nu) = \sum_{\substack{l_1,\ldots,l_j \\ \text{distinct } \in[N]}} \sum_{\boldsymbol{b}\in\Delta_j^k} \binom{N}{k}^{-1} \prod_{m=1}^{j} \binom{\nu_{l_m}}{b_m}.$$

Taking expectations and using the exchangeability assumption completes the proof. □

Kingman's celebrated result gives conditions under which the genealogy of a sample is approximated by the coalescent. He showed (Kingman (1982b)) that if

(i) $\sigma_N^2 \equiv \mathrm{Var}(\nu_1) \to \sigma^2 \in (0,\infty)$ as $N \to \infty$;
(ii) $\sup_N \mathbb{E}(\nu_1^k) < \infty \quad k = 3, 4, \ldots$.

and time is measured in units of $N\sigma^{-2}$ generations, then in the limit as $N \to \infty$, the genealogical structure of the sample is well approximated by the coalescent. Thus any result which follows from the fact that sample genealogies are described by an n-coalescent will be approximately true for any large population evolving according to an exchangeable model. The assumption of large population is reasonable in many genetics applications.

Note that the variance of the offspring distribution plays a role in the approximation of genealogy by the coalescent. If time is scaled in units of N generations, then the ancestral process appropriate for the sample is given by $A_n(\sigma^2 t), t \geq 0$. On this time scale, the waiting time T_j while the sample has j distinct ancestors has an exponential distribution with mean

$$\mathbb{E}T_j = \frac{2}{\sigma^2 j(j-1)}$$

in coalescent units, or

$$\frac{2N}{\sigma^2 j(j-1)}$$

in units of generations. It should be clear that when inferring properties of the ancestral tree from data, the parameter σ^2 has to be estimated.

Remark. As noted in Kingman (2000)), his attempt to understand the structure of the Ewens sampling formula led directly to his development of the coalescent. Kingman (1982c) derives the Ewens Sampling Formula in (3.5.3) directly from the effects of mutation in the coalescent. Define a relation $\mathcal{R} \in \mathcal{E}_n$ which contains (i,j) if, on watching the equivalence classes of $C(t)$ containing i and j until the time they coincide, we observe no mutations to either. Kingman gives the distribution of $\mathcal{R}$ as

$$\mathbb{P}(\mathcal{R} = \xi) = \frac{\theta^k}{\theta_{(n)}} \prod_{j=1}^{k} (\lambda_j - 1)!, \tag{4.3.6}$$

where $\lambda_1, \ldots, \lambda_k$ are the sizes of the equivalence classes of $\mathcal{R}$. If we multiply this by the number of ξ that have the given sizes, namely

$$\frac{n!}{\lambda_1! \cdots \lambda_k! c_1! \cdots c_n!},$$

where c_j is the number of the λ_i equal to j, we obtain the ESF. Thus the ESF is indeed a consequence of mutation in the coalescent.

4.4 Generalizations

Since the introduction of Kingman's coalescent several authors have studied related approximations. For populations of constant size, Möhle (1998) has phrased the approximations in terms of the parameter

$$c_N = \frac{\mathrm{Var}(\nu_1)}{N-1},$$

which is the probability that two individuals chosen at random without replacement from the same generation have the same parent; cf. (4.3.4). The natural time scale is then in units of $\lfloor c_N^{-1} \rfloor$ generations.

We assume in what follows that $c_N > 0$ for sufficiently large N, and that, for integers $k_1 \geq \cdots \geq k_j \geq 2$ the limits

$$\phi_j(k_1, \ldots, k_j) = \lim_{N \to \infty} \frac{\mathbb{E}((\nu_1)_{[k_1]} \cdots (\nu_j)_{[k_j]})}{N^{k_1 + \cdots + k_j - j} c_N} \tag{4.4.1}$$

exist, and that

$$c = \lim_{N \to \infty} c_N \tag{4.4.2}$$

exists.

A complete classification of the limiting behavior of the finite population coalescent process (run on the new time scale) is given by Möhle and Sagitov (2001). In the case

$$c = 0, \quad \phi_j(k_1, \ldots, k_j) = 0 \text{ for } j \geq 2$$

the limiting process is Kingman's coalescent described earlier.

More generally, when $c = 0$ the limiting process is a continuous time Markov chain on the space of equivalence relations $\mathcal{E}_n$, with transition rates given by

$$q_{\alpha\beta} = \begin{cases} \phi_a(b_1, \ldots, b_a) & \text{if } \alpha \subseteq \beta, \\ 0 & \text{otherwise} \end{cases} \tag{4.4.3}$$

In (4.4.3), a is the number of equivalence classes in α, $b_1 \geq b_2 \geq \cdots \geq b_a$ are the ordered sizes of the groups of merging equivalence classes of β, and b is the number of equivalence classes of β. Note that $\phi_1(2) = 1$, so this does indeed reduce to the transition rates in (4.1.1) in the Kingman case. For rates

of convergence of such approximations see Möhle (2000), and for analogous results in the case of variable population size see Möhle (2002).

When $c > 0$, the limit process is a discrete time Markov chain on $\mathcal{E}_n$, with transition matrix P given by $P = I + cQ$, where Q has entries given in (4.4.3). This case obtains, for example, when some of the family sizes are of order N with positive probability. In these limits many groups of individuals can coalesce at the same time, and the resulting coalescent tree need not be bifurcating. Examples of this type arise when a small number of individuals has a high chance of producing most of the offspring, as is the case in some fish populations. For related material, see also Pitman (1999), Sagitov (1999) and Schweinsberg (2000).

4.5 Coalescent reviews

Coalescents have been devised for numerous other population genetics settings, most importantly to include recombination (Hudson (1983)), a subject we return to later in the notes. There have been numerous reviews of aspects of coalescent theory over the years, including Hudson (1991, 1992), Ewens (1990), Tavaré (1993), Donnelly and Tavaré (1995), Fu and Li (1999), Li and Fu (1999) and Neuhauser and Tavaré (2001). Nordborg (2001) has the most comprehensive review of the structure of the coalescent that includes selfing, substructure, migration, selection and much more.

5 The Infinitely-many-sites Model

We begin this section by introducing a data set that will motivate the developments that follow. The data are part of a more extensive mitochondrial data set obtained by Ward *et al.* (1991). Table 3 describes the segregating sites (those nucleotide positions that are not identical in all individuals in the sample) in a collection of sequences of length 360 base pairs sampled from the D-loop of 55 members of the Nuu Chah Nulth native American Indian tribe. The data exhibit a number of important features. First, each segregating site is either *purine* (A, G) or *pyrimidine* (C, T); no transversions are observed in the data. Thus at each segregating site one of two possible nucleotides is present. The segregating sites are divided into 5 purine sites and 13 pyrimidine sites. The right-most column in the table gives the multiplicity of each distinct allele (here we call each distinct sequence an allele). Notice that some alleles, such as e and j, appear frequently whereas others, such as c and n appear only once. We would like to explore the nature of the mutation process that gave rise to these data, to estimate relevant genetic parameters and to uncover any signal the data might contain concerning the demographic history of the sample. Along the way, we introduce several aspects of the theory of the infinitely-many-sites model.

The mutations represented on a tree

In our example, there are $n = 14$ distinct sequences, and each column consists of two possible characters, labelled 0 and 1 for simplicity. In order to summarize these data, we compute the numbers $\Pi(i, j)$ giving the number of coordinates at which the ith and jth of the n sequences differ. $\Pi(i, j)$ is the Hamming distance between sequences i and j. This results in a symmetric $n \times n$ matrix Π with 0 down the diagonal. For our example, the off-diagonal elements of Π are given in Table 4

It is known (cf. Buneman (1971), Waterman (1995) Chapter 14, Gusfield (1997) Chapter 17) that if an $n \times s$ data matrix representing n sequences each of k binary characters, satisfies the *four-point condition*

$$\begin{aligned} &\text{For every pair of columns, not more than three} \\ &\text{of the patterns } 00, 01, 10, 11 \text{ occur} \end{aligned} \tag{5.0.1}$$

then there is an unrooted tree linking the n sequences in such a way that the distance from sequence i to sequence j is given by the elements of the matrix D. Our example set does indeed satisfy this condition.

If the character state 0 corresponds to the ancestral base at each site, then we can check for the existence of a rooted tree by verifying the *three-point condition*

$$\begin{aligned} &\text{For every pair of columns, not more than two} \\ &\text{of the patterns } 01, 10, 11 \text{ occur} \end{aligned} \tag{5.0.2}$$

Table 3. Segregating sites in a sample of mitochondrial sequences

Position	1 0 6	1 9 0	2 5 1	2 9 6	3 4 4	8	9	1 2 4	1 4 9	1 6 2	1 6 6	1 9 4	2 3 3	2 6 7	2 7 1	2 7 5	3 1 9	3 3 9	allele freqs.
Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
allele																			
a	A	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	2
b	A	G	G	A	A	T	C	C	T	T	T	T	C	T	C	T	T	C	2
c	G	A	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	1
d	G	G	A	G	A	C	C	C	C	C	T	T	C	C	C	T	T	C	3
e	G	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	19
f	G	G	G	A	G	T	C	C	T	C	T	T	C	T	C	T	T	C	1
g	G	G	G	G	A	C	C	C	T	C	C	C	C	C	C	T	T	T	1
h	G	G	G	G	A	C	C	C	T	C	C	C	T	C	C	T	T	T	1
i	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	C	T	4
j	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	T	T	8
k	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	C	5
l	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	4
m	G	G	G	G	A	C	C	T	T	C	T	T	C	C	C	T	T	C	3
n	G	G	G	G	A	C	T	C	T	C	T	T	C	C	T	T	T	C	1

Mitochondrial data from Ward *et al.* (1991). Variable purine and pyrimidine positions in the control region. Position 69 corresponds to position 16,092 in the human reference sequence published by Anderson *et al.* (1981)

It is known that if the most frequent type at each site is labelled 0 (ancestral), then the unrooted tree exists if and only if the rooted tree exists. Gusfield (1991) gives a $O(ns)$ time algorithm for finding a rooted tree:

Algorithm 5.1 Algorithm to find rooted tree for binary data matrix

1. Remove duplicate columns in the data matrix.
2. Consider each column as a binary number. Sort the columns into decreasing order, with the largest in column 1.
3. Construct paths from the leaves to the root in the tree by labelling nodes by mutation column labels and reading vertices in paths from right to left where 1s occur in rows.

Table 4. Distance between sequences for the Ward data

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*
a														
b	1													
c	6	7												
d	6	7	4											
e	1	2	5	5										
f	2	3	6	6	1									
g	7	8	3	5	6	7								
h	8	9	4	6	7	8	1							
i	7	8	3	5	6	7	4	5						
j	6	7	2	4	5	6	3	4	1					
k	4	5	2	2	3	4	3	4	3	2				
l	5	6	1	3	4	5	2	3	2	1	1			
m	5	6	3	3	4	5	4	5	4	3	1	2		
n	6	7	4	4	5	6	5	6	5	4	2	3	3	

Figure 5.1 shows the resulting rooted tree for the Ward data, and Figure 5.2 shows corresponding unrooted tree. Note that the distances between any two sequences in the tree is indeed given by the appropriate entry of the matrix in Table 4. We emphasize that these trees are equivalent representations of the original data matrix.

In this section we develop a stochastic model for the evolution of such trees, beginning with summary statistics such as the number of segregating sites seen in the data.

5.1 Measures of diversity in a sample

We begin our study by describing some simple measures of the amount of diversity seen in a sample of DNA sequences. For a sample of n sequences of length s base pairs, write $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{is})$ for the sequence of bases from sequence i, $1 \le i \le n$, and define $\Pi(i,j)$ to be the number of sites at which sequences i and j differ:

$$\Pi(i,j) = \sum_{l=1}^{s} \mathbb{1}(y_{il} \neq y_{jl}), \ i \neq j. \tag{5.1.1}$$

The *nucleotide diversity* Π_n in the sample is the mean pairwise difference defined by

$$\Pi_n = \frac{1}{n(n-1)} \sum_{i \neq j} \Pi(i,j), \tag{5.1.2}$$

and the per site nucleotide diversity is defined as

Fig. 5.1. Rooted tree for the Ward data found from Gusfield's algorithm

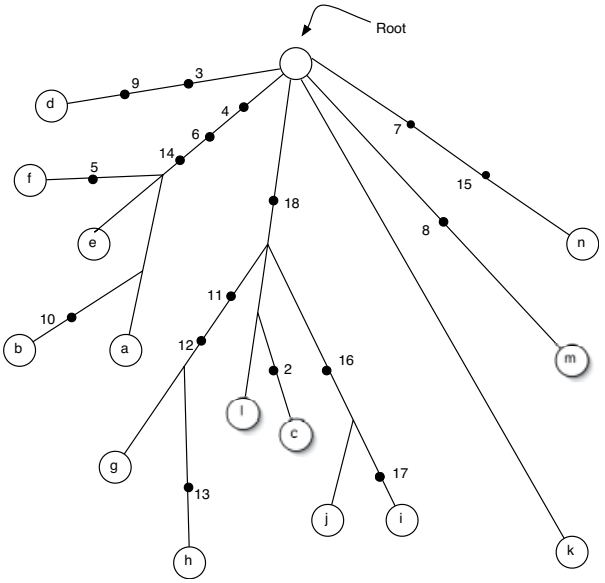
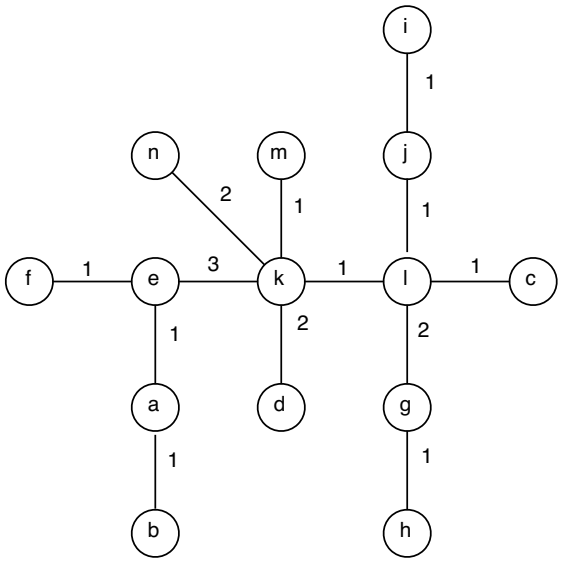


Fig. 5.2. Unrooted tree for the Ward data found from Figure 5.1. The numbers on the branches correspond to the number of sites on that branch.



$$\pi_n = \Pi_n / s.$$

Suppose that each position in the sequences being compared is from an alphabet $\mathcal{A}$ having α different letters (so that $\alpha = 4$ in the usual nucleotide alphabet), and write n_{la} for the number of times the letter a appears in site l in the sample. Then it is straightforward to show that

$$\Pi_n = \frac{1}{n(n-1)} \sum_{l=1}^{s} \sum_{a \in \mathcal{A}} n_{la}(n - n_{la}) := \frac{n}{n-1} \sum_{l=1}^{s} H_l, \tag{5.1.3}$$

where H_l is the heterozygosity at site l, defined by

$$H_l = \sum_{a \in \mathcal{A}} \frac{n_{la}}{n} \left(1 - \frac{n_{la}}{n}\right).$$

Thus, but for the correction factor $n/(n-1)$, the per site nucleotide diversity is just the average heterozygosity across the region; that is,

$$\pi_n = \frac{n}{n-1} \frac{1}{s} \sum_{l=1}^{s} H_l.$$

The sampling distribution of Π_n depends of course on the mutation mechanism that operates in the region. In the case of the infinitely-many-sites mutation model, we have

$$\begin{aligned} \mathbb{E}\Pi_n &= \frac{1}{n(n-1)} \sum_{i \neq j} \Pi(i,j) \\ &= \mathbb{E}\Pi(1,2) \qquad (\text{ by symmetry}) \\ &= \mathbb{E}(\#\text{ of segregating sites in sample of size 2}) \\ &= \theta \mathbb{E}(T_2), \end{aligned}$$

where T_2 is the time taken to find the MRCA of a sample of size two. In the case of constant population size, we have

$$\mathbb{E}\Pi_n = \theta. \tag{5.1.4}$$

The variance of Π_n was found by Tajima (1983), who showed that

$$\mathrm{Var}(\Pi_n) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2. \tag{5.1.5}$$

The nucleotide diversity statistic is a rather crude summary of the variability in the data. In the next section, we study pairwise difference curves.

5.2 Pairwise difference curves

The random variables $\Pi(i,j)$ are identically distributed, but they are of course not independent. Their common distribution can be found from the observation, exploited several times already, that

$$\mathbb{P}(\Pi(1,2)=k)=\mathbb{E}\mathbb{P}(\Pi(1,2)=k|T_2),$$

Conditional on T_2, $\Pi(1,2)$ has a Poisson distribution with parameter $2T_2\theta/2=\theta T_2$, so that for a population varying with rate function $\lambda(t)$,

$$\mathbb{P}(\Pi(1,2)=k)=\int_0^\infty e^{-\theta t}\frac{(\theta t)^k}{k!}\lambda(t)e^{-\Lambda(t)}dt. \tag{5.2.1}$$

In the case of a constant size, when $\lambda(t)=1$ and $\Lambda(t)=t$, the integral can be evaluated explicitly, giving

$$\mathbb{P}(\Pi(1,2)=k)=\frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^k,\ k=0,1,\ldots. \tag{5.2.2}$$

Thus $\Pi(1,2)$ has a geometric distribution with mean θ.

The *pairwise difference curve* is obtained by using the empirical distribution of the set $\Pi(i,j),1\le i\ne j\le n\}$ to estimate the probabilities in (5.2.1). Define

$$\Pi_{nk}=\frac{1}{n(n-1)}\sum_{i\ne j}\mathbb{1}(\Pi(i,j)=k), \tag{5.2.3}$$

the fraction of pairs of sequences separated by k segregating sites. By symmetry, we have

$$\mathbb{E}(\Pi_{nk})=\mathbb{P}(\Pi(1,2)=k),k=0,1,\ldots. \tag{5.2.4}$$

5.3 The number of segregating sites

The basic properties of the infinitely-many-sites model were found by Watterson (1975). Because each mutation is assumed to produce a new segregating site, the number of segregating sites observed in a sample is just the total number of mutations S_n since the MRCA of the sample. Conditional on L_n, S_n has a Poisson distribution with mean $\theta L_n/2$. We say that S_n has a *mixed Poisson distribution*, written $S_n\sim \mathrm{Po}(\theta L_n/2)$. It follows that

$$\begin{aligned}\mathbb{E}(S_n)&=\mathbb{E}(\mathbb{E}(S_n|L_n))\\&=\mathbb{E}\left(\theta L_n/2\right)\\&=\frac{\theta}{2}\sum_{j=2}^{n}j\frac{2}{j(j-1)}=\theta\sum_{j=1}^{n-1}\frac{1}{j}.\end{aligned} \tag{5.3.1}$$

Notice that for large n, $\mathbb{E}(S_n) \sim \theta \log(n)$.

We can write $S_n = Y_2 + \cdots + Y_n$ where Y_j is the number of mutations that arise while the sample has j ancestors. Since the T_j are independent, the Y_j are also independent. As above, Y_j has a mixed Poisson distribution, $\text{Po}(\theta j T_j/2)$. It follows that

$$\begin{aligned}\mathbb{E}(s^{Y_j}) &= \mathbb{E}(\mathbb{E}(s^{Y_j}|T_j)) \\ &= \mathbb{E}(\exp(-[\theta j T_j/2](1-s))) \\ &= \frac{j-1}{j-1+\theta(1-s)}, \end{aligned} \tag{5.3.2}$$

showing (Watterson (1975)) that Y_j has a geometric distribution with parameter $(j-1)/(j-1+\theta)$:

$$\mathbb{P}(Y_j = k) = \left(\frac{\theta}{\theta+j-1}\right)^k \left(\frac{j-1}{\theta+j-1}\right) \quad k = 0, 1, \ldots \tag{5.3.3}$$

Since the Y_j are independent for different j, it follows that

$$\text{Var}(S_n) = \sum_{j=2}^{n} \text{Var}(Y_j) = \theta \sum_{j=1}^{n-1} \frac{1}{j} + \theta^2 \sum_{j=1}^{n-1} \frac{1}{j^2}. \tag{5.3.4}$$

The probability generating function of S_n satisfies

$$\mathbb{E}(s^{S_n}) = \prod_{j=2}^{n} \mathbb{E}(s^{Y_j}) = \prod_{j=2}^{n} \frac{j-1}{j-1+\theta(1-s)} \tag{5.3.5}$$

from which further properties may be found. In particular, it follows from this that for $m = 0, 1, \ldots$

$$\mathbb{P}(S_n = m) = \frac{n-1}{\theta} \sum_{l=1}^{n-1} (-1)^{l-1} \binom{n-2}{l-1} \left(\frac{\theta}{l+\theta}\right)^{m+1}. \tag{5.3.6}$$

Estimating θ

It follows from (5.3.1) that

$$\theta_W = S_n \Big/ \sum_{j=1}^{n-1} \frac{1}{j} \tag{5.3.7}$$

is an unbiased estimator of θ. From (5.3.4) we see that the variance of θ_W is

$$\text{Var}(\theta_W) = \left[\theta \sum_{j=1}^{n-1} \frac{1}{j} + \theta^2 \sum_{j=1}^{n-1} \frac{1}{j^2}\right] \left[\sum_{j=1}^{n-1} \frac{1}{j}\right]^{-2}. \tag{5.3.8}$$

Notice that as $n \to \infty$, $\mathrm{Var}(\theta_W) \to 0$, so that the estimator θ_W is weakly consistent for θ.

An alternative estimator of θ is the moment estimator derived from (5.1.4), namely

$$\theta_T = \Pi_n. \tag{5.3.9}$$

The variance of θ_T follows immediately from (5.1.5). In fact, Π_n has a non-degenerate limit distribution as $n \to \infty$, so that θ_T cannot be consistent. This parallels the discussion in Section 3 about estimating θ on the basis of the number K_n of alleles or via the sample homozygosity F_n. The inconsistency of the pairwise estimators arises because these summary statistics lose a lot of information available in the sample.

We used the coalescent simulation algorithm to assess the properties of the estimators θ_W and θ_T for samples of size $n = 100$. The results of 10,000 simulations are given in Tables 5 and 6 for a variety of values of θ. It can be seen that the distribution of θ_W is much more concentrated than that of θ_T. Histograms comparing the two estimators appear in Figure 5.3.

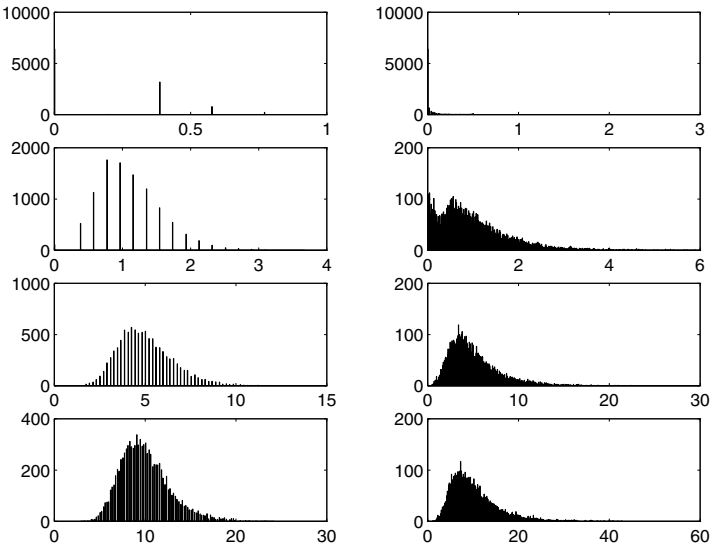
Table 5. Simulated properties of θ_W in samples of size $n = 100$.

	$\theta = 0.1$	$\theta = 1.0$	$\theta = 5.0$	$\theta = 10.0$
mean	0.18	1.10	5.03	9.99
std dev	0.23	0.48	1.53	2.75
median	0.00	0.97	4.83	9.66
5th %ile	0.00	0.39	2.90	6.18
95th %ile	0.39	1.93	7.73	15.07

Table 6. Simulated properties of θ_T in samples of size $n = 100$.

	$\theta = 0.1$	$\theta = 1.0$	$\theta = 5.0$	$\theta = 10.0$
mean	0.10	1.00	4.95	9.97
std dev	0.19	0.75	2.65	4.98
median	0.00	0.84	4.35	8.91
5th %ile	0.00	0.08	1.79	4.13
95th %ile	0.40	2.42	10.16	19.48

Fig. 5.3. Histograms of 10,000 replicates of estimators of θ based on samples of size $n = 100$. Left hand column is θ_W, right hand column is θ_T. First row corresponds to $\theta = 0.1$, second to $\theta = 1.0$, third to $\theta = 5.0$, and fourth to $\theta = 10.0$.



How well can we do?

The estimators θ_W and θ_T are based on summary statistics of the original sequence data. It is of interest to know how well these unbiased estimators might in principle behave. In this section, we examine this question in more detail for the case of constant population size.

If we knew how many mutations had occurred on each of the j branches of length T_j, $j = 2, \ldots, n$ in the coalescent tree, then we could construct a simple estimator of θ using standard results for independent random variables. Let $Y_{jk}, k = 1, \ldots, j; j = 2, \ldots, n$ denote the number of mutations on the k^{th} branch of length T_j and set $Y_j = \sum_{k=1}^{j} Y_{jk}$. Y_j is the observed number of mutations that occur during the time the sample has j distinct ancestors. Since each mutation produces a new segregating site, this is just the number of segregating sites that arise during this time. Since the T_j are independent, so too are the Y_j. We have already met the distribution of Y_j in equation (5.3.3), and it follows that the likelihood for observations Y_j, $j = 2, \ldots, n$ is

$$L_n(\theta) = \prod_{j=2}^{n} \left(\frac{\theta}{j-1+\theta}\right)^{Y_j} \left(\frac{j-1}{j-1+\theta}\right)$$
$$= \theta^{S_n}(n-1)! \prod_{j=2}^{n} (j-1+\theta)^{-(Y_j+1)},$$

where $S_n = \sum_{j=2}^{n} Y_j$ is the number of segregating sites. The maximum likelihood estimator based on this approach is therefore the solution of the equation

$$\theta = S_n \Big/ \sum_{j=2}^{n} \frac{Y_j + 1}{j-1+\theta}. \tag{5.3.10}$$

Furthermore,

$$\frac{\partial^2 \log L_n}{\partial \theta^2} = -\frac{S_n}{\theta^2} + \sum_{j=2}^{n} \frac{(Y_j+1)}{(j-1+\theta)^2},$$

so that

$$-\mathbb{E}\left(\frac{\partial^2 \log L_n}{\partial \theta^2}\right) = \frac{\theta \sum_1^{n-1} \frac{1}{j}}{\theta^2} - \sum_{j=2}^{n} \left(\frac{\theta}{j-1} + 1\right) \frac{1}{(j-1+\theta)^2}$$
$$= \frac{1}{\theta} \sum_1^{n-1} \frac{1}{j} - \sum_1^{n-1} \frac{1}{j(j+\theta)}$$
$$= \frac{1}{\theta} \sum_1^{n-1} \frac{1}{j+\theta} \tag{5.3.11}$$

Hence the variance of unbiased estimators θ_U of θ satisfies

$$\mathrm{Var}(\theta_U) \geq \theta \Big/ \sum_1^{n-1} \frac{1}{j+\theta}\ ,$$

as shown by Fu and Li (1993). The right-hand side is also the large-sample variance of the estimator θ_F in (5.3.10).

How does this bound compare with that in (5.3.8)? Certainly

$$\mathrm{Var}(\theta_F) \leq \mathrm{Var}(\theta_W), \tag{5.3.12}$$

and we can see that if θ is fixed and $n \to \infty$ then

$$\frac{\mathrm{Var}(\theta_F)}{\mathrm{Var}(\theta_W)} \to 1.$$

If, on the other hand, n is fixed and θ is large, we see that

$$\frac{\mathrm{Var}(\theta_F)}{\mathrm{Var}(\theta_W)} \to \left(\sum_1^{n-1} \frac{1}{j}\right)^2 \Big/ (n-1) \sum_1^{n-1} \frac{1}{j^2}\ ,$$

so that there can be a marked decrease in efficiency in using the estimator θ_W when θ is large. We cannot, of course, determine the numbers Y_j from data; this is more information than we have in practice. However, it does suggest that we explore the MLE of θ using the likelihoods formed from the full data rather than summary statistics. Addressing this issue leads us to study the underlying tree structure of infinitely-many-sites data in more detail, as well as to develop some computational algorithms for computing MLEs.

5.4 The infinitely-many-sites model and the coalescent

The infinitely-many-sites model is an early attempt to model the evolution of a completely linked sequence of sites in a DNA sequence. The term 'completely linked' means that no recombination is allowed. Each mutation on the coalescent tree of the sample introduces a mutant base at a site that has not previously experienced a mutation. One formal description treats the type of an individual as an element $(x_1, x_2, \ldots)$ of $E = \cup_{r\geq 1}[0,1]^r$. If a mutation occurs in an offspring of an individual of type $(x_1, x_2, \ldots, x_r)$, then the offspring has type $(x_1, x_2, \ldots, x_r, U)$, where U is a uniformly distributed random variable independent of the past history of the process.

Figure 3.1 provides a trajectory of the process. It results in a sample of five sequences, their types being (U_1, U_2), (U_1, U_2), (U_1, U_2, U_4, U_5), (U_0, U_3), (U_0, U_3) respectively.

There are several other ways to represent such sequences, of which we mention just one. Consider the example above once more. Each sequence gives a mutational path from the individual back to the most recent common ancestor of the sample. We can think of these as labels of locations at which new mutant sites have been introduced. In this sample there are six such sites, each resulting in a new segregating site. We can therefore represent the sequences as strings of 0s and 1s, each of length six. At each location, a 1 denotes a mutant type and a 0 the original or 'wild' type. Arbitrarily labelling the sites 1, 2, ..., 6 corresponding to the mutations at $U_0, U_1, \ldots, U_5$, we can write the five sample sequences as

$$\begin{aligned}(U_1, U_2, U_4, U_5) &= 011011\\ (U_1, U_2) &= 011000\\ (U_1, U_2) &= 011000\\ (U_0, U_3) &= 100100\\ (U_0, U_3) &= 100100\end{aligned}$$

These now look more like aligned DNA sequences! Of course, in reality we do not know which type at a given segregating site is ancestral and which is mutant, and the ordering of sites by time of mutation is also unknown.

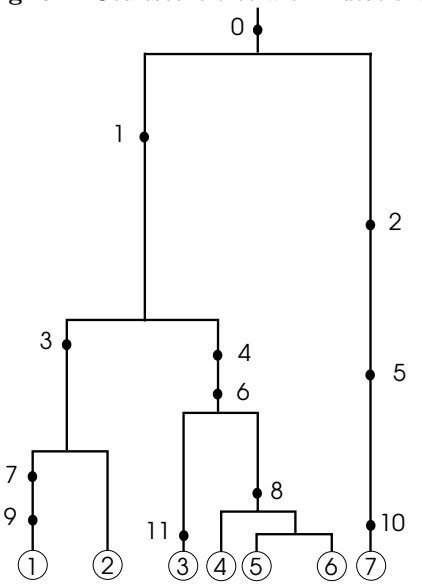
5.5 The tree structure of the infinitely-many-sites model

We have just seen that in the infinitely-many-sites model, each gene can be thought of as an infinite sequence of completely linked sites, each labelled 0 or 1. A 0 denotes the ancestral (original) type, and a 1 the mutant type. The mutation mechanism is such that a mutant offspring gets a mutation at a single new site that has never before seen a mutation. This changes the 0 to a 1 at that site, and introduces another segregating site into the sample. By way of example, a sample of 7 sequences might have the following structure:

gene 1 . . . 1 0 1 0 0 0 1 0 1 0 0 . . .
gene 2 . . . 1 0 1 0 0 0 0 0 0 0 0 . . .
gene 3 . . . 1 0 0 1 0 1 0 0 0 0 1 . . .
gene 4 . . . 1 0 0 1 0 1 0 1 0 0 0 . . .
gene 5 . . . 1 0 0 1 0 1 0 1 0 0 0 . . .
gene 6 . . . 1 0 0 1 0 1 0 1 0 0 0 . . .
gene 7 . . . 0 1 0 0 1 0 0 0 0 1 0 . . .

the dots indicating non-segregating sites. Many different coalescent trees can give rise to a given set of sequences. Figure 5.4 shows one of them.

Fig. 5.4. Coalescent tree with mutations



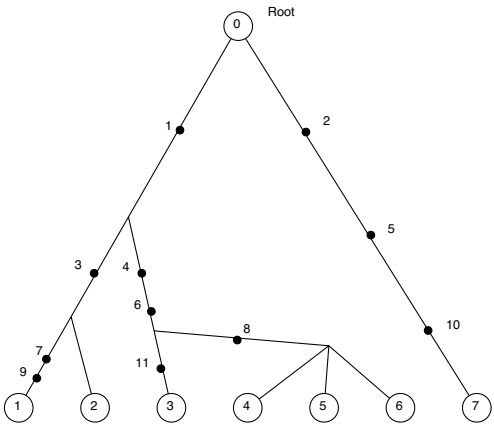
The coalescent tree with mutations can be condensed into a genealogical tree with no time scale by labelling each sequence by a list of mutations up

to the common ancestor. For the example in Figure 5.4, the sequences may be represented as follows:

gene 1 (9,7,3,1,0)
gene 2 (3,1,0)
gene 3 (11,6,4,1,0)
gene 4 (8,6,4,1,0)
gene 5 (8,6,4,1,0)
gene 6 (8,6,4,1,0)
gene 7 (10,5,2,0)

The condensed genealogical tree is shown in Figure 5.5. The leaves in the tree

Fig. 5.5. Genealogical tree corresponding to Figure 5.4



are the tips, corresponding to the sequences in the sample. The branches in the tree are the internal links between different mutations. The 0s in each sequence are used to indicate that the sequences can be traced back to a common ancestor.

Thus we have three ways to represent the sequences in the sample: (i) as a list of paths from the sequence to the root; (ii) as a *rooted* genealogical tree; and (iii) as a matrix with entries in $\{0, 1\}$ where a 0 corresponds to the ancestral type at a site, and a 1 the mutant type. In our example, the 0-1 matrix given above is equivalent to the representations in Figures 5.4 and 5.5. Finally, the number of segregating sites is precisely the number of mutations in the tree. In the next section, we discuss the structure of these tree representations in more detail.

5.6 Rooted genealogical trees

Following Ethier and Griffiths (1987), we think of the ith gene in the sample as a sequence $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots)$ where each $x_{ij} \in \mathbb{Z}_+$. (In our earlier parlance, the type space E of a gene is the space $\mathbb{Z}_+^\infty$.) It is convenient to think of $x_{i0}, x_{i1}, \ldots$ as representing the most recently mutated site, the next most recently, and so on. A sample of n genes may therefore be represented as n sequences $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$. The assumption that members of the sample have an ancestral tree and that mutations never occur at sites that have previously mutated imply that the sequences $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ satisfy:

(1) Coordinates within each sequence are distinct
(2) If for some $i, i' \in \{1, \ldots, n\}$ and $j, j' \in \mathbb{Z}_+$ we have $x_{ij} = x_{i'j'}$, then $x_{i,j+k} = x_{i',j'+k},\ k = 1, 2, \ldots$
(3) there is a coordinate common to all n sequences.

Rules (2) and (3) above say that the part of the sequences inherited from the common ancestor appears at the right-hand end of the sequences. In practice we can discard from each $\boldsymbol{x}$ sequence those entries that are common to all of the sequences in the sample; these are the coordinates after the value common to all the sequences. It is the segregating sites, and not the non-segregating sites, that are important to us. In what follows, we use these representations interchangeably.

Trees are called *labelled* if the sequences (tips) are labelled. Two labelled trees are identical if there is a renumbering of the sites that makes the labelled trees the same. More formally, let $\mathcal{T}_n = \{(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is a tree$\}$. Define an equivalence relation $\sim$ by writing $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \sim (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ if there is a bijection $\xi : \mathbb{Z}_+ \to \mathbb{Z}_+$ with $y_{ij} = \xi(x_{ij}), i = 1, \ldots, n,\ j = 0, 1, \ldots$. Then $\mathcal{T}_n/\sim$ corresponds to labelled trees. Usually, we do not distinguish between an equivalence class and a typical member.

An *ordered labelled* tree is one where the sequences are labelled, and considered to be in a particular order. Visually this corresponds to a tree diagram with ordered leaves. An *unlabelled* (and so unordered) tree is a tree where the sequences are not labelled. Visually two unlabelled trees are identical if they can be drawn identically by rearranging the leaves and corresponding paths in one of the trees. Define a second equivalence relation $\approx$ by $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \approx (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ if there is a bijection $\xi : \mathbb{Z}_+ \to \mathbb{Z}_+$ and a permutation σ of $1, 2, \ldots, n$ such that $y_{\sigma(i),j} = \xi(x_{ij}),\ i = 1, \ldots, n,\ j = 0, 1, \ldots$. Then $\mathcal{T}_n/\approx$ corresponds to unlabelled trees.

Usually trees are unlabelled, with sequences and sites then labelled for convenience. However it is easiest to deal with ordered labelled trees in a combinatorial and probabilistic sense, then deduce results about unlabelled trees from the labelled variety. Define

$$(\mathcal{T}_d/\sim)_0 = \{T \in \mathcal{T}_d/\sim:\ \boldsymbol{x}_1, \ldots, \boldsymbol{x}_d \text{ all distinct}\}$$

and similarly for $(\mathcal{T}_d/\approx)_0$. $T \in \cup_{d\geq 1}(\mathcal{T}_d/\sim)_0$ corresponds to the conventional graph theoretic tree, with multiple tips removed. There is a one-to-one correspondence between trees formed from the sequences and binary sequences of sites. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d$ be distinct sequences of sites satisfying (1), (2) and (3), and let $\mathcal{I}$ be the incidence matrix of segregating sites. If $u_1, \ldots, u_k$ are the segregating sites (arranged in an arbitrary order) then

$$\mathcal{I}_{ij} = 1 \text{ if } u_j \in \boldsymbol{x}_i,\ i = 1, \ldots, d,\ j = 1, \ldots, k.$$

The sites which are not segregating do not contain information about the tree.

Deducing the tree from a set of d binary sequences is not a priori simple, because sites where mutations occur are unordered with respect to time and any permutation of the columns of $\mathcal{I}$ produces the same tree. In addition, binary data often have unknown ancestral labelling, adding a further complication to the picture. However, these trees are equivalent to the rooted trees discussed in the introduction. It follows that we can use the three-point condition in (5.0.2) to check whether a matrix of segregating sites is consistent with this model, and if it is, we can reconstruct the tree using Gusfield's algorithm 5.1. We turn now to computing the distribution of such a rooted tree.

5.7 Rooted genealogical tree probabilities

Let $p(T, \boldsymbol{n})$ be the probability of obtaining the alleles $T \in (\mathcal{T}_d/\sim)_0$ with multiplicities $\boldsymbol{n} = (n_1, \ldots, n_d)$ and let $n = \sum_1^d n_i$. This is the probability of getting a particular *ordered* sample of distinct sequences with the indicated multiplicities. Ethier and Griffiths (1987) and Griffiths (1989) established the following:

Theorem 5.1 *$p(T, \boldsymbol{n})$ satisfies the equation*

$$\begin{aligned} n(n-1+\theta)p(T,\boldsymbol{n}) = &\sum_{k:n_k\geq 2} n_k(n_k-1)p(T,\boldsymbol{n}-\boldsymbol{e}_k) \\ &+ \theta \sum_{\substack{k:n_k=1,\ x_{k0}\\ \mathcal{S}\boldsymbol{x}_k\neq\boldsymbol{x}_j\ \forall\ j}}\ \text{distinct,} p(\mathcal{S}_k T,\boldsymbol{n}) \qquad (5.7.1) \\ &+ \theta \sum_{\substack{k:n_k=1,\\ x_{k0}\ \text{distinct.}}} \sum_{j:\mathcal{S}\boldsymbol{x}_k=\boldsymbol{x}_j} p(\mathcal{R}_k T, \mathcal{R}_k(\boldsymbol{n}+\boldsymbol{e}_j)). \end{aligned}$$

In equation (5.7.1), $\boldsymbol{e}_j$ is the jth unit vector, $\mathcal{S}$ is a shift operator which deletes the first coordinate of a sequence, $\mathcal{S}_k T$ deletes the first coordinate of the k^{th} sequence of T, $\mathcal{R}_k T$ removes the k^{th} sequence of T, and 'x_{k0} distinct' means that $x_{k0} \neq x_{ij}$ for all $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$ and $(i,j) \neq (k,0)$. The boundary condition is $p(T_1, (1)) = 1$.

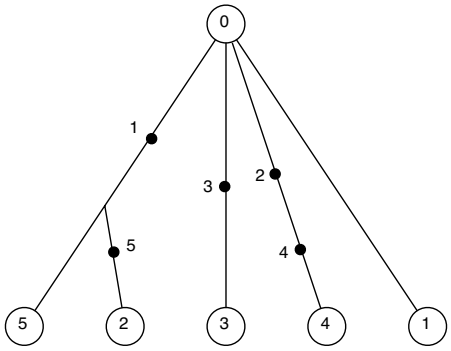
Remark. The system (5.7.1) is recursive in the quantity $\{n - 1 +$ number of vertices in $T\}$.

Proof. Equation (5.7.1) can be validated by a simple coalescent argument, by looking backwards in time for the first event in the ancestry of the sample. The first term on the right of (5.7.1) corresponds to a coalescence occurring first. This event has probability $(n-1)/(\theta+n-1)$. For any k with $n_k \geq 2$, the two individuals who coalesce may come from an allele with n_k copies, and the tree after the coalescence would be $(T, \boldsymbol{n}-\boldsymbol{e}_k)$. The contribution to $(T, \boldsymbol{n})$ form events of this sort is therefore

$$\frac{n-1}{\theta+n-1} \sum_{k:n_k \geq 2} \frac{n_k}{n} \left(\frac{n_k - 1}{n-1} \right) p(T, \boldsymbol{n} - \boldsymbol{e}_k).$$

The second terms on the right of (5.7.1) correspond to events where a mutation occurs first. Suppose then that the mutation gave rise to sequence $\boldsymbol{x}_k$. There are two different cases to consider, these being determined by whether or not the sequence $\mathcal{S}\boldsymbol{x}_k$ that resulted in $\boldsymbol{x}_k$ is already in the sample, or not. These two cases are illustrated in the tree in Figure 5.6. The sequences are

Fig. 5.6. Representative tree



$$\begin{aligned}
\boldsymbol{x}_1 &= (0) \\
\boldsymbol{x}_2 &= (5\ 1\ 0) \\
\boldsymbol{x}_3 &= (3\ 0) \\
\boldsymbol{x}_4 &= (2\ 4\ 0) \\
\boldsymbol{x}_5 &= (1\ 0)
\end{aligned}$$

Note that $\mathcal{S}\boldsymbol{x}_2 = (1\ 0) = \boldsymbol{x}_5$, so the ancestral type of $\boldsymbol{x}_2$ is in the sample. This corresponds to the third term on the right of (5.7.1). On the other hand, $\mathcal{S}\boldsymbol{x}_4 = (4\ 0)$, a type not now in the sample. This corresponds to second term on the right of (5.7.1). The phrase 'x_{k0} distinct' that occurs in these two sums is

required because not all leaves with $n_k = 1$ can be removed; some cannot have arisen in the evolution of the process. The sequence $\boldsymbol{x}_5$ provides an example.

Combining these probabilities gives a contribution to $p(T, \boldsymbol{n})$ of

$$\frac{\theta}{\theta + n - 1}\left\{\sum \frac{1}{n} p(\mathcal{S}_k T, \boldsymbol{n}) + \sum\sum \frac{1}{n} p(\mathcal{R}_k T, \mathcal{R}_k(\boldsymbol{n} + \boldsymbol{e}_j))\right\},$$

and completes the proof. □

It is sometimes more convenient to consider the recursion satisfied by the quantities $p^0(T, \boldsymbol{n})$ defined by

$$p^0(T, \boldsymbol{n}) = \frac{n!}{n_1! \ldots n_d!} p(T, \boldsymbol{n}). \tag{5.7.2}$$

$p^0(T, \boldsymbol{n})$ is the probability of the labelled tree T, without regard to the order of the sequences in the sample. Using (5.7.1), this may be written in the form

$$\begin{aligned} n(n - 1 + \theta) p^0(T, \boldsymbol{n}) = & \sum_{k: n_k \geq 2} n(n_k - 1) p^0(T, \boldsymbol{n} - \boldsymbol{e}_k) \\ & + \theta \sum_{\substack{k: n_k = 1, \ x_{k0} \\ \mathcal{S}\boldsymbol{x}_k \neq \boldsymbol{x}_j \ \forall \ j}} \text{distinct}, \quad p^0(\mathcal{S}_k T, \boldsymbol{n}) \qquad (5.7.3) \\ & + \theta \sum_{\substack{k: n_k = 1, \\ x_{k0} \text{ distinct.}}} \sum_{j: \mathcal{S}\boldsymbol{x}_k = \boldsymbol{x}_j} (n_j + 1)\, p^0(\mathcal{R}_k T, \mathcal{R}_k(\boldsymbol{n} + \boldsymbol{e}_j)). \end{aligned}$$

Let $p^*(T, \boldsymbol{n})$ be the probability of a corresponding unlabelled tree with multiplicity of the sequences given by $\boldsymbol{n}$. p^* is related to p^0 by a combinatorial factor, as follows. Let S_d denote the set of permutations of $(1, \ldots, d)$. Given a tree T and $\sigma \in S_d$, define $T_\sigma = (\boldsymbol{x}_{\sigma(1)}, \ldots, \boldsymbol{x}_{\sigma(d)})$ and $\boldsymbol{n}_\sigma = (n_{\sigma(1)}, \ldots, n_{\sigma(d)})$. Letting

$$a(T, \boldsymbol{n}) = |\{\sigma \in S_d : T_\sigma = T, \boldsymbol{n}_\sigma = \boldsymbol{n}\}|, \tag{5.7.4}$$

we have

$$p^*(T, \boldsymbol{n}) = \frac{1}{a(T, \boldsymbol{n})} p^0(T, \boldsymbol{n}). \tag{5.7.5}$$

Informally, the number of distinct ordered labelled trees corresponding to the unlabelled tree is

$$\frac{n!}{n_1! \cdots n_d! a(T, \boldsymbol{n})}.$$

In the tree shown in Figure 5.5, $a(T, \boldsymbol{n}) = 1$. A subsample of three genes $(9, 7, 3, 1, 0), (11, 6, 4, 1, 0), (10, 5, 2, 0)$, forming a tree T' with frequencies $\boldsymbol{n}' = (1, 1, 1)$, has $a(T', \boldsymbol{n}') = 2$, because the first two sequences are equivalent in an unlabelled tree.

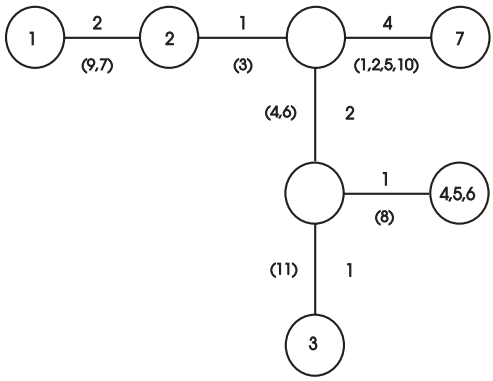
These recursions may be solved for small trees, and the resulting genealogical tree probabilities used to estimate θ by true maximum likelihood methods.

One drawback is that the method depends on knowing the ancestral type at each site, an assumption rarely met in practice. We turn now to the tree structure that underlies the process when the ancestral labelling is unknown.

5.8 Unrooted genealogical trees

When the ancestral base at each site is unknown there is an *unrooted* genealogical tree that corresponds to the sequences. In these unrooted trees, the vertices represent sequences and the number of mutations between sequences are represented by numbers along the edges; see Griffiths and Tavaré (1995). It is convenient to label the vertices to show the sequences they represent. The unrooted tree for the example sequences is shown in Figure 5.7.

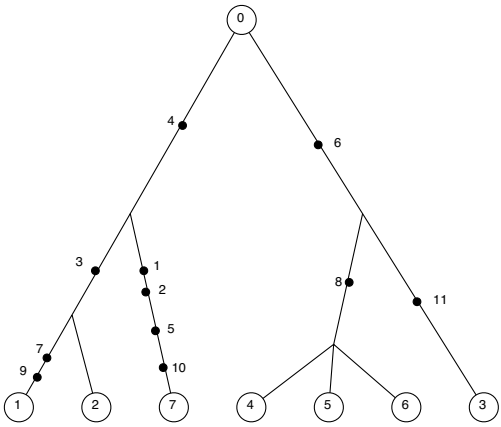
Fig. 5.7. Unrooted genealogical tree corresponding to Figure 5.4



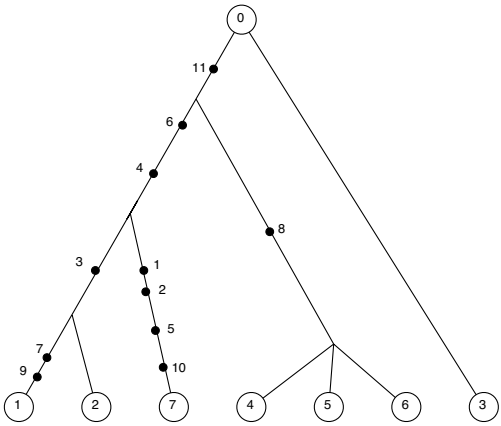
Given a single rooted tree, the unrooted genealogy can be found. The constructive way to do this is to put potential ancestral sequences at the nodes in the rooted tree (ignoring the root). There are three such nodes in the example in Figure 5.5. The ancestral sequence might be represented in the sample (as with sequence 2 in that figure), or it may be an inferred sequence not represented in the sample.

Given a rooted genealogy, we have seen how the corresponding unrooted tree can be found. Conversely, the class of rooted trees produced from an unrooted genealogy may be constructed by placing the root at one of the sequences, or between mutations along an edge. This corresponds to picking up the unrooted tree at that point and shaking it. Two examples are given in Figure 5.8. In the first, the root corresponds to the third sequence, and in the second it is between the two mutations between the two inferred sequences. The unrooted tree constructed from any of these rooted trees is of course unique.

Fig. 5.8. Moving the root



Tree with root between mutations



Tree with root the third sequence

If there are α sequences (including the inferred sequences), with $m_1, m_2, \ldots$ mutations along the edges, and s segregating sites, then there are

$$\alpha + \sum_j (m_j - 1) = s + 1 \tag{5.8.1}$$

rooted trees when the sequences are labelled. There may be fewer unlabelled rooted trees, as some can be identical after unlabelling the sequences. In the example there are 11 segregating sites, and so 12 labelled rooted trees, which correspond to distinct unlabelled rooted trees as well.

The class of rooted trees corresponds to those constructed from toggling the ancestor labels 0 and 1 at sites. The number of the 2^s possible relabellings that are consistent with the sequences having come from a tree is

$$\alpha + \sum_j \sum_{k=1}^{m_j - 1} \binom{m_j}{k} = \alpha + \sum_j (2^{m_j} - 2). \tag{5.8.2}$$

This follows from the observation that if there is a collection of m segregating sites which correspond to mutations between sequences, then the corresponding data columns of the 0-1 sequences (with 0 the ancestral state) are identical or complementary. Any of the $\binom{m}{k}$ configurations of k identical and $m-k$ complementary columns correspond to the same labelled tree with a root placed after the kth mutation. The correspondence between different rooted labelled trees and the matrix of segregating sites can be described as follows: in order to move the root from one position to another, toggle those sites that occur on the branches between the two roots.

The upper tree in Figure 5.8 has incidence matrix

gene 1 0 0 1 1 0 0 1 0 1 0 0
gene 2 0 0 1 1 0 0 0 0 0 0 0
gene 3 0 0 0 0 0 1 0 0 0 0 1
gene 4 0 0 0 0 0 1 0 1 0 0 0
gene 5 0 0 0 0 0 1 0 1 0 0 0
gene 6 0 0 0 0 0 1 0 1 0 0 0
gene 7 1 1 0 1 1 0 0 0 0 1 0

whereas the lower tree in Figure 5.8 has incidence matrix

gene 1 0 0 1 1 0 1 1 0 1 0 1
gene 2 0 0 1 1 0 1 0 0 0 0 1
gene 3 0 0 0 0 0 0 0 0 0 0 0
gene 4 0 0 0 0 0 0 0 1 0 0 1
gene 5 0 0 0 0 0 0 0 1 0 0 1
gene 6 0 0 0 0 0 0 0 1 0 0 1
gene 7 1 1 0 1 1 1 0 0 0 1 1

It can readily be checked that the sites between the two roots are those numbered 6 and 11, and if these are toggled then one tree is converted into the other.

5.9 Unrooted genealogical tree probabilities

A labelled unrooted genealogical tree of a sample of sequences has a vertex set V which corresponds to the labels of the sample sequences and any inferred sequences in the tree. Let $\boldsymbol{Q}$ be the edges of the tree, described by $(m_{ij}, i, j \in V)$, where m_{ij} is the number of mutations between vertices i and

j. Let $\boldsymbol{n}$ denote the multiplicities of the sequences. It is convenient to include the inferred sequences $\ell \in V$ with $n_\ell = 0$. Then the unrooted genealogy is described by $(\boldsymbol{Q}, \boldsymbol{n})$.

Define $p(\boldsymbol{Q}, \boldsymbol{n})$, $p^0(\boldsymbol{Q}, \boldsymbol{n})$, $p^*(\boldsymbol{Q}, \boldsymbol{n})$ analogously to the probabilities for T. The combinatorial factor relating $p^*(\boldsymbol{Q}, \boldsymbol{n})$ and $p^0(\boldsymbol{Q}, \boldsymbol{n})$ is

$$a(\boldsymbol{Q}, \boldsymbol{n}) = |\{\sigma \in S_{|V|} : \boldsymbol{Q}_\sigma = \boldsymbol{Q}, \boldsymbol{n}_\sigma = \boldsymbol{n}\}|. \tag{5.9.1}$$

The quantities $p(\boldsymbol{Q}, \boldsymbol{n})$ and $p^0(\boldsymbol{Q}, \boldsymbol{n})$ satisfy recursions similar to (5.7.1) and (5.7.3), which can be derived by considering whether the last event back in time was a coalescence or a mutation. The recursion for $p(\boldsymbol{Q}, \boldsymbol{n})$ is

$$\begin{aligned} n(n-1+\theta)p(\boldsymbol{Q}, \boldsymbol{n}) = & \sum_{k: n_k \geq 2} n_k(n_k - 1)p(\boldsymbol{Q}, \boldsymbol{n} - \boldsymbol{e}_k) \\ & + \theta \sum_{\substack{k: n_k = 1,\ |k|=1, \\ k \to j,\ m_{kj} > 1}} p(\boldsymbol{Q} - \boldsymbol{e}_{kj}, \boldsymbol{n}) \\ & + \theta \sum_{\substack{k: n_k = 1,\ |k|=1, \\ k \to j,\ m_{kj} = 1}} p(\boldsymbol{Q} - \boldsymbol{e}_{kj}, \boldsymbol{n} + \boldsymbol{e}_j - \boldsymbol{e}_k), \end{aligned} \tag{5.9.2}$$

where $|k| = 1$ means that the degree of the vertex k is 1 (that is, k is a leaf), and $k \to j$ means that vertex k is joined to vertex j. In the last term on the right of (5.9.2), vertex k is removed from $\boldsymbol{Q}$. The boundary conditions in (5.9.2) for $n = 2$ are

$$p((0), 2\boldsymbol{e}_1) = \frac{1}{1+\theta},$$

and

$$p((m), \boldsymbol{e}_1 + \boldsymbol{e}_2) = \left(\frac{\theta}{1+\theta}\right)^m \frac{1}{1+\theta},\ m = 1, 2, \ldots.$$

The probability of a labelled unrooted genealogical tree $\boldsymbol{Q}$ is

$$p(\boldsymbol{Q}, \boldsymbol{n}) = \sum_{T \in C(\boldsymbol{Q})} p(T, \boldsymbol{n}), \tag{5.9.3}$$

where $C(\boldsymbol{Q})$ is the class of distinct labelled rooted trees constructed from $\boldsymbol{Q}$. The same relationship holds in (5.9.3) if p is replaced by p^0.

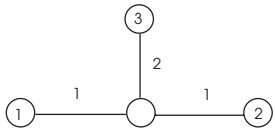
5.10 A numerical example

In this example we suppose that the ancestral states are unknown, and that the sequences, each with multiplicity unity, are:

1 0 0 0
0 0 0 1
0 1 1 0

For convenience, label the segregating sites 1, 2, 3, and 4 from the left. When 0 is the ancestral state, a possible rooted tree for these sequences has paths to the root of $(1,0)$, $(2,3,0)$, and $(4,0)$. It is then straightforward to construct the corresponding unrooted genealogy, which is shown in Figure 5.9. The central sequence is inferred. There are five possible labelled rooted trees

Fig. 5.9. Unrooted Genealogy



constructed from the unrooted genealogy, corresponding to the root being at one of the sequences, or between the two mutations on the edge. These five trees are shown in Figure 5.10, together with their probabilities $p(T, \boldsymbol{n})$, computed exactly from the recursion (5.7.1) when $\theta = 2.0$. $p(\boldsymbol{Q}, \boldsymbol{n})$ is the sum of these probabilities, 0.004973. The factor in (5.9.1) is 2, and the multinomial coefficient $3!/1!1!1! = 6$ so $p^*(\boldsymbol{Q}, \boldsymbol{n}) = 3 \times 0.00497256 = 0.014919$. Note that the trees (b) and (e) are identical unlabelled rooted trees, but are distinct labelled rooted trees, so are both counted in calculating $p^*(\boldsymbol{Q}, \boldsymbol{n})$.

In this small genealogy, the coalescent trees with four mutations can be enumerated to find the probability of the genealogy. The trees which produce the tree in Figure 5.9 are shown in Figure 5.11, with the correspondence to the trees in Figure 5.10 highlighted.

Let T_3 be the time during which the sample has three ancestors, and T_2 the time during which it has two. T_3 and T_2 are independent exponential random variables with respective rates 3 and 1. By considering the Poisson nature of the mutations along the edges of the coalescent tree, the probability of each type of tree can be calculated. For example, the probability $p_{(a1)}$ of the first tree labelled $(a1)$ is

$$\begin{aligned} p_{(a1)} &= \mathbb{E}\left[\left(e^{-\theta T_3/2}\frac{\theta T_3}{2}\right)^2 \; e^{-\theta T_2/2}\; e^{-\theta(T_2+T_3)/2}\frac{1}{2!}(\theta(T_2+T_3)/2)^2\right] \\ &= \frac{\theta^4}{32}\mathbb{E}\left[e^{-\theta(3T_3/2+T_2)}T_3^2(T_2+T_3)^2\right] \\ &= \frac{\theta^4(17\theta^2+46\theta+32)}{27(\theta+1)^3(\theta+2)^5}. \end{aligned}$$

In a similar way the other tree probabilities may be calculated. We obtain

Fig. 5.10. Labelled rooted tree probabilities

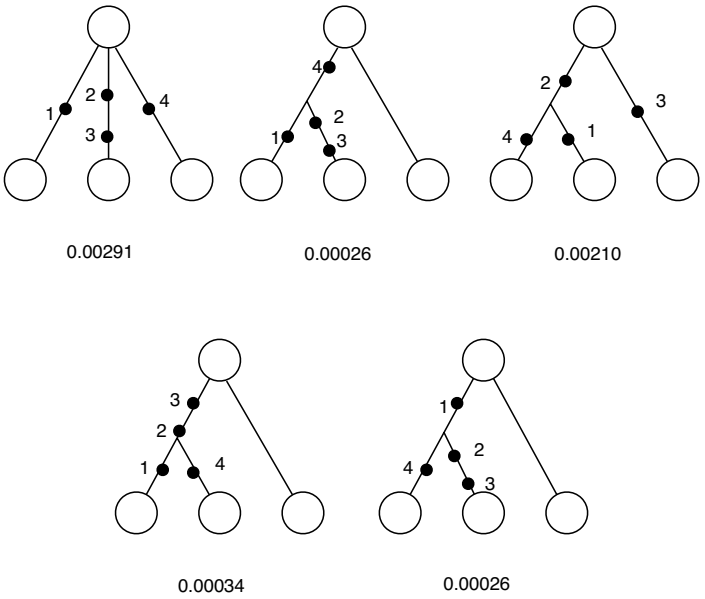
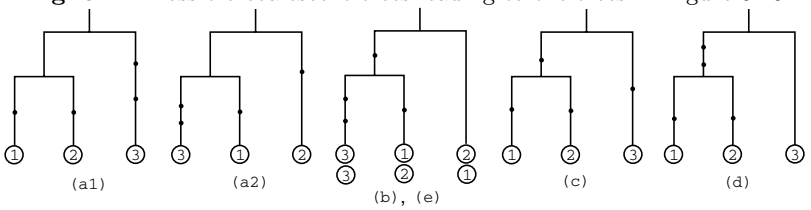


Fig. 5.11. Possible coalescent trees leading to the trees in Figure 5.10



$$
\begin{aligned}
p_{(a2)} &= \frac{\theta^4}{16}\mathbb{E}\left[2e^{-\theta(3T_3/2+T_2)}T_3^3(T_2+T_3)/2\right] \\
&= \frac{2\theta^4(11\theta+14)}{27(\theta+1)^2(\theta+2)^5}, \\
p_{(b)} = p_{(e)} &= \frac{\theta^4}{16}\mathbb{E}\left[e^{-\theta(3T_3/2+T_2)}T_3^3T_2/2\right] \\
&= \frac{\theta^4}{9(\theta+1)^2(\theta+2)^4}, \\
p_{(c)} &= \frac{\theta^4}{16}\mathbb{E}\left[e^{-\theta(3T_3/2+T_2)}(T_2+T_3)T_3^2T_2\right] \\
&= \frac{\theta^4(2\theta+3)}{9(\theta+1)^3(\theta+2)^4}, \\
p_{(d)} &= \frac{\theta^4}{16}\mathbb{E}\left[e^{-\theta(3T_3/2+T_2)}T_3^2T_2^2/2\right] \\
&= \frac{2\theta^4}{9(\theta+1)^3(\theta+2)^3}.
\end{aligned}
$$

Note that there are two coalescent trees that correspond to case $(a2)$, depending on whether 1 coalesced with 3 first, or 2 did. When $\theta = 2$, these probabilities reduce to $p_{(a1)} = 0.004115, p_{(a2)} = 0.004630, p_{(b),(e)} = 0.000772, p_{(c)} = 0.003601, p_{(d)} = 0.001029$. From these we deduce that $p(T(a), \boldsymbol{n}) = (0.004115 + 0.004630)/3 = 0.002915$, $p(T(b), \boldsymbol{n}) = p(T(e), \boldsymbol{n}) = 0.000772/3 = 0.000257$, $p(T(c), \boldsymbol{n}) = 0.003601/3 = 0.001203$, and $p(T(d), \boldsymbol{n}) = 0.001029/3 = 0.000343$, so that $p(\boldsymbol{Q}, \boldsymbol{n}) = 0.004973$, in agreement with the recursive solution.

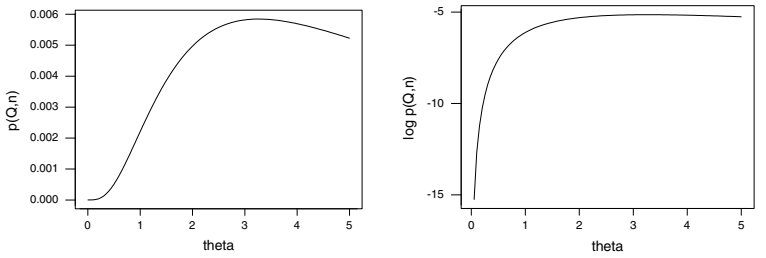
5.11 Maximum likelihood estimation

For the example in the previous section, it can be shown that the likelihood is

$$p(\boldsymbol{Q}, \boldsymbol{n}) = \frac{4\theta^4(5\theta^2 + 14\theta + 10)}{27(\theta + 1)^3(\theta + 2)^5}.$$

This has the value 0.004973 when $\theta = 2$, as we found above. The maximum likelihood estimator of θ is $\hat{\theta} = 3.265$, and the approximate variance (found from the second derivative of the log-likelihood) is 8.24. The likelihood curves are plotted in Figure 5.12.

Fig. 5.12. Likelihood $p(\boldsymbol{Q}, \boldsymbol{n})$ plotted as a function of θ, together with log-likelihood.

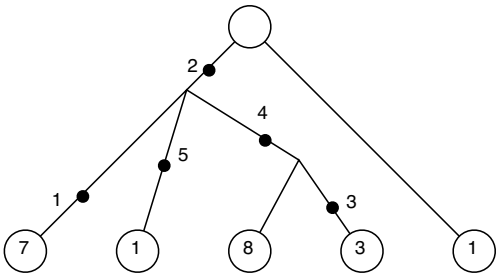


As might be expected, there is little information in such a small sample. Now consider a data set with 20 sequences, 5 segregating sites and multiplicities given below. The reduced genealogical tree is given in Figure 5.13.

$$\begin{array}{l} 0\,1\,0\,1\,0 \;:\; 8 \\ 0\,1\,1\,1\,0 \;:\; 3 \\ 0\,0\,0\,0\,0 \;:\; 1 \\ 0\,1\,0\,0\,1 \;:\; 1 \\ 1\,1\,0\,0\,0 \;:\; 7 \end{array}$$

Assuming that the ancestral labels are known, the probabilities $p^*(T, \boldsymbol{n})$ may be found using the recursion in (5.7.1), and they give a value of the MLE as $\hat{\theta} = 1.40$.

Fig. 5.13. Rooted genealogical tree for example data set. [Here, leaf labels refer to multiplicities of sequences]



To develop a practical method of maximum likelihood we need to be able to solve the recursions for p^0 for large sample sizes and large numbers of segregating sites. A general method for doing this is discussed in the next section.

6 Estimation in the Infinitely-many-sites Model

In this section we describe some likelihood methods for the infinitely-many-sites model, with a view to estimation of the compound mutation parameter θ. The method described here originated with Griffiths and Tavaré (1994), and has since been revisited by Felsenstein *et al.* (1999) and Stephens and Donnelly (2000). As we saw at the end of the previous section, exact calculation using the recursion approach is possible for relatively small sample sizes. For larger samples a different approach is required. We begin this section with Monte Carlo-based method for approximating these sampling probabilities by simulation backwards along the sample paths of the coalescent. Later in the section we relate this approach to importance sampling and show how to improve the original approach.

6.1 Computing likelihoods

Griffiths and Tavaré's approach is based on an elementary result about Markov chains given below.

Lemma 6.1 *Let $\{X_k; k \geq 0\}$ be a Markov chain with state space S and transition matrix P. Let A be a set of states for which the hitting time*

$$\eta = \inf\{k \geq 0 : X_k \in A\}$$

is finite with probability one starting from any state $x \in T \equiv S \setminus A$. Let $f \geq 0$ be a function on S, and define

$$u_x(f) = \mathbb{E}_x \prod_{k=0}^{\eta} f(X_k) \tag{6.1.1}$$

for all $X_0 = x \in S$, so that

$$u_x(f) = f(x), x \in A$$

Then for all $x \in T$

$$u_x(f) = f(x) \sum_{y \in S} p_{xy} u_y(f). \tag{6.1.2}$$

Proof.

$$\begin{aligned}
u_x(f) &= \mathbb{E}_x\Big(\prod_{k=0}^{\eta} f(X_k)\Big) \\
&= f(x)\mathbb{E}_x\Big(\prod_{k=1}^{\eta} f(X_k)\Big) \\
&= f(x)\mathbb{E}_x\Big(\mathbb{E}_x\Big(\prod_{k=1}^{\eta} f(X_k)\Big)|X_1\Big) \\
&= f(x)\mathbb{E}_x\Big(\mathbb{E}_{X_1}\Big(\prod_{k=0}^{\eta} f(X_k)\Big)\Big) \text{ (by the Markov property)} \\
&= f(x)\mathbb{E}_x u(X_1) \\
&= f(x)\sum_{y\in S} p_{xy}u_y(f).
\end{aligned}$$

□

This result immediately suggests a simulation method for solving equations like that on the right of (6.1.2): simulate a trajectory of the chain X starting at x until it hits A at time η, compute the value of the product $\prod_{k=0}^{\eta} f(X_k)$, and repeat this several times. Averaging these values provides an estimate of $u_x(f)$.

One application of this method is calculation of the sample tree probabilities $p^0(T, \boldsymbol{n})$ for the infinitely-many-sites model using the recursion in (5.7.3). In this case the appropriate Markov chain $\{X_k, k \geq 0\}$ has a tree state space, and makes transitions as follows:

$$(T, \boldsymbol{n}) \to (T, \boldsymbol{n} - \boldsymbol{e}_k) \text{ with probability } \frac{(n_k - 1)}{f(T, \boldsymbol{n})(n + \theta - 1)} \tag{6.1.3}$$

$$\to (\mathcal{S}_k T, \boldsymbol{n}) \text{ with probability } \frac{\theta}{f(T, \boldsymbol{n})n(n + \theta - 1)} \tag{6.1.4}$$

$$\to (\mathcal{R}_k T, \mathcal{R}_k(\boldsymbol{n} + \boldsymbol{e}_j)) \text{ with prob. } \frac{\theta(n_j + 1)}{f(T, \boldsymbol{n})n(n + \theta - 1)} \tag{6.1.5}$$

The first type of transition is only possible if $n_k > 1$, and the second or third if $n_k = 1$. In the last two transitions a distinct singleton first coordinate in a sequence is removed. The resulting sequence is still distinct from the others in (6.1.4), but in (6.1.5) the shifted kth sequence is equal to the jth sequence. The scaling factor is

$$f(T, \boldsymbol{n}) \equiv f_\theta(T, \boldsymbol{n}) = \sum_{k=1}^{d} \frac{(n_k - 1)}{(n + \theta - 1)} + \frac{\theta m}{n(n + \theta - 1)},$$

where m is given by

$$m = |\{k : n_k = 1,\ x_{k,0} \text{ distinct},\ \S \boldsymbol{x}_k \neq \boldsymbol{x}_j\ \forall\ j\}| + \sum_{k:n_k=1,\ x_{k,0}\ \text{distinct}} \sum_{j:\S \boldsymbol{x}_k = \boldsymbol{x}_j} (n_j + 1).$$

The idea is to run the process starting from an initial tree $(T, \boldsymbol{n})$ until the time τ at which there are two sequences $(x_{10}, \ldots, x_{1i})$ and $(x_{20}, \ldots, x_{2j})$ with $x_{1i} = x_{2j}$ (corresponding to the root of the tree) representing a tree T_2. The probability of such a tree is

$$p^0(T_2) = (2 - \delta_{i+j,0}) \binom{i+j}{j} \left[\frac{\theta}{2(1+\theta)}\right]^{i+j} \frac{1}{1+\theta}.$$

The representation of $p^0(T, \boldsymbol{n})$ is now

$$p^0(T, \boldsymbol{n}) = \mathbb{E}_{(T,\boldsymbol{n})} \left[\prod_{l=0}^{\tau-1} f(T(l), \boldsymbol{n}(l))\right] p^0(T_2), \tag{6.1.6}$$

where $X(l) \equiv (T(l), \boldsymbol{n}(l))$ is the tree at time l. Equation (6.1.6) may be used to produce an estimate of $p^0(T, \boldsymbol{n})$ by simulating independent copies of the tree process $\{X(l), l = 0, 1, \ldots\}$, and computing $\left[\prod_{l=0}^{\tau-1} f(T(l), \boldsymbol{n}(l))\right] p^0(T_2)$ for each run. The average over all runs is then an unbiased estimator of $p^0(T, \boldsymbol{n})$. An estimate of $p^*(T, \boldsymbol{n})$ can then be found by dividing by $a(T, \boldsymbol{n})$.

6.2 Simulating likelihood surfaces

The distribution $p^0(T, \boldsymbol{n})$ provides the likelihood of the data $(T, \boldsymbol{n})$, and so can be exploited for maximum likelihood approaches. One way to do this is to simulate the likelihood *independently* at a grid of points, and examine the shape of the resulting curve. In practice, this can be a very time consuming approach. In this section we describe another approach, based on importance sampling, for approximating a likelihood surface at a grid of points using just one run of the simulation algorithm.

The method uses the following lemma, a generalization of Lemma 6.1. The proof is essentially the same, and is omitted.

Lemma 6.2 *Let $\{X_k; k \geq 0\}$ be a Markov chain with state space S and transition matrix P. Let A be a set of states for which the hitting time*

$$\eta \equiv \eta_A = \inf\{k \geq 0 : X_k \in A\}$$

is finite with probability one starting from any state $x \in T \equiv S \setminus A$. Let $h \geq 0$ be a given function on A, let $f \geq 0$ be a function on $S \times S$ and define

$$u_x(f) = \mathbb{E}_x h(X_\eta) \prod_{k=0}^{\eta-1} f(X_k, X_{k+1}) \tag{6.2.1}$$

for all $X_0 = x \in S$, *so that*

$$u_x(f) = h(x), x \in A.$$

Then for all $x \in T$

$$u_x(f) = \sum_{y \in S} f(x,y) p_{xy} u_y(f). \tag{6.2.2}$$

It is convenient to recast the required equations in a more generic form, corresponding to the notation in Lemma 6.2. We denote by $q_\theta(x)$ the probability of the data x when the unknown parameters have value θ, which might be vector-valued. Equations such as (5.7.3) can then be recast in the form

$$q_\theta(x) = \sum_y f_\theta(x) p_\theta(x,y) q_\theta(y) \tag{6.2.3}$$

for some appropriate transition matrix $p_\theta(x,y)$. Now suppose that θ_0 is a particular set of parameters satisfying

$$f_\theta(x) p_\theta(x,y) > 0 \Rightarrow p_{\theta_0}(x,y) > 0.$$

We can recast the equations (6.2.3) in the form

$$q_\theta(x) = \sum_y f_\theta(x) \frac{p_\theta(x,y)}{p_{\theta_0}(x,y)} p_{\theta_0}(x.y)\, q_\theta(y) \tag{6.2.4}$$

so that from Lemma 6.2

$$q_\theta(x) = \mathbb{E}_x q_\theta(X(\eta)) \prod_{j=0}^{\eta-1} f_{\theta,\theta_0}(X(j), X(j+1)) \tag{6.2.5}$$

where $\{X(k), k \geq 0\}$ is the Markov chain with parameters θ_0 and

$$f_{\theta,\theta_0}(x,y) = f_\theta(x) \frac{p_\theta(x,y)}{p_{\theta_0}(x,y)}. \tag{6.2.6}$$

It follows that $q_\theta(x)$ can be calculated from the realizations of a single Markov chain, by choosing a value of θ_0 to drive the simulations, and evaluating the functional $q(X(\eta)) \prod_{j=0}^{\eta-1} f_{\theta,\theta_0}(X(j), X(j+1))$ along the sample path for each of the different values of θ of interest.

6.3 Combining likelihoods

It is useful to use independent runs for several values of θ_0 to estimate $q_\theta(x)$ on a grid of θ-values. For each such θ, the estimates for different θ_0 have the required mean $q_\theta(x)$, but they have different variances for different θ_0. This

raises the question about how estimated likelihoods from different runs might be combined. Suppose then that we are approximating the likelihood on a set of g grid points, $\theta_1, \ldots, \theta_g$, using r values of θ_0 and t runs of each simulation. Let $\hat{q}_{ij}$ be the sample average of the t runs at the jth grid point for the ith value of θ_0. For large t, the vectors $\hat{\boldsymbol{q}}_i \equiv (\hat{q}_{i1}, \ldots, \hat{q}_{ig}), i = 1, \ldots, r$ have independent and approximately multivariate Normal distributions with common mean vector $(q_{\theta_1}(x), \ldots, q_{\theta_g}(x))$ and variance matrices $t^{-1}\Sigma_1, \ldots, t^{-1}\Sigma_r$ respectively. The matrices $\Sigma_1, \ldots, \Sigma_r$ are unknown but may be estimated in the conventional way from the simulations. Define the log-likelihood estimates $\hat{\boldsymbol{l}}_i \equiv (\hat{l}_{ij}, j = 1, 2, \ldots, g)$ by

$$\hat{l}_{ij} = \log \hat{q}_{ij}, \ j = 1, \ldots, g, \ i = 1, \ldots, r.$$

By the delta method, the vectors $\hat{\boldsymbol{l}}_i, i = 1, \ldots, r$ are independent, asymptotically Normal random vectors with common mean vector $\boldsymbol{l} \equiv (l_1, \ldots, l_g)$ given by

$$l_i = \log q_{\theta_i}(x),$$

and covariance matrices $t^{-1}\Sigma_i^*$ determined by

$$(\Sigma_i^*)_{lm} = \frac{(\Sigma_i)_{lm}}{q_{\theta_l}(x)\, q_{\theta_m}(x)}. \tag{6.3.1}$$

If the Σ_j^* were assumed known, the minimum variance unbiased estimator of $\boldsymbol{l}$ would be

$$\hat{\boldsymbol{l}} = \left(\sum_{j=1}^{r} \left(\Sigma_j^*\right)^{-1} \right)^{-1} \sum_{j=1}^{r} \left(\Sigma_j^*\right)^{-1} \hat{\boldsymbol{q}}_j'. \tag{6.3.2}$$

If the observations for different Θ_j are not too correlated, it is useful to consider the simpler estimator with $\Sigma_j' \equiv \operatorname{diag} \Sigma_j^*$ replacing Σ_j^* in (6.3.2). This estimator requires a lot less computing than that in (6.3.2). In practice, we use the estimated values $\hat{q}_{il}$ and $\hat{q}_{im}$ from the ith run to estimate the terms in the denominator of (6.3.1).

6.4 Unrooted tree probabilities

The importance sampling approach can be used to find the likelihood of an unrooted genealogy. However it seems best to proceed by finding all the possible rooted labelled trees corresponding to an unrooted genealogy, and their individual likelihoods. Simulate the chain $\{(T(l), \boldsymbol{n}(l)), l = 0, 1, \ldots\}$ with a particular value θ_0 as parameter, and obtain the likelihood surface for other values of θ using the representation

$$p_\theta^0(T, \boldsymbol{n}) = \mathbb{E}_{(T,\boldsymbol{n})}^{\theta_0} \left[\prod_{l=0}^{\tau-1} h((T(l), \boldsymbol{n}(l)), (T(l+1), \boldsymbol{n}(l+1))) \right] p_\theta^0(T_2), \tag{6.4.1}$$

where $(T(l), \boldsymbol{n}(l))$ is the tree at time l, and h is determined by

$$h((T,\boldsymbol{n}),(T,\boldsymbol{n}-\boldsymbol{e}_k)) = f_{\theta_0}(T,\boldsymbol{n})\frac{n+\theta_0-1}{n+\theta-1},$$

and

$$h((T,\boldsymbol{n}),(T',\boldsymbol{n}')) = f_{\theta_0}(T',\boldsymbol{n}')\frac{\theta(n+\theta_0-1)}{\theta_0(n+\theta-1)}.$$

where the last form holds for both transitions (6.1.4), when $(T',\boldsymbol{n}') = (\mathcal{S}_k T, \boldsymbol{n})$, and (6.1.5), when $(T',\boldsymbol{n}') = (\mathcal{R}_k T, \mathcal{R}_k(\boldsymbol{n}+\boldsymbol{e}_j))$.

Example

To illustrate the method we consider the following set of 30 sequences, with multiplicities given in parentheses:

0 0 1 0 0 0 1 (3)
0 0 0 0 0 0 1 (4)
0 0 0 0 0 0 0 (4)
1 0 0 1 0 0 0 (11)
1 0 0 0 0 0 0 (1)
0 1 0 0 0 0 0 (2)
0 0 0 0 1 0 1 (2)
0 0 0 0 1 1 1 (3)

Simulations of the process on a grid of θ-values $\theta = 0.6(0.2)3.0$ for $\theta_0 = 1.0, 1.8$, and 2.6 were run for 30,000 replicates each. The curves of $\log p^0$ were combined as described earlier. This composite curve is compared with the true curve, obtained by direct numerical solution of the recursion, in Figure 6.1.

6.5 Methods for variable population size models

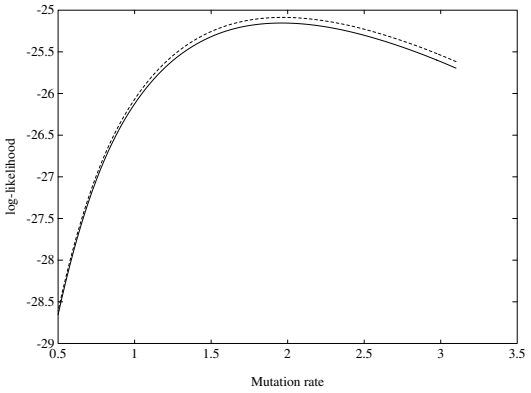
The present approach can also be used when the population size varies, as shown by Griffiths and Tavaré (1996, 1997). The appropriate recursions have a common form that may be written

$$q(t,x) = \int_t^\infty \sum_y r(s;x,y)q(s,y)g(t,x;s)ds \tag{6.5.1}$$

where $r(s;x,y) \geq 0$ and $g(t,x;s)$ is the density of the time to the first event in the ancestry of the sample after time t:

$$g(t,x;s) = \gamma(s,x)\exp\left(-\int_t^s \gamma(u,x)du\right). \tag{6.5.2}$$

Fig. 6.1. Log-likelihood curves. Dashed line: exact values. Solid line: Monte Carlo approximant.



Define

$$f(s;x) = \sum_y r(s;x,y)$$
$$P(s;x,y) = \frac{r(s;x,y)}{f(s;x)}, \tag{6.5.3}$$

and rewrite (6.5.1) as

$$q(t,x) = \int_t^\infty f(s;x) \sum_y P(s;x,y) q(s,y)\ g(t,x;s) ds. \tag{6.5.4}$$

We associate a non-homogeneous Markov chain $\{X(t), t \geq 0\}$ with (6.5.4) as follows: Given that $X(t) = x$, the time spent in state x has density $g(t,x;s)$, and given that a change of state occurs at time s, the probability that the next state is y is $P(s;x,y)$. The process $X(\cdot)$ has a set of absorbing states, corresponding to those x for which $q(\cdot,x)$ is known. $X(\cdot)$ may be used to give a probabilistic representation of $q(t,x)$ analogous to the result in Lemma 6.1 in the following way: Let $\tau_1 < \tau_2 \cdots < \tau_k = \tau$ be the jump times of $X(\cdot)$, satisfying $\tau_0 \equiv t < \tau_1$, where τ is the time to hit the absorbing states. Then

$$q(t,x) = \mathbb{E}_{(t,x)} q(\tau, X(\tau)) \prod_{j=1}^{k} f(\tau_j; X(\tau_{j-1})), \tag{6.5.5}$$

where $\mathbb{E}_{(t,x)}$ denotes expectation with respect to $X(t) = x$.

Once more, the representation in (6.5.5) provides a means to approximate $q(x) \equiv q(0,x)$: Simulate many independent copies of the process $\{X(t), t \geq 0\}$

starting from $X(0) = x$, and compute the observed value of the functional under the expectation sign in (6.5.5) for each of them. The average of these functionals is an unbiased estimate of $q(x)$, and we may then use standard theory to see how accurately $q(x)$ has been estimated.

We have seen that it is important, particularly in the context of variance reduction, to have some flexibility in choosing the stopping time τ. Even in the varying environment setting, there are cases in which $q(\cdot, x)$ can be computed (for example by numerical integration) for a larger collection of states x, and then it is useful to choose τ to be the hitting time of this larger set.

The probability $q(t, x)$ is usually a function of some unknown parameters, which we denote once more by θ; we write $q_\theta(t, x)$ to emphasize this dependence on θ. Importance sampling may be used as earlier to construct a single process $X(\cdot)$ with parameters θ_0, from which estimates of $q_\theta(t, x)$ may be found for other values of θ. We have

$$q_\theta(t,x) = \int_t^\infty \sum_y f_{\theta,\theta_0}(t,x;s,y) P_{\theta_0}(s;x,y) q_\theta(s,y)\ g_{\theta_0}(t,x;s) ds \qquad (6.5.6)$$

where

$$f_{\theta,\theta_0}(t,x;s,y) = \frac{f_\theta(s;x) g_\theta(t,x;s) P_\theta(s;x,y)}{g_{\theta_0}(t,x;s) P_{\theta_0}(s;x,y)}$$

and $f_\theta(s;x)$ and $P_\theta(s;x,y)$ are defined in (6.5.3). The representation analogous to (6.5.5) is

$$q_\theta(t,x) = \mathbb{E}_{(t,x)} q(\tau, X(\tau)) \prod_{j=1}^{k} f_{\theta,\theta_0}(\tau_{j-1}, X(\tau_{j-1}); \tau_j, X(\tau_j)), \qquad (6.5.7)$$

and estimates of $q_\theta(t, x)$ may be simulated as described earlier in the Section.

6.6 More on simulating mutation models

The genetic variability we observe in samples of individuals is the consequence of mutation in the ancestry of these individuals. In this section, we continue the description of how mutation processes may be superimposed on the coalescent. We suppose that genetic types are labelled by elements of a set E, the 'type space'. As mutations occur, the labels of individuals move around according to a mutation process on E.

We model mutation by supposing that a particular offspring of an individual of type $x \in E$ has a type in the set $B \subseteq E$ with probability $\Gamma(x, B)$. The mutation probabilities satisfy

$$\int_E \Gamma(x, dy) = 1, \quad \text{for all } x \in E.$$

When E is discrete, it is more usual to specify a transition matrix $\Gamma = (\gamma_{ij})$, where γ_{ij} is the probability that an offspring of an individual of type i is of type j. Such a mutation matrix Γ satisfies

$$\gamma_{ij} \geq 0, \quad \sum_{j \in E} \gamma_{ij} = 1 \text{ for each } i.$$

We assume that conditional its parent's type, the type of a particular offspring is independent of the types of other offspring, and of the demography of the population. In particular, the offspring of different individuals mutate independently.

In Section 3.4 we described a way to simulate samples from an infinitely-many-alleles model. This method can be generalized easily to any mutation mechanism. Generate the coalescent tree of the sample, sprinkle Poisson numbers of mutations on the branches at rate $\theta/2$ per branch, and superimpose the effects of the mutation process at each mutation. For discrete state spaces, this amounts to changing from type $i \in E$ to $jinE$ with probability γ_{ij} at each mutation. This method works for variable population size, by running from the bottom up to generate the ancestral history, then from top down to add mutations.

When the population size is constant, it is possible to perform the simulation from the top down in one sweep.

Algorithm 6.1 To generate a stationary random sample of size n.

1. Choose a type at random according to the stationary distribution π of Γ. Copy this type, resulting in 2 lines.
2. If there are currently k lines, wait a random amount of time having exponential distribution with parameter $k(k+\theta-1)/2$ and choose one of the lines at random. Split this line into 2 (each with same type as parent line) with probability $(k-1)/(k+\theta-1)$, and otherwise mutate the line according to Γ.
3. If there are fewer than $n+1$ lines, return to step 2. Otherwise go back to the last time at which there were n lines and stop.

This algorithm is due to Ethier and Griffiths (1987); See also Donnelly and Kurtz (1996). Its nature comes from the 'competing exponentials' world, and it only works in the case of constant population size. For the infinitely-many-alleles and infinitely-many-sites models, the first step has to be modified so that the MRCA starts from an arbitrary label.

6.7 Importance sampling

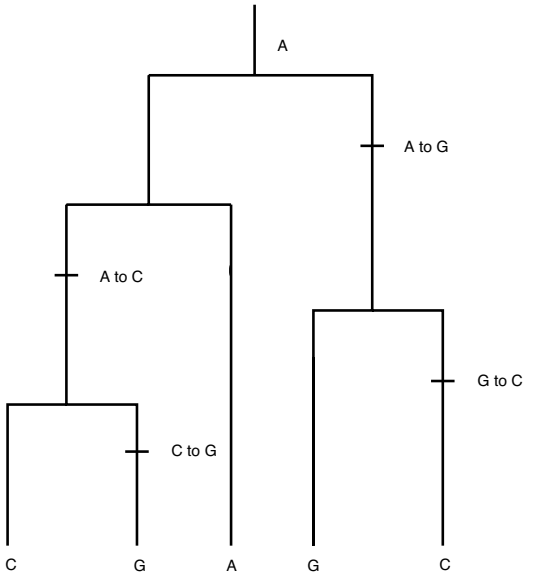
The next two sections are based on the papers of Felsenstein *et al.* (1999), and Stephens and Donnelly (2000). The review article of Stephens (2001) is also useful. In what follows, we assume a constant size population.

The *typed ancestry* $\mathcal{A}$ of the sample is its genealogical tree G, together with the genetic type of the most recent common ancestor (MRCA) and the details and positions of the mutation events that occur along the branches of

G. An example is given in Figure 6.2. Algorithm 6.1 can be used to simulate observations having the distribution of $\mathcal{A}$.

The *history* $\mathcal{H}$ is the typed ancestry $\mathcal{A}$ with time and topology information removed. So $\mathcal{H}$ is the type of the MRCA together with an ordered list of the split and mutation events which occur in $\mathcal{A}$ (including the details of the types involved in in each event, but not including which line is involved in each event). The history $\mathcal{H}$ contains a record of the states $(H_{-m}, H_{-m+1}, \ldots, H_{-1}, H_0)$ visited by the process beginning with the type $H_{-m} \in E$ of the MRCA and ending with genetic types $H_0 \in E^n$ of the sample. Here m is random, and the H_i are unordered lists of genetic types. Think of $\mathcal{H}$ as $(H_{-m}, H_{-m+1}, \ldots, H_{-1}, H_0)$, although it actually contains the details of which transitions occur between these states. In Figure 6.2, we have $\mathcal{H} = (\{A\}, \{A, A\}, \{A, G\}, \{A, A, G\}, \{A, C, G\}, \{A, C, G, G\}, \{A, C, C, G\}, \{A, C, C, C, G\}, \{A, C, C, G, G\})$.

Fig. 6.2. Genealogical tree G, typed ancestry $\mathcal{A}$ and history $\mathcal{H}$



If H_i is obtained from H_{i-1} by a mutation from α to β, write $H_i = H_{i-1} - \alpha + \beta$, whereas if H_i is obtained from H_{i-1} by the split of a line of type α, write $H_i = H_{i-1} + \alpha$. The distribution $P_\theta(\mathcal{H})$ of $\mathcal{H}$ is determined by the distribution π of the type of the MRCA, by the stopping rule in Algorithm 6.1, and by the Markov transition probabilities

$$\tilde{p}_\theta(H_i \mid H_{i-1}) = \begin{cases} \dfrac{n_\alpha}{n}\dfrac{\theta}{n-1+\theta}\Gamma_{\alpha\beta} \text{ if } H_i = H_{i-1} - \alpha + \beta \\ \dfrac{n_\alpha}{n}\dfrac{n-1}{n-1+\theta} \qquad \text{if } H_i = H_{i-1} + \alpha \\ 0 \qquad\qquad\qquad\quad \text{otherwise} \end{cases} \tag{6.7.1}$$

where n_α is the number of chromosomes of type α in H_{i-1} and $n = \sum n_\alpha$.

We want to compute the distribution $q_\theta(\cdot)$ of the genetic types $\mathcal{D}_n = (a_1, \ldots, a_n)$ in a random ordered sample. A sample from $\mathcal{H}$ provides, through H_0, a sample from q_θ. To get the ordered sample, we have to label the elements of H_0, so that

$$q_\theta(\mathcal{D}_n \mid \mathcal{H}) = \begin{cases} (\prod_{\alpha \in E} n_\alpha!)/n! \text{ if } H_0 \text{ is consistent with } \mathcal{D}_n \\ 0 \qquad\qquad\qquad\quad \text{otherwise.} \end{cases} \tag{6.7.2}$$

We regard $L(\theta) \equiv q_\theta(\mathcal{D}_n)$ as the likelihood of the data $\mathcal{D}_n$. The Griffiths-Tavaré method uses the representation

$$L(\theta) = \mathbb{E}\left(\prod_{j=0}^{\tau} F(B_j) \mid B_0 = \mathcal{D}_n\right), \tag{6.7.3}$$

where $B_0, B_1, \ldots$ is a particular Markov chain and τ a stopping time for the chain; recall (6.1.6). Using (6.7.2), we can calculate

$$L(\theta) = \int q_\theta(\mathcal{D}_n \mid \mathcal{H}) P_\theta(\mathcal{H}) d\mathcal{H} \tag{6.7.4}$$

This immediately suggests a naive estimator of $L(\theta)$:

$$L(\theta) \approx \frac{1}{R}\sum_{i=1}^{R} q_\theta(\mathcal{D}_n \mid \mathcal{H}_i) \tag{6.7.5}$$

where $\mathcal{H}_i, i = 1, \ldots, R$ are independent samples from $P_\theta(\mathcal{H})$. Unfortunately each term in the sum is with high probability equal to 0, so reliable estimation of $L(\theta)$ will require *enormous* values of R.

The importance sampling approach tries to circumvent this difficulty. Suppose that $Q_\theta(\cdot)$ is a distribution on histories that satisfies $\{\mathcal{H} : Q_\theta(\mathcal{H}) > 0\} \supset \{\mathcal{H} : P_\theta(\mathcal{H}) > 0\}$. Then we can write

$$L(\theta) = \int q_\theta(\mathcal{D}_n \mid \mathcal{H}) \frac{P_\theta(\mathcal{H})}{Q_\theta(\mathcal{H})} Q_\theta(\mathcal{H}) d\mathcal{H} \tag{6.7.6}$$

$$\approx \frac{1}{R}\sum_{i=1}^{R} q_\theta(\mathcal{D}_n \mid \mathcal{H}) \frac{P_\theta(\mathcal{H}_i)}{Q_\theta(\mathcal{H}_i)} := \frac{1}{R}\sum_{i=1}^{R} w_i, \tag{6.7.7}$$

where $\mathcal{H}_1, \ldots, \mathcal{H}_R$ are independent samples from $Q_\theta(\cdot)$.

We call the distribution Q_θ the IS proposal distribution, and the w_i are called the IS weights. The idea of course is to choose the proposal distribution in such a way that the variance of the estimator in (6.7.7) is much smaller than that of the estimator in (6.7.5). The optimal choice Q_θ^* of Q_θ is

$$Q_\theta^*(\mathcal{H}) = P_\theta(\mathcal{H} \mid \mathcal{D}_n); \tag{6.7.8}$$

in this case

$$q_\theta(\mathcal{D}_n \mid \mathcal{H}) \frac{P_\theta(\mathcal{H})}{Q_\theta^*(\mathcal{H})} = L(\theta),$$

so the variance of the estimator is 0. Unfortunately, the required conditional distribution of histories is not known, so something else has to be tried.

In Section 6.2 we mentioned that estimating $L(\theta)$ on a grid of points can be done independently at each grid point, or perhaps by importance sampling, which in the present setting reduces to choosing the driving value θ_0, and calculating

$$L(\theta) \approx \frac{1}{R} \sum_{i=1}^{R} q_\theta(\mathcal{D}_n \mid \mathcal{H}) \frac{P_\theta(\mathcal{H}_i)}{Q_{\theta_0}(\mathcal{H}_i)} \tag{6.7.9}$$

where $\mathcal{H}_1, \ldots, \mathcal{H}_R$ are independent samples from $Q_{\theta_0}(\cdot)$.

6.8 Choosing the weights

A natural class of proposal distributions on histories arises by considering randomly reconstructing histories backward in time in a Markovian way, from the sample $\mathcal{D}_n$ back to an MRCA. So a random history $\mathcal{H} = (H_{-m}, \ldots, H_{-1}, H_0)$ may be sampled by choosing $H_0 = \mathcal{D}_n$, and successively generating $H_{-1}, \ldots, H_{-m}$ according to prespecified backward transition probabilities $p_\theta(H_{i-1} \mid H_i)$. The process stops at the first time that the configuration H_{-m} consists of a single chromosome.

In order for (6.7.6) to hold, we need to look at the subclass $\mathcal{M}$ of these distributions for which, for each i, the support of $p_\theta(\cdot \mid H_i)$ is the set

$$\{H_{i-1} : \tilde{p}_\theta(H_i \mid H_{i-1}) > 0\}$$

where $\tilde{p}_\theta$ is given in (6.7.1). Such a p_θ then specifies a distribution Q_θ whose support is the set of histories consistent with the data $\mathcal{D}_n$.

Felsenstein *et al.* (1999) showed that the Griffiths-Tavaré scheme in (6.7.3) is a special case of this strategy, with

$$p_\theta(H_{i-1} \mid H_i) \propto \tilde{p}_\theta(H_{i-1} \mid H_i). \tag{6.8.1}$$

The optimal choice of Q_θ^* turns out to be from the class $\mathcal{M}$. Stephens and Donnelly (2000) prove the following result:

Theorem 6.3 *Define $\pi(\cdot \mid \mathcal{D})$ to be the conditional distribution of the type of an $(n+1)$th sampled chromosome, given the types $\mathcal{D}$ of the first n sampled chromosomes. Thus*

$$\pi(\alpha \mid \mathcal{D}) = \frac{q_\theta(\{\mathcal{D}, \alpha\})}{q_\theta(\mathcal{D})}.$$

The optimal proposal distribution Q^*_θ is in the class $\mathcal{M}$, with

$$p^*_\theta(H_{i-1} \mid H_i) = \begin{cases} C^{-1}\dfrac{\theta}{2} n_\alpha \dfrac{\pi(\beta \mid H_i - \alpha)}{\pi(\alpha \mid H_i - \alpha)} \Gamma_{\beta\alpha} & \text{if } H_{i-1} = H_i - \alpha + \beta, \\ C^{-1}\dbinom{n_\alpha}{2} \dfrac{1}{\pi(\alpha \mid H_i - \alpha)} & \text{if } H_{i-1} = H_i - \alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{6.8.2}$$

where n_α is the number of chromosomes of type α in H_i, and $C = n(n-1+\theta)/2$ where n is the number of chromosomes in H_i.

It is clear that knowing p^*_θ is equivalent to knowing Q^*_θ, which in turn is equivalent to knowing $L(\theta)$. So it should come as no surprise that the conditional probabilities are unknown for most cases of interest. The only case that is known explicitly is that in which $\Gamma_{\alpha\beta} = \Gamma_\beta$ for all α, β. In this case

$$\pi(\beta \mid \mathcal{D}) = \frac{n_\beta + \theta\Gamma_\beta}{n + \theta}. \tag{6.8.3}$$

Donnelly and Stephens argue that under the optimal proposal distribution there will be a tendency for mutations to occur towards the rest of the sample, and that coalescences of unlikely types are more likely than those of likely types. This motivated their choice of approximation $\hat{\pi}(\cdot \mid \mathcal{D})$ to the sampling probabilities $\pi(\cdot \mid \mathcal{D})$. They define $\hat{\pi}(\cdot \mid \mathcal{D})$ by choosing an individual from $\mathcal{D}$ at random, and mutating it a geometric number of times according to the mutation matrix Γ. So

$$\hat{\pi}(\beta \mid \mathcal{D}) = \sum_{\alpha \in E} \frac{n_\alpha}{n} \sum_{m=0}^{\infty} \left(\frac{\theta}{\theta + n}\right)^m \frac{n}{\theta + n} \Gamma^m_{\alpha\beta} \tag{6.8.4}$$

$$\equiv \sum_{\alpha \in E} \frac{n_\alpha}{n} M^{(n)}_{\alpha\beta}. \tag{6.8.5}$$

$\hat{\pi}$ has a number of interesting properties, among them the fact that when $\Gamma_{\alpha\beta} = \Gamma_\beta$ for all α, β we have $\hat{\pi}(\cdot \mid \mathcal{D}) = \pi(\cdot \mid \mathcal{D})$ and the fact that $\hat{\pi}(\cdot \mid \mathcal{D}) = \pi(\cdot \mid \mathcal{D})$ when $n = 1$ and Γ is reversible.

The proposal distribution $\hat{Q}^*_\theta$, an approximation to Q^*_θ, is defined by substituting $\hat{\pi}(\cdot \mid \mathcal{D})$ into (6.8.2):

$$\hat{p}_\theta(H_{i-1} \mid H_i) = \begin{cases} C^{-1}\dfrac{\theta}{2} n_\alpha \dfrac{\hat{\pi}(\beta \mid H_i - \alpha)}{\hat{\pi}(\alpha \mid H_i - \alpha)} \Gamma_{\beta\alpha} & \text{if } H_{i-1} = H_i - \alpha + \beta, \\ C^{-1}\binom{n_\alpha}{2} \dfrac{1}{\hat{\pi}(\alpha \mid H_i - \alpha)} & \text{if } H_{i-1} = H_i - \alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{6.8.6}$$

In order to sample from $\hat{p}_\theta$ efficiently, one can use the following algorithm.

Algorithm 6.2

1. Choose a chromosome uniformly at random from those in H_i, and denote its type by α.
2. For each type $\beta \in E$ for which $\Gamma_{\beta\alpha} > 0$, calculate $\hat{\pi}(\beta \mid H_i - \alpha)$ from equation (6.8.5).
3. Sample H_i by setting

$$H_{i-1} = \begin{cases} H_i - \alpha + \beta \text{ w.p. } \propto \theta \hat{\pi}(\beta \mid H_i - \alpha) \Gamma_{\beta\alpha} \\ H_i - \alpha \text{ w.p. } \propto n_\alpha - 1. \end{cases}$$

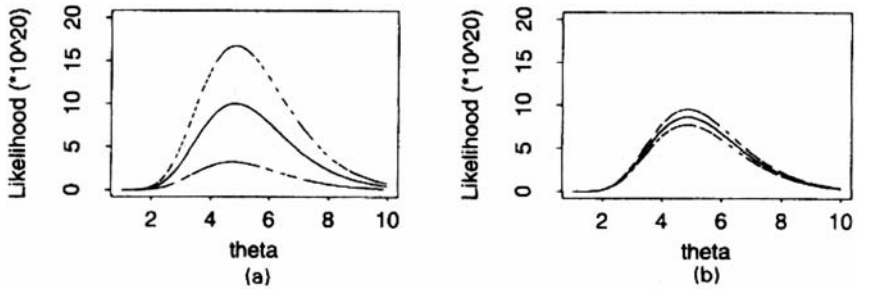
Example

Stephens and Donnelly give a number of examples of the use of their proposal distribution, including for the infinitely-many-sites model. In this case, the foregoing discussion has to be modified, because the type space E is uncountably infinite. However the principles behind the derivation of the proposal distribution $\hat{Q}_\theta$ can be used here too. Namely, we choose a chromosome uniformly at random from those present, and assume this chromosome is involved in the most recent event back in time. As we have seen (recall Theorem 5.1), the configuration of types H_i is equivalent to an unrooted genealogical tree, and the nature of mutations on that tree means that the chromosomes that can be involved in the most recent event backwards in time from H_i are limited:

(a) any chromosome which is not the only one of its type may coalesce with another of that type;
(b) any chromosome which is the only one of its type and has only one neighbor on the unrooted tree corresponding to H_i may have arisen from a mutation to that neighbor.

So their proposal distribution chooses the most recent event back in time by drawing a chromosome uniformly at random from those satisfying (a) or (b). Notice that this distribution does not depend on θ. In Figure 6.3 are shown a comparison of the Griffiths-Tavaré method with this new proposal distribution.

Fig. 6.3. (a) Likelihood surface estimate with ±2 standard deviations from 100,000 runs of GT method, with $\theta_0 = 4$. (b) the same using 100,000 runs of the SD IS function. This is Fig. 7 from Stephens and Donnelly (2000).



It is an open problem to develop other, perhaps better, IS distributions for rooted and unrooted trees as well. The method presented here is also not appropriate for variable population size models, where the simple Markov structure of the process is lost. The representation of the Griffiths-Tavaré method as importance sampling, together with the results for the constant population size model, suggest that the development of much more efficient likelihood algorithms in that case. See Chapter 2 of Liu (2001) for an introduction to sequential importance sampling in this setting. The paper of Stephens and Donnelly has extensive remarks from a number of discussants on the general theme of computational estimation of likelihood surfaces.

7 Ancestral Inference in the Infinitely-many-sites Model

The methods in this section are motivated by the problem of inferring properties of the time to the most recent common ancestor of a sample given the data from that sample. For example, Dorit *et al.* (1996) sequenced a 729 bp region of the ZFY gene in a sample of $n = 38$ males and observed no variability; the number of segregating sites in the data is then $S_{38} = 0$. What can be said about the time to the MRCA (TMRCA) given the observation that $S_{38} = 0$?

Note that the time to the MRCA is an unobservable random variable in the coalescent setting, and so the natural quantity to report is the conditional distribution of W_n given the data $\mathcal{D}$, which in this case is just just the event $\{S_n = 0\}$. In this section we derive some of properties of such conditional distributions. In later sections we consider much richer problems concerning inference about the structure of the coalescent tree conditional on a sample. The main reference for the material in this section is Tavaré *et al.* (1997).

7.1 Samples of size two

Under the infinitely-many-sites assumption, all of the information in the two sequences is captured in S_2, the number of segregating sites. Our goal, then, is to describe T_2, the time to the most recent common ancestor of the sample in the light of the data, which is the observed value of S_2.

One approach is to treat the realized value of T_2 as an unknown parameter which is then naturally estimated by $\tilde{T}_2 = S_2/\theta$, since $\mathbb{E}(S_2|T_2) = \theta T_2$. Such an approach, however, does not use all of the available information. In particular, the information available about T_2 due to the effects of genealogy and demography are ignored.

Under the coalescent model, when $n = 2$ the coalescence time T_2 has an exponential distribution with mean 1 before the data are observed. As Tajima (1983) noted, it follows from Bayes Theorem that after observing $S_2 = k$, the distribution of T_2 is gamma with parameters $1 + k$ and $1 + \theta$, which has probability density function

$$f_{T_2}(t|S_2{=}k) = \frac{(1+\theta)^{1+k}}{k!} t^k e^{-(1+\theta)t}, \quad t \geq 0. \tag{7.1.1}$$

In particular,

$$\mathbb{E}(T_2|S_2{=}k) = \frac{1+k}{1+\theta}, \tag{7.1.2}$$

$$\mathrm{var}(T_2|S_2{=}k) = \frac{1+k}{(1+\theta)^2}. \tag{7.1.3}$$

The pdf (7.1.1) conveys all of the information available about T_2 in the light of both the data and the coalescent model.

If a point estimate were required, equation (7.1.2) suggests the choice $\hat{T}_2 = (1+S_2)/(1+\theta)$. Perhaps not surprisingly, the estimator $\hat{T}_2$, which is based on all of the available information, is superior to $\tilde{T}_2$ which ignores the pre-data information. For example, writing MSE for the mean square error of an estimator, straightforward calculations show that

$$\mathrm{MSE}(\hat{T}_2) = \frac{1}{1+\theta} < \frac{1}{\theta} = \mathrm{MSE}(\tilde{T}_2).$$

The difference in mean square errors could be substantial for small θ. In addition, the estimator $\tilde{T}_2$ is clearly inappropriate when $S_2 = 0$.

7.2 No variability observed in the sample

We continue to assume the infinitely-many-sites mutation model with parameter θ, and derive the distribution of $W_n := T_n + \cdots + T_2$ given $S_n = 0$ for the case of constant population size. Several authors have been motivated to study this particular problem, among them Fu and Li (1996), Donnelly *et al.* (1996) and Weiss and von Haeseler (1996). Because mutations occur according to independent Poisson processes on the branches of the coalescent tree, we see that

$$\begin{aligned}
\mathbb{E}(\exp(-uW_n)\mathbb{1}(S_n = 0)) &= \mathbb{E}[\mathbb{E}(\exp(-uW_n)\mathbb{1}(S_n = 0) \mid T_n, \ldots, T_2)] \\
&= \mathbb{E}[\exp(-uW_n)\mathbb{E}(\mathbb{1}(S_n = 0) \mid T_n, \ldots, T_2)] \\
&= \mathbb{E}[\exp(-uW_n)\exp(-\theta L_n/2)] \\
&= \prod_{j=2}^{n} \mathbb{E}\exp(-(u+\theta j/2)T_j) \\
&= \prod_{j=2}^{n} \frac{\binom{j}{2}}{\binom{j}{2} + u + \frac{\theta j}{2}}
\end{aligned}$$

Since

$$\mathbb{P}(S_n = 0) = \prod_{j=1}^{n-1} \frac{j}{j+\theta},$$

we see that

$$\mathbb{E}(\exp(-uW_n)|S_n = 0) = \prod_{j=2}^{n} \frac{j(j+\theta-1)/2}{u + j(j+\theta-1)/2}. \tag{7.2.1}$$

Let $\tilde{W}_n$ denote a random variable with the same distribution as the conditional distribution of W_n given $S_n = 0$. Equation (7.2.1) shows that we can write

$$\tilde{W}_n = \tilde{T}_n + \cdots + \tilde{T}_2 \tag{7.2.2}$$

where the $\tilde{T}_i$ are independent exponential random variables with parameters $\binom{i}{2} + \frac{i\theta}{2}$ respectively. Many properties of $\tilde{W}_n$ follow from this. In particular

$$\mathbb{E}(W_n | S_n = 0) = \sum_{j=2}^{n} \frac{2}{j(j+\theta-1)}. \tag{7.2.3}$$

The conditional density function of W_n may be calculated from a partial fraction expansion, resulting in the expression

$$f_{W_n}(t|S_n = 0) = \sum_{j=2}^{n} (-1)^j \frac{(2j+\theta-1)n_{[j]}(\theta+1)_{(j)}}{2(j-2)!(\theta+n)_{(j)}} e^{-j(\theta+j-1)t/2}. \tag{7.2.4}$$

The corresponding distribution function follows from

$$\mathbb{P}(W_n > t | S_n = 0) = \sum_{j=2}^{n} (-1)^{j-2} \frac{(2j+\theta-1)n_{[j]}(\theta+1)_{(j)}}{(j-2)!j(j+\theta-1)(\theta+n)_{(j)}} e^{-j(\theta+j-1)t/2}.$$

Intuition suggests that given the sample has no variability, the post-data TMRCA of the sample should be stochastically smaller than the pre-data TMRCA. This can be verified by the following simple coupling argument. Let $E_2, \ldots, E_n$ be independent exponential random variables with parameters $\theta, \ldots, n\theta/2$ respectively, and let $T_2, \ldots, T_n$ be independent exponential random variables with parameters $\binom{2}{2}, \ldots, \binom{n}{2}$ respectively, independent of the E_i. Noting that $\tilde{T}_i = \min(T_i, E_i)$, we see that

$$\begin{aligned} \tilde{W}_n &= \tilde{T}_n + \cdots + \tilde{T}_2 \\ &= \min(T_n, E_n) + \cdots + \min(T_2, E_2) \\ &\leq T_n + \cdots + T_2 \\ &= W_n, \end{aligned}$$

establishing the claim.

7.3 The rejection method

The main purpose of this section is to develop the machinery that allows us to find the joint distribution of the coalescent tree $\mathcal{T}$ conditional on the sample of size n having configuration $\mathcal{D}$. Here $\mathcal{D}$ is determined by the mutation process acting on the genealogical tree $\mathcal{T}$ of the sample. Such conditional distributions lead directly to the conditional distribution of the height W_n of the tree.

The basic result we exploit to study such quantities is contained in

Lemma 7.1 *For any real-valued function g for which $\mathbb{E}|g(\mathcal{T})| < \infty$, we have*

$$\mathbb{E}(g(\mathcal{T})|\mathcal{D}) = \frac{\mathbb{E}(g(\mathcal{T})\mathbb{P}(\mathcal{D}|\mathcal{T}))}{\mathbb{P}(\mathcal{D})}. \tag{7.3.1}$$

Proof. We have

$$\begin{aligned}\mathbb{E}(g(\mathcal{T})\mathbb{1}(\mathcal{D})) &= \mathbb{E}(\mathbb{E}(g(\mathcal{T})\mathbb{1}(\mathcal{D}|\mathcal{T})) \\ &= \mathbb{E}(g(\mathcal{T})\mathbb{E}(\mathbb{1}(\mathcal{D})|\mathcal{T})) \\ &= \mathbb{E}(g(\mathcal{T})\mathbb{P}(\mathcal{D}|\mathcal{T})).\end{aligned}$$

Dividing this by $\mathbb{P}(\mathcal{D})$ completes the proof. □

For most mutation mechanisms, explicit results are not available for these expectations, but we can develop a simple simulation algorithm. The expectation in (7.3.1) has the form

$$\mathbb{E}(g(\mathcal{T})|\mathcal{D}) = \int g(t)\frac{\mathbb{P}(\mathcal{D}|t)}{\mathbb{P}(\mathcal{D})}f_n(t)dt, \tag{7.3.2}$$

where $f_n(t)$ denotes the density of $\mathcal{T}$. The expression in (7.3.2) is a classical set-up for the rejection method:

Algorithm 7.1 To simulate from the distribution of $\mathcal{T}$ given $\mathcal{D}$.

1. Simulate an observation t from the coalescent distribution of $\mathcal{T}$.
2. Calculate $u = \mathbb{P}(\mathcal{D}|t)$.
3. Keep t with probability u, else go to Step 1.

The joint distribution of the *accepted* trees t is precisely the conditional distribution of $\mathcal{T}$ given $\mathcal{D}$.

The average number of times the rejection step is repeated per output observation is $1/\mathbb{P}(\mathcal{D})$, so that for small values of $\mathbb{P}(\mathcal{D})$ the method is likely to be inefficient. It can be improved in several ways. If, for example, there is a constant c such that

$$\mathbb{P}(\mathcal{D}|t) \leq c \quad \text{for all values of } t,$$

then u in Step 2 of the algorithm can be replaced by u/c.

Note that if properties of W_n are of most interest, observations having the conditional distribution of W_n given $\mathcal{D}$ can be found from the trees generated in algorithm 7.1. When the data are summarized by the number S_n of segregating sites, these methods become somewhat more explicit, as is shown in the next section.

7.4 Conditioning on the number of segregating sites

In this section we consider events of the form

$$\mathcal{D} \equiv \mathcal{D}_k = \{S_n = k\},$$

corresponding to the sample of size n having k segregating sites. Since each mutation in the coalescent tree corresponds to a segregating site, it follows that

$$\mathbb{P}(\mathcal{D}|\mathcal{T}) = \mathbb{P}(\mathcal{D}_k|L_n) = \mathrm{Po}(\theta L_n/2)\{k\},$$

where $L_n = 2T_2 + \cdots + nT_n$ is the total length of the ancestral tree of the sample and $\mathrm{Po}(\lambda)\{k\}$ denotes the Poisson point probability

$$\mathrm{Po}(\lambda)\{k\} = e^{-\lambda}\frac{\lambda^k}{k!},\ k = 0, 1, \ldots.$$

Therefore

$$\mathbb{E}(g(W_n)|\mathcal{D}_k) = \frac{\mathbb{E}(g(W_n)\mathrm{Po}(\theta L_n/2)\{k\})}{\mathbb{E}(\mathrm{Po}(\theta L_n/2)\{k\})} \tag{7.4.1}$$

The simulation algorithm 7.1 then becomes

Algorithm 7.2 To simulate from the joint density of $T_2, \ldots, T_n$ given $\mathcal{D}_k$.

1. Simulate an observation $\boldsymbol{t} = (t_n, \ldots, t_2)$ from the joint distribution of $\boldsymbol{T}_n = (T_n, \ldots, T_2)$. Calculate $l = 2t_2 + \cdots + nt_n$.
2. Calculate $u = \mathbb{P}(\mathcal{D}_k|\boldsymbol{t}) = \mathrm{Po}(\theta l/2)\{k\}$.
3. Keep $\boldsymbol{t}$ with probability u, else go to Step 1.

The joint distribution of the *accepted* vectors $\boldsymbol{t}$ is precisely the conditional distribution of $\boldsymbol{T}_n$ given $\mathcal{D}_k$.

Since

$$\mathbb{P}(S_n = k|\boldsymbol{t}) = \mathrm{Po}(\theta l_n/2)\{k\} \le \mathrm{Po}(k)\{k\},$$

where we define Po(0,0) = 1, the modified algorithm becomes:

Algorithm 7.3 To simulate from the joint density of $T_2, \ldots, T_n$ given $S_n = k$.

1. Simulate an observation $\boldsymbol{t} = (t_n, \ldots, t_2)$ from the joint distribution of $\boldsymbol{T}_n = (T_n, \ldots, T_2)$.
2. Calculate $l = 2t_2 + \cdots + nt_n$, and set

$$u = \frac{\mathrm{Po}(l\theta/2)\{k\}}{\mathrm{Po}(k)\{k\}}$$

3. Keep $\boldsymbol{t}$ with probability u, else go to Step 1.

Values of $w_n = t_2 + \cdots + t_n$ calculated from accepted vectors $\boldsymbol{t}$ have the conditional distribution of W_n given $S_n = k$.

Notice that nowhere have we assumed a particular form for the distribution of $\boldsymbol{T}_n$. In particular, the method works when the population size is variable so long as $\boldsymbol{T}_n$ has the distribution specified by (2.4.8). For an analytical approach to the constant population size case, see Fu (1996).

Remark. In these examples, we have simulated the ancestral process back to the common ancestor. It is clear, however, that the same approach can be used to simulate observations for any fixed time t into the past. All that is required is to simulate coalescence times back into the past until time t, and then the effects of mutation (together with the genetic types of the ancestors at time t) can be superimposed on the coalescent forest.

Example

We use this technique to generate observations from the model with variable population size when the conditioning event is $\mathcal{D}_0$. The particular population size function we use for illustration is

$$f(x) = \alpha^{\min(t/v,1)}, \tag{7.4.2}$$

corresponding to a population of constant relative size α more than (coalescent) time v ago, and exponential growth from time v until the present relative size of 1.

In the illustration, we chose $V = 50,000$ years, $N = 10^8$, a generation time of 20 years and $\alpha = 10^{-4}$. Thus $v = 2.5 \times 10^{-5}$. We compare the conditional distribution of W_n given $\mathcal{D}_0$ to that in the constant population size case with $N = 10^4$. Histograms of 5000 simulated observations are given in Figures 7.1 and 7.2. The mean of the conditional distribution in the constant population size case is 313,200 years, compared to 358,200 years in the variable case. Examination of other summary statistics of the simulated data (Table 7) shows that the distribution in the variable case is approximately that in the constant size case, plus about V years. This observation is supported by the plot of the empirical distribution functions of the two sets in Figure 7.3.

The intuition behind this is clear. Because of the small sample size relative to the initial population size N, the sample of size n will typically have about n distinct ancestors at the time of the expansion, V. These ancestors themselves form a random sample from a population of size αN.

Table 7. Summary statistics from 5000 simulation runs

	constant	variable
mean	313,170	358,200
std dev	156,490	158,360
median	279,590	323,210
5%	129,980	176,510
95%	611,550	660,260

Fig. 7.1. Histogram of 5000 replicates for constant population size, $N = 10^4$

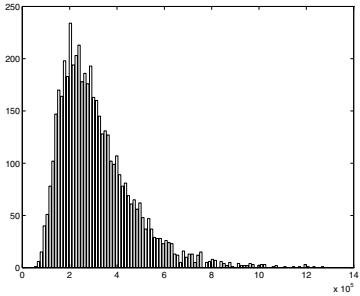


Fig. 7.2. Histogram of 5000 replicates for variable population size, $N = 10^8, T = 50,000, \alpha = 10^{-4}$

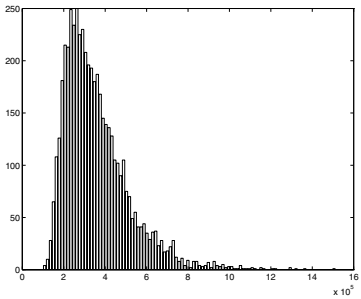
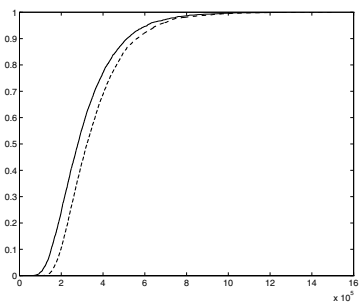


Fig. 7.3. Empirical distribution function. Solid line is constant population size case.



7.5 An importance sampling method

If moments of the post-data distribution of W_n, say, are required, then they can be found in the usual way from observations generated by Algorithm 7.2. As an alternative, an importance sampling scheme can be used. This is best illustrated by an example. Consider then the expression in (7.4.1). We have

$$\mathbb{E}(g(W_n)|\mathcal{D}_k) = \frac{\mathbb{E}(g(W_n)\mathrm{Po}(\theta L_n/2)\{k\})}{\mathbb{E}(\mathrm{Po}(\theta L_n/2)\{k\})}.$$

Point estimates of this quantity can be found by simulating independent copies $(W_n^{(j)}, L_n^{(j)}), j = 1, 2, \ldots, R$ of the height and length of the ancestral tree and computing the ratio estimator

$$r_R = \frac{\sum_{j=1}^{R} g(W_n^{(j)})\mathrm{Po}(\theta L_n^{(j)}/2)\{k\}}{\sum_{j=1}^{R} \mathrm{Po}(\theta L_n^{(j)}/2)\{k\}}. \tag{7.5.1}$$

One application provides an estimate of the conditional distribution function of W_n given $\mathcal{D}_k$: Suppose that we have ordered the points $W_n^{(j)}$ and listed them as $W_n^{[1]} < W_n^{[2]} < \cdots < W_n^{[R]}$. Let $L_n^{[1]}, \ldots, L_n^{[R]}$ be the corresponding L-values. The empirical distribution function then has jumps of height

$$\frac{e^{-\theta L_n^{[l]}/2}}{\sum_{j=1}^{R} e^{-\theta L_n^{[j]}/2}}$$

at the points $W_n^{[l]},\ l = 1, 2, \ldots, R$.

This approach uses all the simulated observations, but requires either knowing which g are of interest, or storing a lot of observations. Asymptotic properties of the ratio estimator can be found from standard theory.

7.6 Modeling uncertainty in N and μ

In this section, we use prior information about the distribution of μ, as well as information that captures our uncertainty about the population size N. We begin by describing some methods for generating observations from the posterior distribution of the vector (W_n, N, μ) given the data $\mathcal{D}$. We use this to study the posterior distribution of the time W_n to a common ancestor, measured in years:

$$W_n^y = N \times G \times W_n.$$

The rejection method is based on the analog of (7.3.1):

$$\mathbb{E}(g(\boldsymbol{T}_n, N, \mu)|\mathcal{D}) = \frac{\mathbb{E}(g(\boldsymbol{T}_n, N, \mu)\mathbb{P}(\mathcal{D}|\boldsymbol{T}_n, N, \mu))}{\mathbb{P}(\mathcal{D})}. \tag{7.6.1}$$

This converts once more into a simulation algorithm; for definiteness we suppose once more that $\mathcal{D} = \{S_n = k\}$.

Algorithm 7.4 To simulate from conditional distribution of $\boldsymbol{T}_n, N, \mu$ given $S_n = k$.

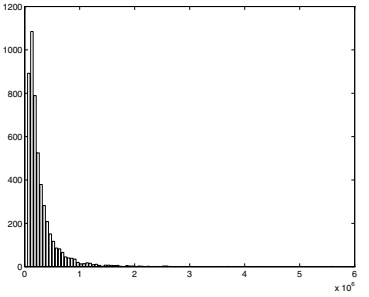
1. Generate an observation $\boldsymbol{t}, N, \mu$ from the joint distribution of $\boldsymbol{T}_n, N, \mu$.
2. calculate $l = 2t_2 + \cdots + nt_n$, and
$$u = \frac{\text{Po}(lN\mu)\{k\}}{\text{Po}(k)\{k\}}$$
3. accept $\boldsymbol{t}, N, \mu$ with probability u, else go to Step 1.

Usually we assume that N and μ are independent of $\boldsymbol{T}_n$, and that N and μ are themselves independent.

Examples

Suppose that no variation is observed in the data, so that $\mathcal{D}_0$. Suppose that N has a lognormal distribution with parameters (10, 1), and that μ has a Gamma distribution with mean μ_0 and standard deviation $C\mu_0$. A constant size population is assumed. In the example, we took $\mu_0 = 2 \times 10^{-5}$ and $C = 1/20$ and $C = 1.0$. Histograms appear in Figures 7.4 and 7.5, and some summary statistics are given in Table 8.

Fig. 7.4. Histogram of 5000 replicates $C = 1/20$



Here we illustrate for the exponential growth model described earlier, with initial population size $N = 10^8$, and $\alpha = 10^{-4}$. We took N lognormally distributed with parameters 17.92, 1. (The choice of 17.92 makes the mean of $N = 10^8$.) For μ we took the Gamma prior with mean $= \mu_0$, and standard deviation $C\mu_0$. In the simulations, we used $C = 1$ and $C = 1/20$. Histograms of 5000 simulated observations are given in Figures 7.6 and 7.7. Some summary statistics are given in Table 9.

The importance sampling method also readily adapts to this Bayesian setting: apply the approach outlined in (7.5.1) to the expectation formula in (7.6.1).

Fig. 7.5. Histogram of 5000 replicates $C = 1$

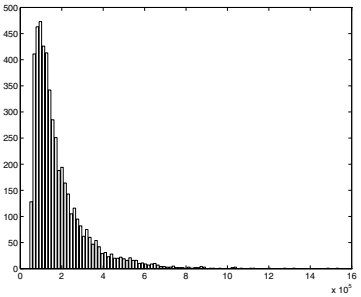


Table 8. Summary statistics from 5000 simulation runs. Prior mean $\mu_0 = 2 \times 10^{-5}, \mathcal{D} = \mathcal{D}_0$

	$C = 1.0$	$C = 1/20$
mean	647,821	262,590
median	369,850	204,020
5%	68,100	52,372
95%	2,100,000	676,890

Fig. 7.6. Histogram of 5000 replicates. Variable size model. $C = 1/20$

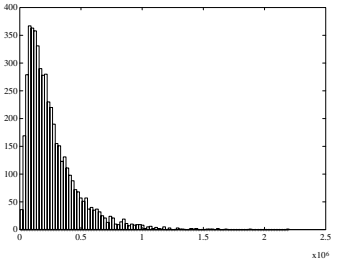


Fig. 7.7. Histogram of 5000 replicates. Variable size model. $C = 1$

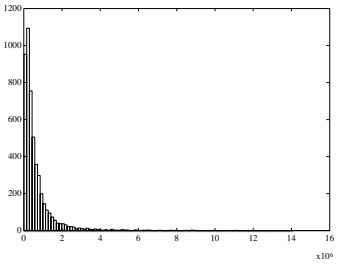


Table 9. Summary statistics from 5000 simulation runs. Prior mean $\mu_0 = 2 \times 10^{-5}$, $\mathcal{D} = \mathcal{D}_0$

	$C = 1$	$C = 1/20$
mean	292,000	186,000
median	194,000	141,490
5%	70,600	65,200
95%	829,400	462,000

7.7 Varying mutation rates

These rejection methods can be employed directly to study the behavior of the infinitely-many-sites model that allows for several regions with different mutation rates. Suppose then that there are r regions, with mutation rates $\mu_1, \ldots, \mu_r$. The analysis also applies, for example, to r different types of mutations within a given region. We sample n individuals, and observe k_1 segregating sites in the first region, k_2 in the second, $\ldots$, and k_r in the r^{th}. The problem is to find the conditional distribution of $\mathcal{T}$, given the vector $(k_1, \ldots, k_r)$.

When N and the μ_i are assumed known, this can be handled by a modification of Algorithm 7.2. Conditional on L_n, the probability of $(k_1, \ldots, k_r)$ is

$$h(L_n) = \mathrm{Po}(k_1, L_n\theta_1/2) \times \cdots \times \mathrm{Po}(k_r, L_n\theta_r/2),$$

where $\theta_i = 2N\mu_i, i = 1, 2, \ldots, r$. It is easy to check that $h(L_n) \leq h(k/\theta)$, where

$$k = k_1 + \cdots + k_r, \quad \theta = \theta_1 + \cdots + \theta_r.$$

Therefore in the rejection algorithm we may take $u = h(L_n)/h(k/\theta)$ which simplifies to

$$u = h(L_n)/h(k/\theta) = \frac{\mathrm{Po}(L_n\theta/2)\{k\}}{\mathrm{Po}(k)\{k\}}. \tag{7.7.1}$$

Equation (7.7.1) establishes the perhaps surprising fact that the conditional distribution of W_n given $(k_1, \ldots, k_r)$ and $(\theta_1, \ldots, \theta_r)$ depends on these values only through their respective totals: the total number of segregating sites k and the total mutation rate θ. Thus Algorithm 7.2 can be employed directly with the appropriate values of k and θ. This result justifies the common practice of analyzing segregating sites data through the total number of segregating sites, even though these sites may occur in regions of differing mutation rate.

If allowance is to be made for uncertainty about the μ_i, then this simplification no longer holds. However, Algorithm 7.3 can be employed with the rejection step replaced by (7.7.2):

$$u = \frac{\mathrm{Po}(L_n\theta_1/2)\{k_1\}}{\mathrm{Po}(k_1)\{k_1\}} \cdots \frac{\mathrm{Po}(L_n\theta_r/2)\{k_r\}}{\mathrm{Po}(k_r)\{k_r\}}. \tag{7.7.2}$$

In this case, Step 2 requires generation of a vector of rates $\mu = (\mu_1, \ldots, \mu_r)$ from the joint prior π_μ. Furthermore, the algorithm immediately extends to the case of variable population size.

7.8 The time to the MRCA of a population given data from a sample

In this section, we show how the rejection technique can be used to study the time T_m to the MRCA of a sample of m individuals, conditional on the number of segregating sites in a subsample of size n. In many applications of ancestral inference, the real interest is on the time to the MRCA of the *population*, given data on a *sample*. This can be obtained by setting $m = N$ below. See Tavaré (1997) and Tavaré *et al.* (1997) for further details and examples.

The quantities of interest here are A_m (the number of distinct ancestors of the sample), A_n (the number of distinct ancestors of the subsample), and W_n (the time to the MRCA of the subsample). The results of Saunders *et al.* (1984) justify the following algorithm:

Algorithm 7.5 Rejection algorithm for $f_{W_m}(t|S_n{=}k)$.

1. Set $A_m = m, A_n = n, W_n = 0, L_n = 0$
2. Generate E, exponential of rate $A_m(A_m - 1)/2$. Set $W_n = W_n + W, L_n = L_n + A_n \cdot E$.
3. Set $p = \frac{A_n(A_n-1)}{A_m(A_m-1)}$. Set $A_m = A_m - 1$. With probability p set $A_n = A_n - 1$. If $A_n > 1$ go to 2.
4. Set $u = \text{Po}(\theta L_n/2)\{k\}/\text{Po}(k)\{k\}$. Accept (A_m, W_n) with probability u, else go to 1.
5. If $A_m = 1$, set $T_{nm} = 0$, and return $W_m = W_n$. Else, generate independent exponentials E_j with parameter $j(j-1)/2$, for $j = 2, 3, \ldots, A_m$, and set $T_{nm} = E_2 + \cdots + E_{A_m}$. Return $W_m = W_n + T_{nm}$.

Many aspects of the joint behavior of the sample and a subsample can be be studied using this method. In particular, values of (A_m, W_n) accepted at step 5 have the joint conditional distribution of the number of ancestors of the sample at the time the subsample reaches its common ancestor and the time of the MRCA of the subsample, conditional on the number of segregating sites in the subsample. In addition, values of T_{nm} produced at step 5 have the conditional distribution of the time between the two most recent common ancestors. It is straightforward to modify the method to cover the case of variable population size, and the case where uncertainty in N and μ is modeled. With high probability, the sample and the subsample share a common ancestor and therefore a common time to the MRCA. However, if the two common ancestors differ then the times to the MRCA can differ substantially. This is explored further in the examples below.

Examples

Whitfield *et al.* (1995) describe another Y chromosome data set that includes a sample of $n = 5$ humans. The 15,680 bp region has three polymorphic nucleotides that once again are consistent with the infinitely-many-sites model. They estimated the coalescence time of the sample to be between 37,000 and 49,000 years. Again, we present several reanalyses, each of which is based on the number of segregating sites in the data. The results are summarized in Table 10 and illustrated in Figure 7.8.

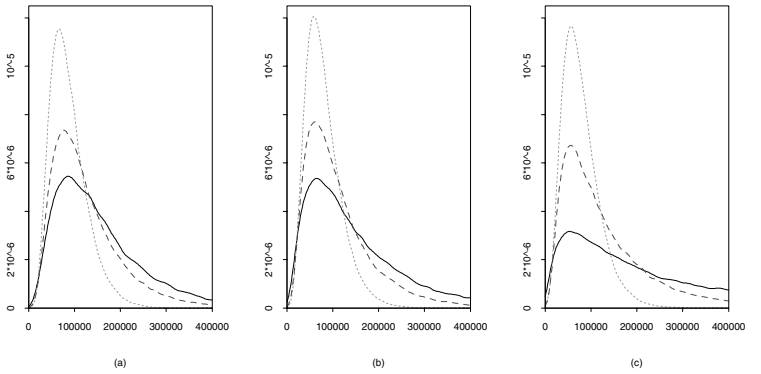
Table 10. Results of re-analyses of the data of Whitfield *et al.* In each case the data are $S_5 = 3$. Line (a) gives the interval reported by the authors (but note that they assigned no probability to their interval). Mean and 95% interval are estimated from samples of size 10,000. Details of the gamma and lognormal distributions are given in the text.

	Model	Mean of W_5 ($\times 10^3$)		95% Interval ($\times 10^3$)	
		pre-data	post-data	pre-data	post-data
(a)	Whitfield *et al.*				37 – 49
(b)	$N = 4,900$	157	87	31 – 429	30 – 184
	$\mu_S = 3 \cdot 52 \times 10^{-4}$				
(c)	$N = 4,900$	157	125	31– 429	32 – 321
	μ_S gamma				
(d)	N gamma	159	80	21 – 517	26 – 175
	$\mu_S = 3 \cdot 52 \times 10^{-4}$				
(e)	N gamma	159	117	21– 517	25 – 344
	μ_S gamma				
(f)	N lognormal	428	149	19 – 2,200	22 – 543
	μ_S gamma				

In estimating the coalescence time, Whitfield *et al.* adopt a method which does not use population genetics modeling. While the method is not systematically biased, it may be inefficient to ignore pre-data information about plausible values of the coalescence time. In addition, the method substantially underrepresents the uncertainty associated with the estimates presented. Here, we contrast the results of such a method with those of one which does incorporate background information.

To determine the mutation rate, we use the average figure of $1 \cdot 123 \times 10^{-9}$ substitutions per nucleotide position per year given in Whitfield *et al.*, and a

Fig. 7.8. Probability density curves for W_5. In each panel the three curves correspond to: solid, pre-data; dashed, post-data, assuming μ_S gamma; dotted, post-data assuming $\mu_S = 3 \cdot 52 \times 10^{-4}$. The three panels correspond to (a) $N = 4,900$; (b) N gamma; (c) N lognormal.



generation time of 20 years, to give $\mu = 15,680 \times 1 \cdot 123 \times 10^{-9} \times 20 = 3 \cdot 52 \times 10^{-4}$ substitutions per generation. For these parameter values, the post-data mean of W_5 is 87,000 years.

As noted in the previous section, the appropriate values of the parameters are not known. Analysis (c) incorporates uncertainty about μ, in the form of a gamma distribution with shape parameter 2 and mean $3 \cdot 52 \times 10^{-4}$, while continuing to assume that N is known to be 4,900. The effect is to greatly increase the post-data mean of W_5. Allowing N to be uncertain while μ_S is known has, on the other hand, the effect of slightly reducing the post-data estimates of W_5, compared with the case that N and μ_S are both known. This may be attributed to the data favoring values of N smaller than 4,900.

Analyses (e) and (f) incorporate uncertainty about both N and μ_S. They use the same prior distributions as analyses (g) and (i) respectively of the previous section. Note that, as should be expected, the uncertainty about T is larger than when one or both of N and μ_S are assumed known exactly.

Whitfield *et al.* (1995) point to their estimated coalescence time as being substantially shorter than those published for the human mitochondrial genome. In contrast, the ranges in each of our analyses (b) – (e) overlap with recent interval estimates for the time since mitochondrial Eve. In addition, recall that the quantity W_5 being estimated in Table 10 is the coalescence time of the sample of 5 males sequenced in the study. This time may be different from, and substantially shorter than, the coalescence time of *all* existing Y chromosomes. Under the assumption that $N = 4,900$ and $\mu = 3 \cdot 52 \times 10^{-4}$, Algorithm 7.5 can be used to show that the mean time to the common ancestor

of the male *population*, given $S_5 = 3$, is 157,300 years, with a corresponding 95% interval of (58,900 – 409,800) years. These figures differ markedly from the corresponding values for the sample, given at line (b) of Table 10. It is the population values which are likely to be of primary interest.

7.9 Using the full data

The approach that conditions on the number of segregating sites in the data is convenient primarily because the rejection methods are quick and easy to program. However, it does not make full use of the data. In this section, we discuss how we can approximate the conditional distribution of TMRCA given the infinitely-many-sites rooted tree $(T, \boldsymbol{n})$ that corresponds to the data, or the corresponding unrooted tree $(\boldsymbol{Q}, \boldsymbol{n})$. See Griffiths and Tavaré (1994, 1999) for further details.

Consider first the rooted case. The probability $q(t, x)$ that a sample taken at time t has configuration x satisfies an equation of the form

$$q(t,x) = \int_t^\infty \sum_y r(s;x,y)q(s,y)g(t,x;s)ds$$

for a positive kernel r. For the case of an unrooted tree, we have $x = (T, \boldsymbol{n})$. Now define

$$q(t,x,w) = \mathbb{P}(\text{sample taken at time } t \text{ has configuration } x \\ \text{and TMRCA } \leq t+w)$$

By considering the time of the first event in the history of the sample, it can be seen that $q(t, x, w)$ satisfies the equation

$$q(t,x,w) = \int_t^\infty \sum_y r(s;x,y)q(s,y,t+w-s)g(t,x;s)ds \tag{7.9.1}$$

where we assume that $q(t, x, y) = 0$ if $y < t$. Recursions of this type can be solved using the Markov chain simulation technique described in Section 6. The simplest method is given in (6.5.3): we define

$$f(s;x) = \sum_y r(s;x,y)$$
$$P(s;x,y) = \frac{r(s;x,y)}{f(s;x)},$$

and rewrite (7.9.1) in the form

$$q(t,x,w) = \int_t^\infty f(s;x) \sum_y P(s;x,y)q(s,y,t+w-s)g(t,x;s)ds. \tag{7.9.2}$$

The Markov chain associated with the density g and the jump matrix P is once again denoted by $X(\cdot)$. The representation we use is then

$$q(t,x,w) = \mathbb{E}_{(t,x)} q(\tau, X(\tau), t+w-\tau) \prod_{j=1}^{k} f(\tau_j; X(\tau_{j-1})), \tag{7.9.3}$$

where $t = \tau_0 < \tau_1 < \cdots < \tau_k = \tau$ are the jump times of $X(\cdot)$, and τ is the time taken to reach the set A that corresponds to a sample configuration x for a single individual. For the infinitely-many-sites tree, this corresponds to a tree of the form $(T, \boldsymbol{e}_1)$.

The natural initial condition is

$$q(t,x,w) = \mathbb{1}(w \geq 0), \ x \in A,$$

so that

$$q(\tau, X(\tau), t+w-\tau) = \mathbb{1}(\tau < t+w).$$

The Monte Carlo method generates R independent copies of the X process, and for the ith copy calculates the observed value

$$F_i = \prod_{j=1}^{k_i} f(\tau_j^i; X^i(\tau_{j-1}^i)).$$

and estimates $q(t,x,w)$ by

$$\hat{q}(t,x,w) = \frac{\sum_{i=1}^{R} F_i \mathbb{1}(\tau^i \leq t+w)}{\sum_{i=1}^{R} F_i}.$$

The distribution function of TMRCA given the data (t,x) can be therefore be approximated by a step function that jumps a height $F_{(l)}/\sum F_i$ at the point $\tau_{(l)}$, where the $\tau_{(l)}$ are the increasing rearrangement of the times τ^i, and the $F_{(l)}$ are the corresponding values of the F_i.

This method can be used immediately when the data correspond to a rooted tree $(T, \boldsymbol{n})$. When the data correspond to an unrooted tree $(\boldsymbol{Q}, \boldsymbol{n})$ we proceed slightly differently. Corresponding to the unrooted tree $(\boldsymbol{Q}, \boldsymbol{n})$ are rooted trees $(T, \boldsymbol{n})$. An estimator of $\mathbb{P}(TMRCA \leq t+w, (T, \boldsymbol{n}))$ is given by

$$\frac{1}{R} \sum_{i=1}^{R} F_i(T) \mathbb{1}(\tau_i(T) \leq t+w),$$

the T denoting a particular rooted tree. Recalling (5.9.3), an estimator of $q(t, (\boldsymbol{Q}, \boldsymbol{n}), w)$ is therefore given by

$$\sum_{T} \frac{1}{R} \sum_{i=1}^{R} F_i(T) \mathbb{1}(\tau_i(T) \leq t+w),$$

and the conditional probability $q(t,(\boldsymbol{Q},\boldsymbol{n}),w)/q(t,(\boldsymbol{Q},\boldsymbol{n}))$ is estimated by

$$\frac{\sum_T \sum_{i=1}^R F_i(T)\mathbb{1}(\tau_i(T) \le t+w)}{\sum_T \sum_{i=1}^R F_i(T)}.$$

The distribution of TMRCA given data $(\boldsymbol{Q},\boldsymbol{n})$ taken at time t is found by ranking all the times $\tau_j(T)$ over different T to get the increasing sequence $\tau_{(j)}$, together with the corresponding values $F_{(j)}$, and then approximating the distribution function by jumps of height $F_{(j)}/\sum F_{(j)}$ at the point $\tau_{(j)}$. Usually we take $t=0$ in the previous results.

8 The Age of a Unique Event Polymorphism

In this section we study the age of an allele observed in a sample of chromosomes. Suppose then that a particular mutation Δ has arisen just once in the history of the population of interest. This mutation has an age (the time into the past at which it arose), and we want to infer its distribution given data $\mathcal{D}$. These data can take many forms:

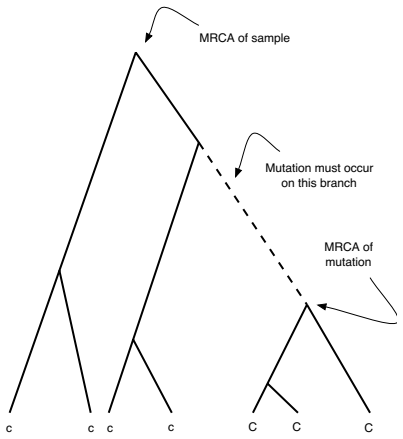
- the number of copies, b, of Δ observed in a sample of size n. Here we assume that $1 \leq b < n$, so that the mutation is segregating in the sample.
- the number of copies of Δ together with other molecular information about the region around Δ. For example, we might have an estimate of the number of mutations that have occurred in a linked region containing Δ.
- in addition, we might also have molecular information about the individuals in the sample who do not carry Δ.

The unique event polymorphism (UEP) assumption leads to an interesting class of coalescent trees that we study in the next section.

8.1 UEP trees

Suppose that the mutation Δ is represented b times in the sample. The UEP property means that the b sequences must coalesce together before any of the non-Δ sequences share any common ancestors with them. This situation is illustrated in Figure 8.1 for $n = 7$ and $b = 3$.

Fig. 8.1. Tree with UEP. The individuals carrying the special mutation Δ are labeled C, those not carrying the mutation are labeled c.



To understand the structure of these trees, we begin by studying the properties of trees that have the property $\mathcal{E}$ that a particular b sequences coalesce together before any of the other $n-b$ join their subtree. To this end, let $n > J_{b-1} > \cdots > J_1$ be the total number of distinct ancestors of the sample at the time the b first have $b-1, \ldots, 1$ distinct ancestors, and let J_0 $(1 \le J_0 < J_1)$ be the number of ancestors in the sample at the time the first of the other $n-b$ sequences shares a common ancestor with an ancestor of the b. In Figure 8.1, we have $J_2 = 5, J_1 = 4, J_0 = 2$.

It is elementary to find the distribution of $J_{b-1}, \ldots, J_0$. Recalling that in a coalescent tree joins are made at random, we find that

$$\begin{aligned}\mathbb{P}(J_r = j_r, r = b-1, \ldots, 0) = \prod_{r=2}^{b} \left\{ \frac{\binom{j_r - r}{2}}{\binom{j_r}{2}} \cdots \frac{\binom{j_{r-1}+2-r}{2}}{\binom{j_{r-1}}{2}} \frac{\binom{r}{2}}{\binom{j_{r-1}+1}{2}} \right\} \\ \times \frac{\binom{j_1 - 1}{2}}{\binom{j_1}{2}} \cdots \frac{\binom{j_0+2-1}{2}}{\binom{j_0+2}{2}} \frac{j_0}{\binom{j_0+1}{2}}\end{aligned}$$

where we have defined $j_b = n$, and where $1 \le j_0 < j_1 < \cdots < j_{b-1} < n$. This expression can be simplified to give

$$\mathbb{P}(J_r = j_r, r = b-1, \ldots, 0) = \frac{2b!(b-1)!(n-b)!(n-b-1)!j_0}{n!(n-1)!}. \tag{8.1.1}$$

We can find $\mathbb{P}(\mathcal{E})$ by summing $1 \le j_0 < j_1 < \cdots < j_{b-1} < n$. Note that

$$\begin{aligned}\sum_{j_0=1}^{n-b} \sum_{j_0 < j_1 < \cdots < j_{b-1} < n} 1 &= \sum_{j_0=1}^{n-b} j_0 \binom{n - j_0 - 1}{b-1} \\ &= \sum_{l=0}^{n-b-1} (l+1) \binom{n-1-l-1}{n-b-1-l} \\ &= \binom{n}{n-b-1},\end{aligned}$$

the last equality coming from the identity

$$\sum_{k=1}^{c} \binom{c}{k} \binom{d+k}{d+1} = \binom{d}{c-1},$$

valid for integral c, d with $c = b, d = 2$. It follows that

$$\begin{aligned}\mathbb{P}(\mathcal{E}) &= \frac{2b!(b-1)!(n-b)!(n-b-1)!}{n!(n-1)!} \binom{n}{n-b-1} \\ &= \frac{2}{b+1} \binom{n-1}{b-1}^{-1},\end{aligned} \tag{8.1.2}$$

as found by Wiuf and Donnelly (1999).

Now we can compute the conditional distribution of 'everything' given $\mathcal{E}$. For example it follows that for $1 \leq j_0 < j_1 < \cdots < j_{b-1} < n$

$$\mathbb{P}(J_r = j_r, r = b-1, \ldots, 0 \mid \mathcal{E}) = j_0 \binom{n}{b+1}^{-1}, \tag{8.1.3}$$

while for $1 \leq j_0 < j_1 < n$,

$$\mathbb{P}(J_1 = j_1, J_0 = j_0 \mid \mathcal{E}) = j_0 \binom{n-j_1-1}{b-2} \binom{n}{b+1}^{-1} \tag{8.1.4}$$

and for $1 < j_1 < j_2 \cdots < j_{b-1} < n$,

$$\mathbb{P}(J_r = j_r, r = b-1, \ldots, 2 \mid J_1 = j_1, J_0 = j_0, \mathcal{E}) = \binom{n-j_1-1}{b-2}^{-1}. \tag{8.1.5}$$

Having discussed the topological properties of UEP coalescent trees, we move on to the age of the mutation itself.

The distribution of J_{Δ}

Suppose that Δ mutations occur at rate $\mu/2$ on the branches of the coalescent tree. The random variable J_{Δ} gives the number of ancestors of the sample of size n when the mutation Δ occurs. Clearly, $J_0 < j_{\Delta} \leq J_1$. Its distribution can be found as follows. To get $J_{\Delta} = k$, a single mutation must arise on the branch of length T_k, and no other mutations must occur in the remainder of the coalescent tree. It follows from (8.1.4) that for $1 \leq j_0 < k \leq j_1 \leq n-b+1$,

$$\mathbb{P}(J_1 = j_1, J_{\Delta} = k, J_0 = j_0 \mid \boldsymbol{T}, \mathcal{E}) = \frac{\mu}{2} T_k e^{-L_n \mu/2} j_0 \binom{n-j_1-1}{b-2} \binom{n}{b+1}^{-1},$$

where $\boldsymbol{T} = (T_n, \ldots, T_2)$ and $L_n = nT_n + \cdots + 2T_2$ is the total length of the tree. Using the fact that for integral k,

$$\sum_{j=0}^{k} \binom{c+k-j-1}{k-j} \binom{d+j-1}{j} = \binom{c+d+k-1}{k}$$

we see that

$$\sum_{j_1=k}^{n-b+1} \binom{n-j_1-1}{b-2} = \binom{n-k}{b-1},$$

so that

$$\sum_{j_0=1}^{k-1} \sum_{j_1=k}^{n-b+1} j_0 \binom{n-j_1-1}{b-2} = \frac{k(k-1)}{2} \binom{n-k}{b-1}.$$

Hence

$$\mathbb{P}(J_\Delta = k \mid \boldsymbol{T}, \mathcal{E}) = \frac{\mu}{2} T_k e^{-L_n \mu/2} \frac{k(k-1)}{2} \binom{n-k}{b-1} \binom{n}{b+1}^{-1}, \tag{8.1.6}$$

and

$$\mathbb{P}(J_\Delta = k \mid \mathcal{E}) = \mathbb{E}\left(\frac{\mu}{2} T_k e^{-L_n \mu/2}\right) \frac{k(k-1)}{2} \binom{n-k}{b-1} \binom{n}{b+1}^{-1}.$$

Letting $\mathcal{U}$ denote the event that there is indeed a UEP, we have

$$\mathbb{P}(\mathcal{U} \mid \mathcal{E}) = \sum_{k=2}^{n-b+1} \mathbb{P}(J_\Delta = k \mid \mathcal{E}),$$

so that for $k = 2, \ldots, n-b+1$,

$$\mathbb{P}(J_\Delta = k \mid \mathcal{U} \cap \mathcal{E}) = \frac{k(k-1)\binom{n-k}{b-1} \mathbb{E}\left[T_k e^{-L_n \mu/2}\right]}{\sum_{l=2}^{n-b+1} l(l-1)\binom{n-l}{b-1} \mathbb{E}\left[T_l e^{-L_n \mu/2}\right]}. \tag{8.1.7}$$

Remark. In the constant population size case, this gives

$$\mathbb{P}(J_\Delta = k \mid \mathcal{U} \cap \mathcal{E}) = \frac{(k-1)\binom{n-k}{b-1} \frac{1}{k-1+\mu}}{\sum_{l=2}^{n-b+1} (l-1)\binom{n-l}{b-1} \frac{1}{l-1+\mu}},$$

as given by Stephens (2000).

Similar arguments show that for $k \le j_1 < j_2 \cdots < j_{b-1} < n$,

$$\mathbb{P}(J_1 = j_1, \ldots, J_{b-1} = j_{b-1} \mid J_\Delta = k, \mathcal{U} \cap \mathcal{E}) = \binom{n-k}{b-1}^{-1}, \tag{8.1.8}$$

so that given $J_\Delta = k$, the places where the subtree has joins form a random (ordered)$(b-1)-$ subset of the integers $k, k+1, \ldots, n-1$. Hence for $1 \le i \le b-1$ and $k \le j_1 < \cdots < j_i < n-i+b$,

$$\mathbb{P}(J_1 = j_1, \ldots, J_i = j_i \mid J_\Delta = k, \mathcal{U} \cap \mathcal{E}) = \binom{n-j_i-1}{b-i-1} \binom{n-k}{b-1}^{-1}. \tag{8.1.9}$$

8.2 The distribution of T_Δ

We let J_Δ be the number of ancestors of the sample at the time the unique Δ mutation occurs. Clearly $J_0 < J_\Delta \le J_1$. We can find the conditional distribution of the age T_Δ as follows. We have

$$
\begin{aligned}
&\mathbb{E}(e^{-\phi T_\Delta} \mid \mathcal{E}) \\
&= \sum_{k=2}^{n-b+1} \mathbb{E}(e^{-\phi T_\Delta} \mathbb{1}(J_\Delta = k) \mid \mathcal{E}) \\
&= \sum_{k=2}^{n-b+1} \mathbb{E}(\mathbb{E}(e^{-\phi T^{[k]}} \mathbb{1}(J_\Delta = k) \mid \boldsymbol{T}, \mathcal{E})) \\
&= \sum_{k=2}^{n-b+1} \mathbb{E}(e^{-\phi T^{[k]}} \mathbb{P}(J_\Delta = k \mid \boldsymbol{T}, \mathcal{E})) \\
&= \sum_{k=2}^{n-b+1} \mathbb{E}(e^{-\phi T^{[k]}} \frac{T_k \mu}{2} e^{-L_n \mu/2}) \frac{k(k-1)}{2} \binom{n-k}{b-1} \binom{n}{b+1}^{-1} \qquad (8.2.1)
\end{aligned}
$$

where

$$T^{[k]} = T_n + \cdots + T_{k+1} + UT_k, \qquad (8.2.2)$$

and U is uniformly distributed on $(0,1)$, independent of $\boldsymbol{T}$. The penultimate inequality comes from (8.1.6). This gives us:

Theorem 8.1 *The Laplace transform of the conditional distribution of the age T_Δ of a UEP observed b times in a sample of size n (where $0 < b < n$) is given by*

$$
\begin{aligned}
&\mathbb{E}(e^{-\phi T_\Delta} \mid \mathcal{U} \cap \mathcal{E}) \\
&= \frac{\sum_{k=2}^{n-b+1} k(k-1)\binom{n-k}{b-1} \mathbb{E}\left[e^{-\phi T^{[k]}} T_k e^{-L_n \mu/2}\right]}{\sum_{k=2}^{n-b+1} k(k-1)\binom{n-k}{b-1} \mathbb{E}\left[T_k e^{-L_n \mu/2}\right]} \\
&= \sum_{k=2}^{n-b+1} \mathbb{P}(J_\Delta = k \mid \mathcal{U} \cap \mathcal{E}) \, \frac{\mathbb{E}(e^{-\phi T^{[k]}} T_k e^{-L_n \mu/2})}{\mathbb{E}(T_k e^{-L_n \mu/2})}, \qquad (8.2.3)
\end{aligned}
$$

where $T^{[k]}$ is defined in (8.2.2).

Proof. This follows from the previous steps and (8.1.7).

Remark. The representation in (8.2.3) provides a useful way to simulate observations from T_Δ; this is exploited later. Note that the original random variables $\boldsymbol{T}$ can be tilted by the size-biasing function $e^{-L_n \mu/2}$, so that

$$\mathbb{E}_\mu f(T_n, \ldots, T_2) = \frac{\mathbb{E}(f(T_n, \ldots, T_2) e^{-L_n \mu/2})}{\mathbb{E}(e^{-L_n \mu/2})}.$$

In what follows we refer to this as *μ-biasing*. The previous results can then be written in terms of these μ-biased times:

$$\mathbb{P}_\mu(J_\Delta = k \mid \mathcal{U} \cap \mathcal{E}) = \frac{k(k-1)\binom{n-k}{b-1} \mathbb{E}_\mu T_k}{\sum_{l=2}^{n-b+1} l(l-1)\binom{n-l}{b-1} \mathbb{E}_\mu T_l}, \qquad (8.2.4)$$

and

$$\mathbb{E}_\mu(e^{-\phi T_\Delta} \mid \mathcal{U} \cap \mathcal{E}) = \sum_{k=2}^{n-b+1} \mathbb{P}_\mu(J_\Delta = k \mid \mathcal{U} \cap \mathcal{E}) \frac{\mathbb{E}_\mu(e^{-\phi T^{[k]}} T_k)}{\mathbb{E}_\mu T_k}. \tag{8.2.5}$$

8.3 The case $\mu = 0$

It is of great interest in practice to consider the limiting case in which the mutation rate at the special locus is extremely small. In this case rather more can be said about the age of a neutral mutation. An immediate specialization of Theorem 8.1 provides a proof of Griffiths and Tavaré's (1998) result:

Lemma 8.2 *The Laplace transform of the conditional distribution of the age T_Δ of a UEP observed b times in a sample of size n has limit as $\mu \to 0$ given by*

$$\mathbb{E}(e^{-\phi T_\Delta} \mid \mathcal{U} \cap \mathcal{E}) = \frac{\sum_{k=2}^{n-b+1} k(k-1)\binom{n-k}{b-1}\mathbb{E}(e^{-\phi T^{[k]}} T_k)}{\sum_{k=2}^{n-b+1} k(k-1)\binom{n-k}{b-1}\mathbb{E}T_k}. \tag{8.3.1}$$

This result provides the distribution of T_Δ in reasonably explicit form. If we define

$$S_k = T_n + \cdots + T_k,$$

then

$$\begin{aligned}
\mathbb{E}(T_k e^{-\phi(UT_k + T_{k+1} + \cdots + T_n)}) &= \mathbb{E}\left[\int_0^1 T_k e^{-\phi u T_k} du\, e^{-\phi(T_{k+1}+\cdots+T_n)}\right] \\
&= \mathbb{E}\left[\phi^{-1}(1 - e^{-\phi T_k}) e^{-\phi(T_{k+1}+\cdots+T_n)}\right] \\
&= \mathbb{E}\left[\phi^{-1} e^{-\phi(T_{k+1}+\cdots+T_n)} - \phi^{-1} e^{-\phi(T_k+\cdots+T_n)}\right] \\
&= \int_0^\infty e^{-\phi t}\{\mathbb{P}(S_{k+1} \le t) - \mathbb{P}(S_k \le t)\} dt \\
&= \int_0^\infty e^{-\phi t} \mathbb{P}(A_n(t) = k) dt,
\end{aligned}$$

the last equality following from the fact that the ancestral process $A_n(t) = k$ if, and only if, $S_k > t$ and $S_{k+1} \le t$. Hence we have

Theorem 8.3 *Assuming the times T_j have continuous distributions, the density of the age T_Δ is given by*

$$f_\Delta(t) = \frac{\sum_{k=2}^{n-b+1} k(k-1)\binom{n-k}{b-1}\mathbb{P}(A_n(t) = k)}{\sum_{k=2}^{n-b+1} k(k-1)\binom{n-k}{b-1}\mathbb{E}T_k}, \; t > 0. \tag{8.3.2}$$

Moments of $T_\Delta$can be found in a similar way, and one obtains

$$\mathbb{E}(T_\Delta^j) = \frac{\sum_{k=2}^n k(k-1)\binom{n-k}{b-1}\frac{1}{j+1}\mathbb{E}\left(S_k^{j+1} - S_{k+1}^{j+1}\right)}{\sum_{k=2}^n k(k-1)\binom{n-k}{b-1}\mathbb{E}(T_k)}, j = 1, 2\ldots, \quad (8.3.3)$$

from which the mean and variance of T_Δ can be obtained. For example, in the constant population size case, we obtain

$$\mathbb{E}(T_\Delta) = 2\binom{n-1}{b}^{-1}\sum_{j=2}^{n}\binom{n-j}{b-1}\frac{n-j+1}{n(j-1)}. \quad (8.3.4)$$

The age of an allele in the population

To derive the population version of (8.3.3), we assume that $\{A_n(t), t \geq 0\}$ converges in distribution to a process $\{A(t), t \geq 0\}$ as $n \to \infty$, and that the time taken for $A(\cdot)$ to reach 1 is finite with probability 1. Then as $n \to \infty$, and $b/n \to x$, $0 < x < 1$, we see that

$$\mathbb{E}(T_\Delta^j) = \frac{\sum_{k=2}^\infty k(k-1)(1-x)^{k-2}\frac{1}{j+1}\mathbb{E}\left(S_k^{j+1} - S_{k+1}^{j+1}\right)}{\sum_{k=2}^\infty k(k-1)(1-x)^{k-2}\mathbb{E}(T_k)}, j = 1, 2\ldots. \quad (8.3.5)$$

In this population limit the density of the age of a mutant gene that has a relative frequency x is, from Theorem (8.2),

$$\begin{aligned} g_x(t) &= \frac{\sum_{k=2}^\infty k(k-1)(1-x)^{k-2}\mathbb{P}(A(t)=k)}{\sum_{k=2}^\infty k(k-1)(1-x)^{k-2}\mathbb{E}(T_k)} \\ &= \frac{\mathbb{E}\left(A(t)(A(t)-1)(1-x)^{A(t)-2}\right)}{\sum_{k=2}^\infty k(k-1)(1-x)^{k-2}\mathbb{E}(T_k)}. \end{aligned} \quad (8.3.6)$$

The mean age of the mutation known to have frequency x in the population follows from (8.3.4) by letting $n \to \infty$, $b/n \to x$:

$$\mathbb{E}(T_\Delta) = \frac{-2x}{1-x}\log x. \quad (8.3.7)$$

Equation (8.3.4) is the well known formula derived by Kimura and Ohta (1973). The density (8.3.6) is also known in various forms (e.g. Watterson (1977) and Tavaré (1984).

Remark. There have been numerous papers written about the ages of alleles over the years, mostly using diffusion theory and reversibility arguments. This section sets the problem in a coalescent framework (although the results are much more general than they seem!). Watterson (1996) discusses Kimura's contribution to this problem. A modern perspective is given by Slatkin and Rannala (2000).

8.4 Simulating the age of an allele

An alternative to the analytical approach is to simulate observations from the joint conditional distribution of those features of the process that are of interest, for example the age T_Δ of the mutation Δ, and the time $T_{MRCA\Delta}$ to the MRCA of the individuals carrying Δ. In order to simulate such times, we can use the following algorithm based on Theorem 8.2.3 and (8.1.7).

Algorithm 8.1 To simulate from conditional distribution of T_Δ and $T_{MRCA\Delta}$.

1. Choose k according to the distribution of J_Δ in (8.1.7).
2. Choose j_1 from the conditional distribution of J_1 given $J_\Delta = k$ in (8.1.9) with $i = 1$.
3. Simulate an observation from the (unconditional) μ-biased joint distribution of the coalescence times $T_n, \ldots, T_{k+1}$.
4. Conditional on the results of step 3, simulate from the random variable Z having the (standard) size-biased distribution of T_k and set $T^* = UZ$, where U is an independent U(0,1) random variable.
5. Return $T_{MRCA\Delta} = T_n + \cdots + T_{j_1+1}, T_\Delta = T_{MRCA\Delta} + T_{j_1} + \cdots + T_{k+1} + T^*$.

Remark. Generating the appropriate size-biased distributions can be difficult when the population size varies. Another way to implement this is to replace steps 3 and 4 above with a rejection step:

3′. Generate $\boldsymbol{T} = (T_n, \ldots, T_2)$ from the coalescent model, and compute $L_n = 2T_2 + \cdots + nT_n$. Accept $\boldsymbol{T}$ with probability

$$\frac{T_k \mu}{2} e^{-T_k \mu/2} e\, e^{-L_n \mu/2}; \tag{8.4.1}$$

otherwise repeat.

4′. Set $T^* = UT_k$, where U is an independent U(0,1) random variable.

5′. Return $T_{MRCA\Delta} = T_n + \cdots + T_{j_1+1}, T_\Delta = T_{MRCA\Delta} + T_{j_1} + \cdots + T_{k+1} + T^*$.

The extra factor of e comes from the fact that $\mathrm{Po}(T_k\mu/2)\{1\} \le \mathrm{Po}(1)\{1\}$. In the limiting case $\mu = 0$ an independence sampler can be used.

8.5 Using intra-allelic variability

Rannala and Slatkin (1997) discussed a method for estimating the age of an allele known to have frequency b in a sample of size n, given an estimate of the number of mutations, m, that have arisen in the region around the mutation locus. There are at least three versions of this problem, depending on where these new mutations are assumed to occur. For example, we might sequence in the region of the mutation Δ and find the number of additional segregating sites in the region. We suppose once more that these additional mutations occur at rate $\theta/2$ on the branches of the coalescent tree.

If one wants to simulate observations from the posterior distribution of trees and times conditional on the number m of segregating sites appearing *in the b individuals carrying the mutation* in a region completely linked to Δ, then a modification of Algorithm 8.1 can be used:

Algorithm 8.2 To simulate from conditional distribution of age of mutation and $T_{MRCA\Delta}$ given m additional segregating sites in the Δ subtree.

1. Choose k according to the distribution of J_Δ in (8.1.7).
2. Choose $j_1, j_2, \ldots, j_{b-1}$ from the conditional distribution of $J_1, J_2, \ldots, J_{b-1}$ given $J_\Delta = k$ in (8.1.8).
3. Simulate an observation from the (unconditional) joint distribution of the coalescence times $T_n, \ldots, T_{k+1}$, and use the indices in step 2 to compute the coalescence times $T_b^*, \ldots, T_2^*$ in the Δ-subtree, together with the length $L_{nb} = \sum_{j=2}^{b} jT_j^*$ of the Δ-subtree.
4. Accept these statistics with probability

$$\mathrm{Po}(\theta L_{nb}/2)\{m\}/\mathrm{Po}(m)\{m\},$$

   else return to step 1.
5. Conditional on the results of step 3, simulate from the random variable Z having the size-biased distribution of T_k and set $T^* = UZ$, where U is an independent U(0,1) random variable.
6. Return $T_{MRCA\Delta} = T_n + \cdots + T_{j_1+1}, T_\Delta = T_{MRCA\Delta} + T_{j_1} + \cdots + T_{k+1} + T^*$.

Example

The conditional distribution of T_Δ and $T_{MRCA\Delta}$ in the constant population size case were simulated using 50,000 runs of Algorithm 8.2 for the case $n = 200, b = 30, \theta = 4.0$ and $m = 5$ segregating sites observed in the subtree. The mean age was 1.01 with standard deviation 0.91, while the mean subtree height was 0.40 with a standard deviation of 0.25. Percentiles of the distributions are given below, together with the estimated densities. For further details and alternative simulation algorithms, see Griffiths and Tavaré (2003).

	2.5%	25%	50%	75%	97.5%
age	0.156	0.412	0.721	1.289	3.544
subtree height	0.099	0.218	0.334	0.514	1.056

Fig. 8.2. Density of age of mutation.

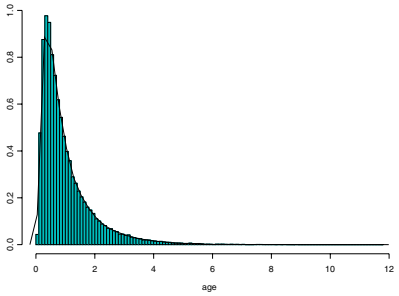
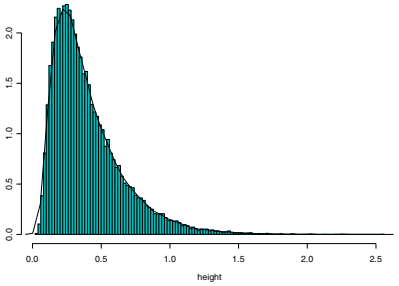


Fig. 8.3. Density of height of subtree.



9 Markov Chain Monte Carlo Methods

In this section we introduce some models for DNA sequence data, and explore some computer intensive methods that can be used to estimate population parameters. The main inference technique discussed here is Markov chain Monte Carlo, introduced into this field by Kuhner *et al.* (1995, 1998).

We assume that mutations occur on the coalescent tree of the sample at rate $\theta/2$, independently in each branch of the tree. Here we study the case in which the type space E is finite, and we suppose that the mutation process is determined by

$$\gamma_{ij} = \mathbb{P}(\text{mutation results in type } j \mid \text{ type was } i)$$

We write $\Gamma = (\gamma_{ij})$, and we note that γ_{ii} may be non-zero.

9.1 *K*-Allele models

One of the first models studied in any depth in this subject was the so-called K-allele model, in which $E = \{A_1, \ldots, A_K\}$ corresponding to K possible alleles in the type space. Let $X_i(t)$ denote the fraction of the population that has allele A_i at time t. Many of the results concern the diffusion model for the process $\{(X_1(t), \ldots, X_K(t)), t \geq 0\}$ with mutations determined according to the transition matrix Γ. The state space of the process is $\{\boldsymbol{x} = (x_1, \ldots, x_K) \in [0,1]^K : \sum_1^K x_i = 1\}$ and its generator has the form

$$L = \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} x_i(\delta_{ij} - x_j) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{j=1}^{K} \left(\sum_{i=1}^{K} x_i r_{ij} \right) \frac{\partial}{\partial x_j},$$

where

$$R = (r_{ij}) = \frac{\theta}{2}(\Gamma - I).$$

When the distribution of the type of a mutant is independent of its parental type, so that

$$\gamma_{ij} = \pi_j, \quad j \in E$$

where $\pi_j > 0, \sum_{j \in E} \pi_j = 1$, we recover the process studied in Section 3.1. The stationary distribution π of the diffusion is the Dirichlet distribution

$$\pi(x_1, \ldots, x_K) = \frac{\Gamma(\theta)}{\Gamma(\theta\pi_1) \cdots \Gamma(\theta\pi_K)} x_1^{\theta\pi_1 - 1} \cdots x_K^{\theta\pi_K - 1}. \tag{9.1.1}$$

Surprisingly perhaps, the distribution is known for essentially no other mutation matrices Γ. Suppose now that we take a sample of n genes from the stationary process with frequencies $(X_1, \ldots, X_K)$. The sample comprises n_i genes of type $i, 1 \leq i \leq K$. Writing $\boldsymbol{n} = (n_1, \ldots, n_K)$, the probability $q(\boldsymbol{n})$ that the sample has configuration $\boldsymbol{n}$ is

$$q(\boldsymbol{n}) = \mathbb{E} \frac{n!}{n_1! \cdots n_K!} X_1^{n_1} \cdots X_K^{n_K}. \tag{9.1.2}$$

For the model (9.1.1), this gives

$$\begin{aligned} q(\boldsymbol{n}) &= \int \cdots \int \frac{n!}{n_1! \cdots n_K!} x_1^{n_1} \cdots x_K^{n_K} \pi(x_1, \ldots, x_K) dx_1 \cdots dx_{K-1} \\ &= \frac{n! \Gamma(\theta) \Gamma(\theta\pi_1 + n_1) \cdots \Gamma(\theta\pi_K + n_K)}{n_1! \cdots n_K! \Gamma(\theta\pi_1) \cdots \Gamma(\theta\pi_K) \Gamma(\theta + n)} \\ &= \binom{\theta + n - 1}{n}^{-1} \prod_{j=1}^{K} \binom{\theta\pi_j + n_j - 1}{n_j}. \end{aligned} \tag{9.1.3}$$

In particular, the mean number of type i in the sample of size n is

$$\mathbb{E}(\text{number of allele } A_i) = n\mathbb{E}X_i = n\pi_i.$$

It is worth pointing out that a sample from any two-allele model can be described by (9.1.3), possibly after rescaling θ and Γ. To see this, suppose the matrix Γ has the form

$$\Gamma = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

Then the stationary distribution is $\boldsymbol{\pi} = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)$. Hence

$$\begin{aligned} R &\equiv \frac{\theta}{2}(\Gamma - I) \\ &= \frac{\theta}{2}\left(\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \\ &= \frac{\theta}{2}\begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} \\ &= \frac{\theta}{2}(\alpha+\beta)\begin{pmatrix} -\frac{\alpha}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & -\frac{\beta}{\alpha+\beta} \end{pmatrix} \\ &= \frac{\theta}{2}(\alpha+\beta)\left(\begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right). \end{aligned}$$

We may therefore use the sampling formula (9.1.3) with $\theta\pi_1$ replaced by $\theta\beta$, and $\theta\pi_2$ replaced by $\theta\alpha$.

The number of real mutations

Suppose that mutations occur at rate $\nu/2$ on the coalescent tree (the switch from θ to ν will be explained shortly). At any mutation point, the current allele is changed according to the transition matrix Γ. We note that not all potential substitutions have to result in changes to the existing allele, as $\gamma_{jj} > 0$ is allowed. The *effective mutation rate* $\theta/2$ is defined to be the expected number of mutations per unit time that result in a change of allele:

$$\frac{\theta}{2} = \frac{\nu}{2}\sum_{j=1}^{K}\pi_j(1-\gamma_{jj}), \tag{9.1.4}$$

where $\pi_j, j = 1, \ldots, K$ denotes the stationary distribution of Γ.

Felsenstein's model

It is convenient to describe here one useful model for the case $K = 4$, corresponding to models for the base at a given site in a DNA sequence. Here, $E = \{A, G, C, T\}$. Because many of our applications focus on mitochondrial

DNA, in which transitions occur with much higher frequency than transversions, we use a model which allows for transition-transversion bias.

Suppose then that mutations arise at rate $\nu/2$. When a potential substitution occurs, it may be one of two types: *general*, in which case an existing base j is substituted by a base of type k with probability $\pi_k, 1 \le j, k \le 4$; or *within-group*, in which case a pyrimidine is replaced by C or T with probability proportional to π_C and π_T respectively, and a purine is replaced by A or G with probability proportional to π_A and π_G respectively. The conditional probability of a general mutation is defined to be $1/(1+\kappa)$, while the conditional probability of a within-group mutation is defined to be $\kappa/(1+\kappa)$, where $\kappa \ge 0$ is the transition-transversion parameter. Thus the mutation matrix Γ is given by

$$\Gamma = \frac{1}{1+\kappa}\Gamma_1 + \frac{\kappa}{1+\kappa}\Gamma_2, \tag{9.1.5}$$

where $\Gamma_{1,ij} = \pi_j, j \in E$ and

$$\Gamma_2 = \begin{pmatrix} \frac{\pi_A}{\pi_A+\pi_G} & \frac{\pi_G}{\pi_A+\pi_G} & 0 & 0 \\ \frac{\pi_A}{\pi_A+\pi_G} & \frac{\pi_G}{\pi_A+\pi_G} & 0 & 0 \\ 0 & 0 & \frac{\pi_C}{\pi_C+\pi_T} & \frac{\pi_C}{\pi_C+\pi_T} \\ 0 & 0 & \frac{\pi_C}{\pi_C+\pi_T} & \frac{\pi_C}{\pi_C+\pi_T} \end{pmatrix}$$

In Γ_1 and Γ_2, the states are written in order A, G, C, T. It is readily checked that he stationary distribution of Γ is $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$. If we define

$$g = \frac{\nu}{2(1+\kappa)},\ w = \kappa g, \tag{9.1.6}$$

then κ is the ratio of the within-class to general substitution rates. From (9.1.4), the effective mutation rate is given by

$$\frac{\theta}{2} = g\left(1 - \sum_{j\in E}\pi_j^2\right) + 2w\left(\frac{\pi_A\pi_G}{\pi_A+\pi_G} + \frac{\pi_C\pi_T}{\pi_C+\pi_T}\right) \tag{9.1.7}$$

The transition matrix e^{Rt} of the mutation process with transition intensity matrix $R = \nu(\Gamma - I)/2$ is known. We denote the jk-th element by $r_{jk}(t)$; this is the probability that a base of type j has changed to a base of type k a time t later. Thorne *et al.* (1992) show that

$$r_{jk}(t) = \begin{cases} e^{-(g+w)t} + e^{-gt}(1-e^{-wt})\frac{\pi_k}{\pi_{H(k)}} + (1-e^{-gt})\pi_k & j = k \\ e^{-gt}(1-e^{-wt})\frac{\pi_k}{\pi_{H(k)}} + (1-e^{-gt})\pi_k,\ j \ne k, & H(j) = H(k) \\ (1-e^{-gt})\pi_k & H(j) \ne H(k) \end{cases}$$

where $\pi_R = \pi_A + \pi_G, \pi_Y = \pi_C + \pi_T$, and $H(i)$ denotes whether base i is a purine or a pyrimidine, so that $H(A) = H(G) = R$ and $H(C) = H(T) = Y$.

9.2 A biomolecular sequence model

Of particular interest to us is the case in which the types represent DNA or protein sequences of length s, say. Then the type space E has the form $E = E_0^s$, where E_0 is the type space of a single position, or site, in the sequence. The sites of the sequence may be labeled in many ways. The DNA alphabet $E_0 = \{A, C, G, T\}$ is one possibility, as is the 20 letter amino-acid sequence alphabet, or the 64 letter codon alphabet. Also common are the purine-pyrimidine alphabet, where $E_0 = \{Y, R\}$ and $Y = \{A, G\}$ denotes purines, $R = \{C, T\}$ the pyrimidines. In many evolutionary studies, transversions are not observed, and it might then be natural to think of sites as being binary, with $E_0 = \{A, G\}$ or $E_0 = \{C, T\}$. There are many possible models for the mutation process Γ, depending on what is assumed about the effects of mutation. Here we suppose that when a mutation occurs, it results in a substitution, the replacement of one element of E_0 by another one. The simplest version of this model supposes that the substitution occurs at site j with probability h_j, where

$$h_j \geq 0, \ \sum_{j=1}^{s} h_j = 1. \tag{9.2.1}$$

The h_j are identical (and so equal to $1/s$) if there are no mutational hotspots, and h_j may be 0 if site j is invariable. Thus the h_j add some flexibility in modeling variable mutation rates across the sequences. A mutation occurring at site j produces substitutions according to transition matrix $P_j = (p_{lm}^{(j)})$. Thus substitutions change a sequence of type $(i_1, \ldots, i_s)$ to one of type $(j_1, \ldots, j_s)$ as follows:

$$(i_1, \ldots, i_s) \to (i_1, \ldots, i_{l-1}, j_l, i_{l+1}, \ldots, i_s)$$

with probability $h_l p_{i_l j_l}^{(l)}, 1 \leq l \leq s$. We may write Γ in the form

$$\Gamma = \sum_{l=1}^{s} h_l I \otimes \cdots \otimes I \otimes P_l \otimes I \otimes \cdots \otimes I \tag{9.2.2}$$

where I denotes the identity matrix, and $\otimes$ denotes direct (or Kronecker) product: $A \otimes B = (a_{ij}B)$. Recall that if A, B, C, D are conformable matrices, then $(A \otimes B)(C \otimes D) = AC \otimes BD$. If $\boldsymbol{\pi}_l$ denotes the stationary distribution of P_l, and $\boldsymbol{\pi}$ denotes the stationary distribution of Γ, then it is easy to show that $\boldsymbol{\pi} = \boldsymbol{\pi}_1 \otimes \cdots \otimes \boldsymbol{\pi}_s$.

Many properties of this process may be studied using the coalescent simulation described in Section 6.6. The previous result shows that for simulating sequences from a stationary population, the ancestral sequence may be generated by simulating independently at each site, according to the stationary distribution of each site.

9.3 A recursion for sampling probabilities

Return now to the K-allele model with mutation matrix $\Gamma = (\gamma_{ij})$, and $R = \frac{\theta}{2}(\Gamma - I)$. Let $q(\boldsymbol{n})$ be the probability that a sample of n genes has a type configuration of $\boldsymbol{n} = (n_1, \ldots, n_K)$, and define $[K] = \{1, 2, \ldots, K\}$. A fundamental recursion is given in

Theorem 9.1

$$q(\boldsymbol{n}) = \frac{\theta}{n+\theta-1}\left(\sum_{i=1}^{K}\frac{n_i}{n}\gamma_{ii}q(\boldsymbol{n}) + \sum_{i,j\in[K],n_j>0,i\neq j}\frac{n_i+1}{n}\gamma_{ij}q(\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j)\right) + \frac{n-1}{n+\theta-1}\sum_{j\in[K],n_j>0}\frac{n_j-1}{n-1}q(\boldsymbol{n}-\boldsymbol{e}_j), \tag{9.3.1}$$

where $\{\boldsymbol{e}_i\}$ are the K unit vectors. Boundary conditions are required to determine the solution to (9.3.1). These have the form

$$q(\boldsymbol{e}_i) = \pi_i^*,\ i = 1, \ldots, K, \tag{9.3.2}$$

where π_i^* is the probability that the most recent common ancestor is of type i.

Proof. To derive (9.3.1) consider the first event back in time that happened in the ancestral tree. Relative rates of mutation and coalescence for n genes are $n\theta/2 : n(n-1)/2$, so the probability that the first event is a mutation is $\theta/(n+\theta-1)$. To obtain a configuration of $\boldsymbol{n}$ after mutation the configuration before must be either $\boldsymbol{n}$, and a transition $i \to i$ takes place for some $i \in [K]$ (the mutation resulted in no observable change), or $\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j, i, j \in [K], n_j > 0, i \neq j$ and a transition $i \to j$ take place. If a coalescence was the first event back in time, then to obtain a configuration $\boldsymbol{n}$ the configuration must be $\boldsymbol{n} - \boldsymbol{e}_j$ for some $j \in [K]$ with $n_j > 0$ and the ancestral lines involved in the coalescence must be of type j. □

The recursion in (9.3.1) is on n, the sample size. Given $\{q(\boldsymbol{m}); m < n\}$, simultaneous equations for the $\binom{n+K-1}{K-1}$ unknown probabilities $\{q(\boldsymbol{m}); m = n\}$ are non-singular, and in theory can be solved; cf. Lundstrom (1990). It is common to assume that

$$\pi_i^* = \pi_i, i = 1, \ldots, K, \tag{9.3.3}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ is the stationary distribution of Γ. With this assumption, $q(\boldsymbol{n})$ is the stationary sampling distribution.

It is worth emphasizing that the probability $q(\boldsymbol{n})$ satisfying (9.3.1) is determined solely by the rate matrix R. Indeed, (9.3.1) can be rewritten in the form

$$q(\boldsymbol{n}) =$$
$$\frac{2}{n(n-1)}\left(\sum_{i=1}^{K} n_i r_{ii} q(\boldsymbol{n}) + \sum_{i,j\in[K],n_j>0,i\neq j} (n_i+1) r_{ij} q(\boldsymbol{n}+\boldsymbol{e}_i+\boldsymbol{e}_j)\right)$$
$$+\frac{1}{n-1}\sum_{j\in[K],n_j>0}(n_j-1)(\boldsymbol{n}-\boldsymbol{e}_j).$$

The point here is that different combinations of θ and Γ can give rise to the same R matrix. Nonetheless, we prefer to think of the model in terms of an overall rate θ and a matrix of substitution probabilities Γ. In practice, we often assume that Γ is known, and the aim might then be to estimate the single parameter θ, which reflects both the effective population size N and the mutation probability u.

Remark. The recursion in (9.3.1) has appeared in a number of guises in the literature, such as Sawyer *et al.* (1987) and Lundstrom *et al.* (1992). In the latter references, a quasi-likelihood approach for estimation of θ in the finitely-many-sites model is developed. The recursion (9.3.1) is used to find the probability distribution at each site, and the quasi-likelihood is computed by assuming independence across the sites.

Griffiths and Tavaré (1994) used the recursion for the finitely-many-sites model to find the likelihood. Conventional numerical solutions in this case are difficult to obtain because of the large number of equations. This prompted them to develop their Markov chain approach. See Forsythe and Leibler (1950) for an early application of Monte Carlo approaches to matrix inversion. We note here that early experience with the Griffiths-Tavaré method suggests it is not feasible for analyzing large amounts of sequence data. In the remainder of this section, we discuss a Markov chain Monte Carlo approach and give a number of examples of its use.

9.4 Computing probabilities on trees

For definiteness, assume we are dealing with DNA sequence data $\mathcal{D}$ having s aligned sites in a sample of size n. We will use Λ to denote the (labeled) coalescent tree topology, and $\boldsymbol{T} = (T_2, \ldots, T_n)$ to denote the coalescence times in the tree. For a given model of substitution at a particular site in the sequence, we will need to compute the probability of the bases in the sample, given a particular value of Λ and $\boldsymbol{T}$. This can be done using a recursive method, known as the *peeling algorithm*, described by Felsenstein (1973, 1981). The idea is to compute the probability of the bases $b_1, \ldots, b_n$ observed at a particular position in sequences 1,...,n. Each node l in the tree is assigned a vector of length 4, the i-th entry of which gives the probability of the data below that node, assuming node l is base i. The algorithm is initialized by assigning the vector associated with a leaf i the vector with elements $\delta_{b_i,j}, j = 1, \ldots, 4$.

The calculation now proceeds recursively. Imagine that the probability vectors $(w_{u1}, \ldots, w_{u4})$ and $(w_{v1}, \ldots, w_{v4})$ have been computed for the descendant nodes u and v respectively of node l. To compute the vector $(w_{l1}, \ldots, w_{l4})$ at node l, we need to calculate the time t_{lu} along the branch from $l \to u$, and the time t_{lv} from $l \to v$. Then we calculate

$$w_{lz} = \left(\sum_x r_{zx}(t_{lu}) w_{ux} \right) \cdot \left(\sum_y r_{zy}(t_{lv}) w_{vy} \right),$$

where $r_{ij}(t)$ is the probability that base i has mutated to base j a time t later.

This scheme allows us to recurse up to the root of the tree. That node has label $l = 2n - 1$ and descendant nodes u and v. We finish the computation of the probability L of the configuration at that site by computing

$$L = \sum_z \pi_z^0 w_{uz} w_{vz}$$

where $\pi_z^0, z = 1, \ldots, 4$ is the distribution of the ancestral base.

Once the likelihood at a single base position is calculated, the likelihood of the set of n sequences can be calculated using the fact that for the mutation model in Section 9.2 the sites evolve independently, conditional on Λ and $\boldsymbol{T}$. Hence if L_i denotes the likelihood of the i-th site, the overall likelihood is

$$\mathbb{P}(\mathcal{D} \mid \Lambda, \boldsymbol{T}) = \prod_{i=1}^{s} L_i. \tag{9.4.1}$$

9.5 The MCMC approach

Here we discuss a version of the Metropolis-Hastings algorithm, due originally to Metropolis *et al.* (1953) and Hastings (1970) that will be exploited for inference on coalescent trees. Our presentation follows that of Markovtsova (2000). The algorithm produces correlated samples from a posterior distribution π of interest, in our case $\pi(G) \equiv f(G \mid \mathcal{D})$, where $G \equiv (\Lambda, \boldsymbol{T}, M)$, M representing the mutation parameters and $\mathcal{D}$ representing the sequence data. We use these samples to make inferences about parameters and statistics of interest. Examples include the effective mutation rate θ, the time to the most recent common ancestor, ages of a particular event in the sample, or population growth rates. We can write

$$f(G \mid \mathcal{D}) = \mathbb{P}(\mathcal{D} \mid G) g_1(\Lambda) g_2(\boldsymbol{T}) g_3(M) / f(\mathcal{D}). \tag{9.5.1}$$

The first term on the right can be computed using the peeling algorithm described in the last section and an appropriate model for mutation among the sequences. The term $g_1(\Lambda)$ on the right of (9.5.1) is the coalescent tree topology distribution, $g_2(\boldsymbol{T})$ is the density of the coalescence times $\boldsymbol{T}$, and $g_3(M)$ is the

prior distribution for the mutation parameters M. The normalizing constant $f(\mathcal{D})$ is unknown and hard to compute. The algorithm starts with an arbitrary choice of Λ, $\boldsymbol{T}$ and M. New realizations of G are then proposed, and accepted or rejected, according to the following scheme.

Algorithm 9.1 Basic Metropolis-Hastings method:

1. Denote the current state by $G = (\Lambda, \boldsymbol{T}, M)$.
2. Output the current value of G.
3. Propose $G' = (\Lambda', \boldsymbol{T}', M')$ according to a kernel $Q(G \to G')$.
4. Compute the Hastings ratio

$$h = \min\left\{1, \quad \frac{\pi(G')Q(G' \to G)}{\pi(G)Q(G \to G')}\right\}. \tag{9.5.2}$$

5. Accept the new state G' with probability h, otherwise stay at G.
6. Return to step 1.

Let $X(t)$ denote the state of this chain after t iterations. Once $X(t)$ has 'reached stationarity' its values represent samples from the distribution $\pi(G) = \pi(\Lambda, \boldsymbol{T}, M)$. The nature of the algorithm is such that consecutive outputs will be correlated. For many problems this might be not a bad thing, however one should be careful with using the output for calculating standard errors. But in some cases it is desirable to simulate approximately independent samples from the posterior distribution of interest, in which case we use output from every m^{th} iteration, for a suitable choice of m.

Current methods

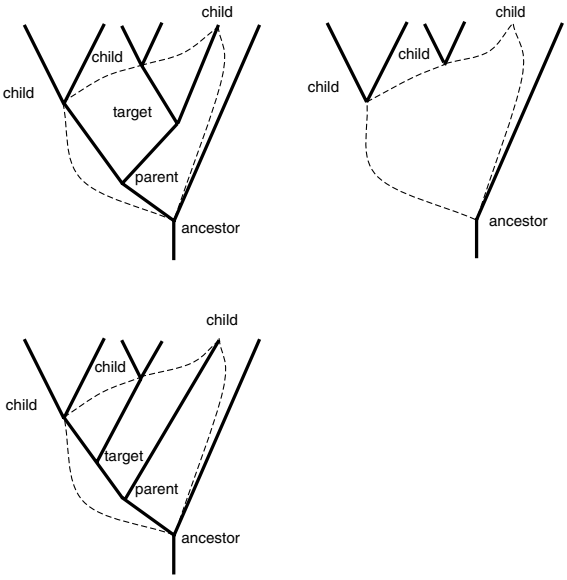
In this section we describe some methods of sampling genealogies. Most of these algorithms are very similar and often differ only in tree representation and useful tricks to speed up the computations. All of them start with an initial genealogy (random or UPGMA) and make small modifications to it. Choices among possible modifications may be random or deterministic.

The first is due to Kuhner *et al.* (1995). As before, the genealogy consists of two parts: the tree topology and a set of times between coalescent events, but time is rescaled in terms of the overall mutation rate in such a way that in one unit of time the expected number of mutations per site is 1. Figure 9.1 shows the updating process: choosing a neighborhood (the region of genealogy to be changed), rearranging the topology in that neighborhood, and choosing new branch lengths within the neighborhood. This fundamental operation is applied repeatedly. To make rearrangements, a node is chosen at random from among all nodes that have both parents and children (i.e., are neither leaves nor the bottom-most node of the genealogy). This node is referred to as the target. The neighborhood of rearrangement consists of the target node, its

child, parent, and parent's other child. A rearrangement makes changes of two types: reassorts the tree children among target and parent, and modifies the branch length within the neighborhood. The lineages to be redrawn are referred to as active lineages, and the lineages outside of the neighborhood as inactive lineages.

The times of the target and parent nodes are drawn from a conditional coalescent distribution with the given mutation rate, conditioned on the number of inactive lineages. For each time interval, the probability of coalescence among the active lineages depends on the number of active and inactive lineages present in the genealogy during that time interval. A random walk, weighted by these probabilities, is used to select a specific set of times.

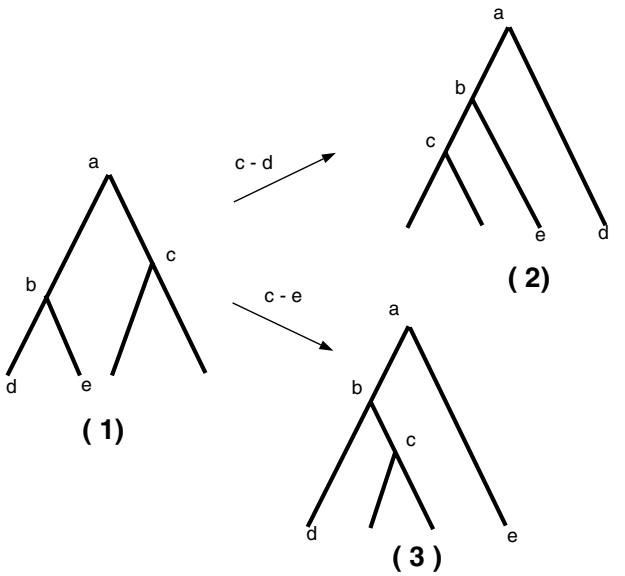
Fig. 9.1. Steps in rearranging a genealogy. Top left: selecting a neighborhood. Top right: erasing the active lineages. Bottom: redrawing the active lineages.



Yang and Rannala (1997) use a stochastic representation of the nearest neighbor interchange (NNI) algorithm as a core of the transition kernel. This algorithm generates two neighboring topologies for each internal branch (see Figure 9.2). Consider an interior branch $a - b$, where a is the ancestral node and b is the descendant node. Node c is the other descendant of a, and nodes d and e are descendants of b. The two neighbors of tree 1 are generated by interchanging node c with node d (tree 2), and node c with node e (tree 3).

Equal probabilities are assigned to each of the neighboring topologies. The NNI algorithm modifies the topology but ignores the ordering of the nodes

Fig. 9.2. NNI algorithm for a rooted binary tree topology.



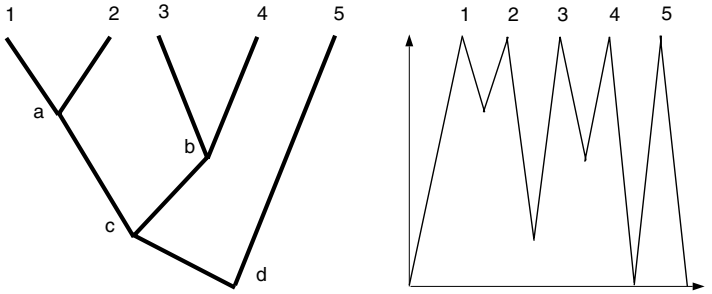
(i.e., labeled history). To modify the NNI algorithm so that the chain moves between labeled histories, they assign an equal probability to each of the possible labeled histories for a nominated topology. This involves enumerating and recording all the labeled histories for that topology. The move to another labeled history that belongs to the current tree topology is allowed with the specified probability if the topology has more than one labeled history. Yang and Rannala use this transition kernel in the study of species data; the time to the MRCA is scaled to be 1 and times between speciation events have different distributions than those specified by the coalescent.

Wilson and Balding (1998) designed an algorithm to deal with microsatellite (or short tandem repeat) data. A step-wise model is chosen for the changes in repeat number at each mutation event. Although calculation of the likelihood via peeling is feasible for problems of moderate size, increasing the dimension of the parameter space by introducing the allelic state of the internal nodes permits much faster likelihood calculations. The algorithm uses a very simple method for generating candidate trees. It involves removing a branch from the tree at random and adding it anywhere in the tree, but locations close to similar allelic types are preferentially chosen.

Larget and Simon (1999) use an algorithm for moving in a tree space that is very close to the one developed by Mau *et al.* (1999). It uses the fact that for a given choice of ordering all sub-trees from left to right there is a unique in-order traversal of the tree. Each internal node is adjacent to two leaves in this traversal, the right-most leaf of its left sub-tree and the left most leaf

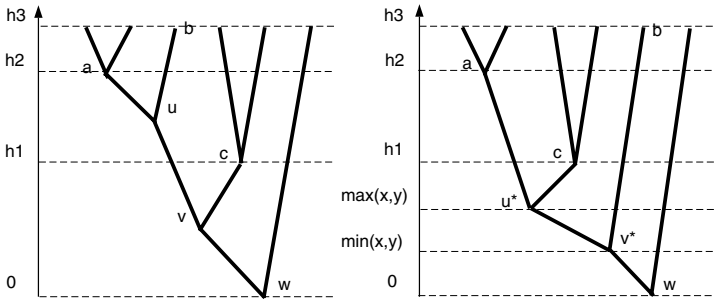
of its right sub-tree. Given the ordering of the nodes and distances between adjacent nodes, the tree topology and branch lengths are uniquely determined. Each taxon appears at a peak of the graph, and each internal node is a valley (see Figure 9.3).

Fig. 9.3. A tree representation.



The transition kernel consists of two different moves: global and local. For a global move one representation of the current tree is selected uniformly at random by choosing left/right orientation of the two sub-trees with equal probability for each internal node. Then the valley depths are simultaneously and independently modified by adding to each a perturbation in either direction, keeping the depth between 0 and a specified maximum. The local move modifies a tree only in a small neighborhood of a randomly chosen internal branch, leaving the remainder of the tree unchanged. Let u and v be the nodes joined by the randomly chosen edge (see Figure 9.4).

Fig. 9.4. A tree before and after a local move.



Leaving positions of a, b, c, and w fixed, new positions for nodes u and v are picked. Let $h_1 < h_2 < h_3$ be the distances between c and w, a and w, and

b and w correspondingly. In the local move, x is chosen uniformly at random from $[0, h_2]$, and y is chosen uniformly at random from $[0, h_1]$. Proposed nodes u^* and v^* will be distances $\max(x, y)$ and $\min(x, y)$ from w, respectively. If $\max(x, y) < h_1$, there are three possible tree topologies. One of the children, a, b, and c, is randomly chosen to be joined to v^*, with the others becoming children of u^*. If v is the root of the tree, the distances between v and the children a, b, and c are changed and the new location of u is chosen. The local move is very similar in character to the method of Kuhner *et al.* (1995).

9.6 Some alternative updating methods

We have some freedom in choosing the proposal kernel $Q(\cdot, \cdot)$. Ideally $Q(\cdot, \cdot)$ is relatively easy to calculate since the scheme above may need to iterated many times in order to converge to stationarity. Furthermore we have to demonstrate that the chain $X(t)$ satisfies the conditions of irreducibility and positive recurrence in order to show that the ergodic theorem applies and so the limiting distribution is indeed $f(\Lambda, \boldsymbol{T}, M \mid \mathcal{D})$.

We define level l of the genealogy to be the first point at which there are l distinct ancestors of the sample. The bottom of a genealogy of n individuals is therefore referred to as level n, and the MRCA of the sample is level 1. Recall that T_l denotes the time between levels l and $l-1$. To propose a new graph $(\Lambda', \boldsymbol{T}')$ we considered three different proposal kernels.

A bad sampler

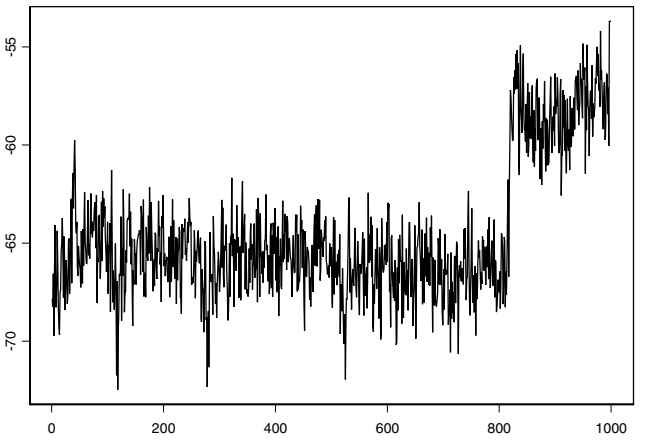
Here is a simple algorithm:

1. Pick a level, l say $(l = n, n-1, \ldots, 2)$, according to an arbitrary distribution F.
2. Delete upper part of the tree starting from level l.
3. Attach a new top of the tree generated according to the coalescent prior for a sample of l individuals.
4. Generate a new time T_l', to replace the old T_l according to an exponential distribution with parameter $l(l-1)/2$.

This algorithm works poorly, mainly because the suggested changes were too global. If we chose level l close to the bottom of the tree and attach a random top to it, then the new tree will be very different from the old one and has small chance of being accepted. As a result our sample will consists of trees with similar topologies and almost the same likelihood. But sometimes quite a different tree might be accepted and our Markov chain would move to other part of state space and stay there for long time. Figure 9.5 is an example of such a chain. This algorithm seems not to be very efficient in exploring the state space of trees.

The following algorithm looks simple and is easy to implement. It makes changes which are more local than the algorithm described above.

Fig. 9.5. Time series plot of log-likelihood



1. Pick a level, l say ($l = n, n-1, \ldots, 2$), according to an arbitrary distribution F.
2. Label the l lines $1, 2, \ldots, l$.
3. Let L_i and L_j be the two lines which coalesce.
4. With probability $1/2$ replace this coalescence by one between L_i and a randomly chosen line (possibly resulting in the same topology as before).
5. Otherwise replace this coalescence by one between L_j and a randomly chosen line (also possibly resulting in the same topology as before).
6. Generate a new time T_l', to replace the old T_l according to an exponential distribution with parameter $l(l-1)/2$.

An example of a possible move, for a genealogy of five individuals, is shown in Figure 9.6.

This algorithm also does not work well, primarily because it is relatively hard to switch the order of two coalescence events. For example, we need several iterations of the algorithm to move from G to G' as illustrated in Figure 9.7.

Theoretically, this kernel has all the required properties, but it is simply not efficient. We might try other distributions for the choice of level l, or for the new time T_l', but it is doubtful these would help. Our experience was that the algorithm became stuck in local maxima which required a re-ordering of coalescences in order to escape.

Fig. 9.6. A move in the sampler

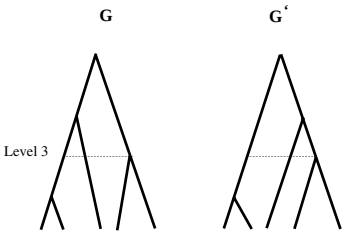
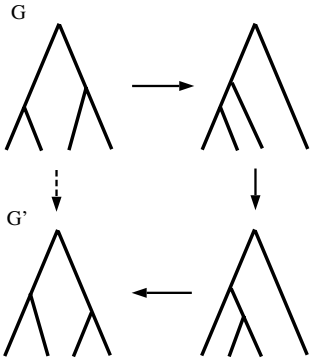


Fig. 9.7. Change of order of two coalescences



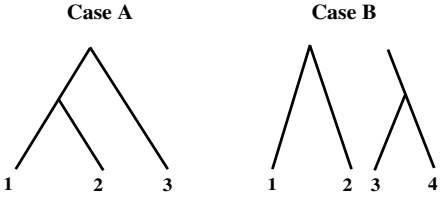
A good sampler

Lack of success with first two algorithms leads to the following approach, described in Markovtsova *et al.* (2000).

Algorithm 9.2 Local updating method.

1. Pick a level, l say ($l = n, n-1, \ldots, 3$), according to an arbitrary distribution F.
2. For the chosen l observe the pattern of coalescence at levels $l-1$ and $l-2$. This pattern falls into two cases, according to whether the coalescence at level $l-2$ involves the line which results from the coalescence at level $l-1$. These are illustrated in Figure 9.8. In Case A our kernel randomly generates a new topology involving the same three lines of ancestry; this new topology will also be Case A and may be the same topology with which we began. These are illustrated in Figure 9.9. In Case B we change

Fig. 9.8. Two possible coalescence patterns

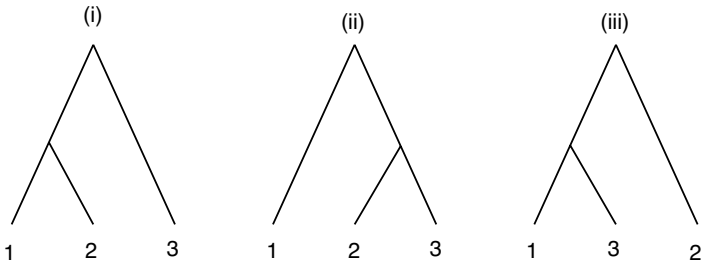


the order of the two coalescence events. So, for the example drawn above, we move to the state shown in Figure 9.10.

3. Generate new times T_l' and T_{l-1}' according to an arbitrary distribution, and leave other times unchanged. Thus we only alter the times corresponding to the levels at which the topology has been changed. This ensures that $(\Lambda', \boldsymbol{T}')$ is similar to $(\Lambda, \boldsymbol{T})$ and therefore has a reasonable probability of being accepted.

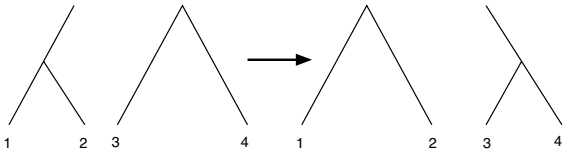
There are several variants of Step 2 of the above scheme. For example, one can allow the topology to remain the same in Case B, but not in Case A. We also tried a variant of Case B in which we proposed a new Case B topology uniformly from the six possible choices in which the four lines are paired randomly. None of these variations impacts significantly on the results.

Fig. 9.9. Possible moves in Case A



There are many possible choices for the updating times T_l' and T_{l-1}'. One might propose new values of T_j' from the pre-data coalescent distribution as it was done in first two algorithms. Second, one might generate times from a Normal distribution with mean equal to the currently accepted value T_j. We chose to truncate the Normal distribution in order to ensure that negative times were not proposed. The variances of the Normal distributions are parameters that can be tuned to get good mixing properties. Unfortunately,

Fig. 9.10. Possible moves in Case B



the optimal choice of variance appears to be highly data-dependent. In principle all choices are valid, but the rate of approach to stationarity, and the correlation between consecutive iterations, can vary significantly. The second approach might work better when trees are much shorter, or longer, than would be expected *a priori*.

Finally, we update the mutation parameter $M = (g)$ every k iterations. There are several ways to do it. First one is to propose new value g' from prior distribution. This updating mechanism works well in the case when the prior for g is very concentrated, i.e. a uniform with narrow support or a Normal distribution with small variance. This approach might be used when some external information is available. Second one is to generate new value g' according to truncated Normal distribution with mean g. The variance of this distribution requires some tuning to ensure well-behaved, i.e. uncorrelated, output. This approach works fine in case of uninformative prior or prior with wide support.

The Hastings ratio

Writing $G = (\Lambda, \boldsymbol{T}, M)$, the kernel Q can be expressed as the product of three terms:

$$Q(G \to G') = Q_1(\Lambda \to \Lambda')\, Q_2(\boldsymbol{T} \to \boldsymbol{T}' \mid \Lambda \to \Lambda')\, Q_3(M \to M').$$

Consequently the Hastings ratio can be written in the form

$$h = \min\left\{1, \quad \frac{\mathbb{P}(\mathcal{D} \mid G')}{\mathbb{P}(\mathcal{D} \mid G)} \frac{g_1(\Lambda')}{g_1(\Lambda)} \frac{g_2(\boldsymbol{T}')}{g_2(\boldsymbol{T})} \frac{g_3(M')}{g_3(M)} \right.$$
$$\left. \times \frac{Q_1(\Lambda' \to \Lambda)}{Q_1(\Lambda \to \Lambda')} \frac{Q_2(\boldsymbol{T}' \to \boldsymbol{T} \mid \Lambda' \to \Lambda)}{Q_2(\boldsymbol{T} \to \boldsymbol{T}' \mid \Lambda \to \Lambda')} \frac{Q_3(M' \to M)}{Q_3(M \to M')} \right\}, \quad (9.6.1)$$

the unknown term $f(\mathbf{D})$ cancelling. We can further simplify (9.6.1) by noting that, since pairs of lines are chosen uniformly to coalesce, all topologies are, *a priori*, equally likely. Hence $g_1(\Lambda') = g_1(\Lambda)$. Furthermore, our transition

kernel changes only two of the times on the tree, T_l and T_{l-1} say. Finally, it is easy to show that $Q_1(\Lambda \to \Lambda') = Q_1(\Lambda' \to \Lambda)$, reducing (9.6.1) to

$$h = \min\left\{1,\ \frac{\mathbb{P}(\mathcal{D} \mid G')}{\mathbb{P}(\mathcal{D} \mid G)}\ \frac{g_2(\boldsymbol{T}')g_3(M')}{g_2(\boldsymbol{T})g_3(M)}\ \frac{f_l(t_l)f_{l-1}(t_{l-1})}{f_l(t'_l)f_{l-1}(t'_{l-1})}\ \frac{Q_3(M' \to M)}{Q_3(M \to M')}\right\}, \tag{9.6.2}$$

where $f_l(\cdot)$ and $f_{l-1}(\cdot)$ are the densities of the time updating mechanism at levels l and $l-1$.

If one uses a transition kernel which proposes new times that are exponential with parameter $l(l-1)/2$ at level l, (*i.e.* the unconditional coalescent distribution for times), then further cross-cancellation reduces (9.6.2) to

$$h = \min\left\{1,\ \frac{\mathbb{P}(\mathcal{D} \mid G')}{\mathbb{P}(\mathcal{D} \mid G)}\ \frac{g_3(M')}{g_3(M)}\ \frac{Q_3(M' \to M)}{Q_3(M \to M')}\right\}. \tag{9.6.3}$$

A similar simplification also follows if one proposes new mutation rates independently of the currently accepted rate and

$$h = \min\left\{1,\ \frac{\mathbb{P}(\mathcal{D} \mid G')}{\mathbb{P}(\mathcal{D} \mid G)}\right\}. \tag{9.6.4}$$

In order to test the algorithm for moving around tree space, we can use a simple mutation model for which there are alternative algorithms. One obvious choice is the infinitely-many-sites model, for which we have already developed some theory in Section 7. The data take the form of the number of segregating sites in the sample, and Algorithm 7.3 can be used to generate observations from the posterior distribution of features of the tree, conditional on the number of segregating sites observed.

9.7 Variable population size

The methods discussed in above can easily be adapted to model populations which are not of a fixed constant size. As in Section 2.4, let $N(t)$ denote the population size at time t, where time is measured in units of $N = N(0)$ generations, and write

$$N(t) = f(t)N(0), \quad \Lambda(t) = \int_0^t \frac{1}{f(u)}du.$$

If $A_n(t)$ is the ancestral process for a sample of size n evolving in a population of constant size, $A_n^v(t) = A_n(\Lambda(t))$ is the coalescent process appropriate for the population of varying size.

We let T_k^v record the coalescent time spent with k lines of descent in a growing population. The algorithm works by manipulating the underlying *coalescent* times, $\{T_i\}$, defined on the original coalescent time-scale, and subsequently transforming them to times in the varying population while calculating probability of data given tree.

Define $S_i = \sum_{j=i+1}^{n} T_j$. S_i represents the amount of standard coalescent time taken to get to a level with i lines of descent present. Similarly, $S_i^v = \sum_{j=i+1}^{n} T_j^v$ in the varying population. We transform the S_i to the S_i^v via $S_i^v = \min\{s : \Lambda(s) = S_i\}$. The proposal kernel works by manipulating the underlying coalescent times, $\{T_i\}$. Assuming we have picked level l in our updating step, new times $T_l^{v'}$, $T_{l-1}^{v\ '}$ are proposed as follows. We begin by generating new times $T_l' = t_l'$ and $T_{l-1}' = t_{l-1}'$. Having done so, we recalculate S_k for all $k \leq l$. From these values we derive the new $\{S_i^{v'}\}$, noting that $S_i^{v'} = S_i^v$ for $i > l$.

9.8 A Nuu Chah Nulth data set

We illustrate our approach with a sample of mitochondrial sequences from the Nuu Chah Nulth obtained by Ward *et al.* (1991). The data D are 360 bp sequences from region I of the control region obtained from a sample of $n = 63$ individuals. The observed base frequencies are $(\pi_A, \pi_G, \pi_C, \pi_T) = (0.3297, 0.1120, 0.3371, 0.2212)$. The data have 26 segregating sites and a mean heterozygosity of 0.0145 per site. There are 28 distinct haplotypes with a haplotype homozygosity of 0.0562. We fit two models to these data, both of which are variants of Felsenstein's model described in Section 9.2:

Model 1. All sites mutate at the same rate, so that $g_i \equiv g$ for all sites i. Here $M = (g, \kappa)$.
Model 2. The special case of Model 1 in which κ is assumed known, so that $M = (g)$.

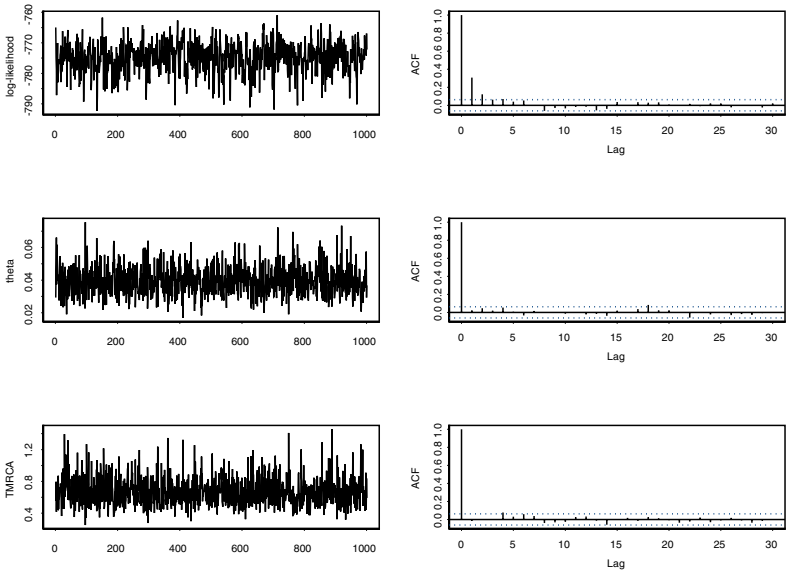
Model 2 above serves as the simplest description of mutation in hypervariable region I of mtDNA. It was used by Kuhner *et al.* (1995) in their analysis of the same data set.

We implemented the MCMC approach described in Algorithm 9.2. One should begin to sample from the process $X(\cdot)$ once it has "reached stationarity". There are many heuristic tests for this, none of which is infallible. For a critique see Gilks *et al.* (1996). Some simple diagnostics are functions of the statistics of interest such as autocorrelations and moving averages. It is also valuable to run the chain from several different, widely spaced, starting points, and compare the long-term behavior.

The output typically appeared to be non-stationary for up to 200,000 iterations of the algorithm. We sampled every 10,000th iteration in order to approximate a random sample from the stationary distribution. In a bid to be very conservative, and since the algorithms run rapidly, we generally discarded the first 2500 samples. After this, our output is typically based on 5000 samples. The acceptance rate was typically around 80%. For runs in which, for example, we needed to tune the variance parameter, the burn-in length varied but the estimated parameter values were unchanged for the different variances we tried.

Figure 9.11 shows the resultant time series for the log-likelihood, the mutation parameter θ, the time to the MRCA and their associated autocorrelation functions. These appear fine, with the proviso that the time series of log-likelihoods is correlated for several lags. While this is not in itself a problem it means one must interpret standard errors with care. As a further check for convergence to stationarity we used the package of diagnostics provided in CODA (Best *et al.* (1995)). All tests were passed.

Fig. 9.11. Example diagnostics



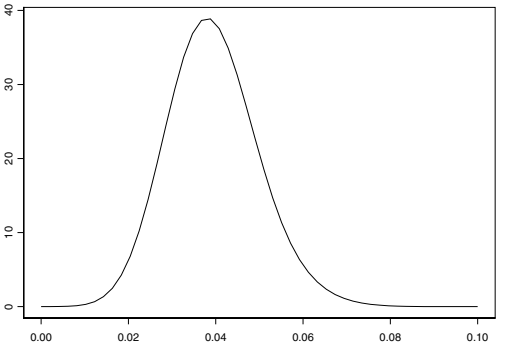
Some time can be saved by starting the process from a genealogy $(\Lambda, \boldsymbol{T})$ for which $\mathbb{P}(\Lambda, \boldsymbol{T} \mid \mathcal{D})$ is relatively high. The rationale for this is that it is sensible to start from a region of the state-space which is well supported by the data. As an example of this one might use the UPGMA tree for the data-set, as described in Kuhner *et al.* (1995). However, we prefer to start from random tree topologies since convergence from different starting points is potentially a useful diagnostic for stationarity.

The analysis of Model 1 gave a median for κ of 65.1, with 25th and 75th percentiles of 32.7 and 162.7 respectively. Note that the data are consistent with no transversions having occurred during the evolution of the sample. Consequently, the posterior distribution for κ has a very long right tail and statistics for the mean, which are strongly influenced by outliers, are poten-

tially misleading and are therefore not presented. The median value of g was 6.87×10^{-4} and the median value for w was 4.47×10^{-2}. These results show that the data are consistent with a value of $\kappa = 100$, as assumed by Kuhner *et al.* (1995).

In what follows we also took $\kappa = 100$, and a uniform prior on (0, 100) for θ. The posterior distribution of the effective mutation rate has a median of 0.038, mean 0.039 and 25th and 75th percentiles of 0.033 and 0.045 respectively. Figure 9.12 shows the posterior distribution of θ.

Fig. 9.12. Posterior density of per site effective mutation rate θ



Since the posterior density of θ is proportional to the likelihood in this case, we may use an estimate of the posterior density to find the maximum likelihood estimate of θ. From the density shown in Figure 9.12, we obtained an MLE of $\hat{\theta} = 0.038$. Kuhner *et al.* (1995) obtained the value $\hat{\theta} = 0.040$ for these data, using the same value of κ. Presumably the difference in the estimates arises from both the parameters chosen for the density estimation, and the different approaches to the optimization. From an estimate of the curvature of the log-density we get an estimate of the standard error of $\hat{\theta}$ of 0.010, resulting in an approximate 95% confidence interval of (0.018, 0.058).

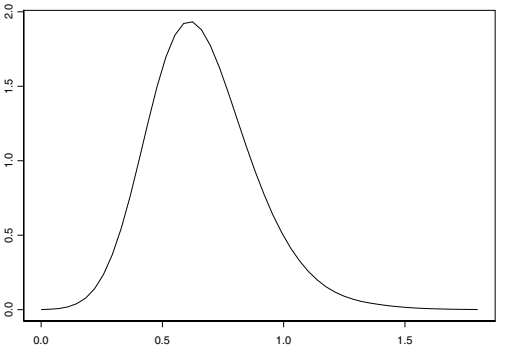
Remark. Estimates of standard errors based on curvature of the log-density should be treated as heuristic. In problems such as these, θ cannot be estimated consistently so the standard theory does not apply.

For comparison, the Watterson estimator (5.3.7) of θ, based on 26 segregating sites in the data, is 0.015 with an estimated standard error of 0.005; the 95% confidence interval for θ is then (0.005, 0.025). The lower MLE obtained

using the Watterson estimator is expected, because multiple mutations at the same site are ignored.

The prior distribution of the time to MRCA of a sample of $n = 63$ has a mean of $2(1 - 1/63) = 1.97$. With an effective size of $N = 600$, a 20 year generation time and a value of $\sigma^2 = 1$ for the variance of the offspring distribution, this is about 23,600 years. The posterior distribution of the time T_{MRCA} to the MRCA (in years) has median 7700, mean 8100 and 25th and 75th percentiles of 6500 and 9300 respectively. The corresponding posterior density appears in Figure 9.13. The joint posterior density of T_{MRCA} and θ is given in Figure 9.14. For a frequentist approach to inference about T_{MRCA}, see Tang *et al.* (2002).

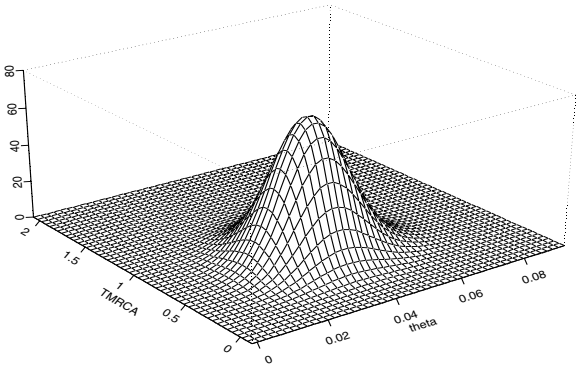
Fig. 9.13. Posterior density of time to MRCA



Testing goodness-of-fit

The adequacy of the fit of models like these can be assessed using the Bayesian posterior predictive distribution. To implement this, we use a variant of the parametric bootstrap. The idea is to simulate observations from the *posterior* distribution of $(\Lambda, \boldsymbol{T}, M)$, and then for each of the trees $(\Lambda, \boldsymbol{T})$ to simulate the mutation process with parameters specified by M. The distribution of certain summary statistics observed in the simulated data is found, and the values of the statistics actually observed in the data are compared to these distributions. We chose to use the number of haplotypes, the maximal haplotype frequency, the haplotype homozygosity, the number of segregating sites and a measure of nucleotide diversity. In practice, we use the output from the MCMC runs to generate the observations on $(\Lambda, \boldsymbol{T}, M)$. In Table 11 we give the results of this comparison for Model 2 using 4000 values from each

Fig. 9.14. Joint posterior density of TMRCA and θ



posterior distribution. There is some evidence that the constant rate model does not fit well, particularly regarding the haplotype distribution. The total number of segregating sites observed in the bootstrap samples gives some evidence of lack-of-fit; the model predicts more segregating sites than are seen in the data. One explanation for this apparent discrepancy might be that the model is not allowing for rate heterogeneity, and therefore does not typically produce enough recurrent mutations. This will lead to a tendency for the mutations which do occur to be spread over a greater number of sites. A model that allows for multiple classes of rates appears in Markovtsova *et al.* (2000b).

Remark. For an implementation of Bayesian methods for the coalescent (and many other species tree problems), using Metropolis-coupled MCMC, see Huelsenbeck and Ronquist's *MrBayes* program, at

`http://morphbank.ebc.uu.se/mrbayes/info.php` .

9.9 The age of a UEP

In this section we provide an MCMC approach that can be used to find the posterior distribution of the age of a unique event polymorphism (UEP). As in the introduction of Section 8, there are several versions of this problem. For the most part, we assume that we have sequenced a region of DNA, and have determined for each of them whether or not the UEP is present. The key figure is given in Figure 8.1. Let $\mathcal{U}$ denote the single event that causes the UEP mutation Δ. The scaled mutation rate at the UEP locus is $\mu/2$. The event that the coalescent tree has the UEP property is, once again, denoted by $\mathcal{E}$. For definiteness we assume that the sequences are evolving according to

Table 11. Assessing goodness-of-fit of Model 2

Statistic	Observed value	Model 2 Fraction of simulations $\leq$ observed value
# haplotypes	28	0.83
max. haplotype frequency	9	0.36
homozygosity	0.0562	0.12
heterozygosity per site	0.0145	0.36
# segregating sites	26	0.05

Felsenstein's model. The material in this section comes from Markovtsova *et al.* (2000a).

Modification of Markov chain Monte Carlo method

The event $\mathcal{U}$ corresponds to a single mutation arising on the branch indicated in Figure 8.1 and no other mutations on the rest of the coalescent tree. Let A denote the age of the UEP, and denote the mutation parameters by $M = (g, \kappa, \mu)$. In what follows we assume a prior distribution for M, and apply an MCMC method for generating observations from the conditional density $f(A, G \mid \mathcal{D}, \mathcal{E} \cap \mathcal{U})$ of A and $G = (\Lambda, \boldsymbol{T}, M)$ given $\mathcal{D}, \mathcal{E}$ and $\mathcal{U}$. To do this we express the required conditional density as a product of simpler terms and describe how each can be calculated.

First we note that

$$f(A, G \mid \mathcal{D}, \mathcal{U} \cap \mathcal{E}) = f(A \mid G, \mathcal{D}, \mathcal{U} \cap \mathcal{E}) f(G \mid \mathcal{D}, \mathcal{U} \cap \mathcal{E}). \tag{9.9.1}$$

The first term on the right of (9.9.1) can be evaluated by considering Figure 8.1 once more. Given that a single mutation occurs on the indicated branch, the Poisson nature of the mutation process for the UEP means that the location of the mutation is uniformly distributed over that branch. Thus we can simulate observations from the conditional distribution of A by simulating from the second term on the right of (9.9.1), reading off the length of the branch on which the UEP mutation occurs, and adding a uniformly distributed fraction of that length to the height of the subtree containing all the chromosomes carrying the UEP. Our task is therefore reduced to simulating from the second term on the right of (9.9.1).

Let $g_1(\Lambda \mid \mathcal{E})$ denote the conditional distribution of the coalescent tree Λ given $\mathcal{E}$, $g_2(\boldsymbol{T})$ the density of the coalescence times $\boldsymbol{T}$, and $g_3(M)$ the prior for the mutation rates $M = (g, \kappa, \mu)$. We can then write

$$f(G \mid \mathcal{D}, \mathcal{U} \cap \mathcal{E}) = \mathbb{P}(\mathcal{D}, \mathcal{U} \mid G, \mathcal{E}) g_1(\Lambda \mid \mathcal{E}) g_2(\boldsymbol{T}) g_3(M) / \mathbb{P}(\mathcal{D}, \mathcal{U} \mid \mathcal{E}). \quad (9.9.2)$$

The term $\mathbb{P}(\mathcal{D}, \mathcal{U} \mid G, \mathcal{E})$ is the product of two terms,

$$\mathbb{P}(\mathcal{D}, \mathcal{U} \mid G, \mathcal{E}) = \mathbb{P}(\mathcal{D} \mid G, \mathcal{E}) \mathbb{P}(\mathcal{U} \mid G, \mathcal{E}).$$

The first of these, the likelihood of D, can be computed using the peeling algorithm and the mutation model described above, while the second is

$$\frac{\mu S}{2} e^{-\mu S/2} \times e^{-\mu(L_n - S)/2} = \frac{\mu S}{2} e^{-\mu L_n/2}, \quad (9.9.3)$$

where S is the length of the branch on which the single UEP mutation must occur, and $L_n = \sum_{i=2}^{n} i T_i$ is the total length of the tree. The normalizing constant $\mathbb{P}(\mathcal{D}, \mathcal{U} \cap \mathcal{E})$ is unknown, and hard to compute. As a consequence, we use a version of the Metropolis-Hastings algorithm to simulate from the required conditional distribution.

Proposal kernel

We make a minor modification to Algorithm 9.2 in order to ensure that new trees are also consistent with the event $\mathcal{E}$. If, when we pick a level, we find we are in case A, and exactly two of the lines carry the UEP, then we cannot change the order in which the two coalescences occur, since such a change would produce a new tree topology which is inconsistent with $\mathcal{E}$. In such a situation we leave the topology unchanged.

Having constructed a new topology, which may be the same as the existing topology, we generate a new set of times in the same way as it was described in Section 9.5. We found that a kernel which proposes new values of T_l' and T_{l-1}' having the pre-data coalescent distribution worked well.

Finally, we update $M = (g, \kappa, \mu)$, where g and κ are the rate parameters for the sequence model and μ is the rate parameter for the UEP. The parameters g and κ were updated every tenth iteration, and μ was updated on each iteration for which g was not updated. These were updated using truncated Normals, whose variances require some tuning.

The Hastings ratio

Writing $G = (\Lambda, \boldsymbol{T}, M)$, the kernel Q can be expressed as the product of three terms:

$$Q(G' \to G) = Q_1(\Lambda' \to \Lambda)\, Q_2(\boldsymbol{T}' \to \boldsymbol{T} \mid \Lambda' \to \Lambda)\, Q_3(M' \to M).$$

Using (9.9.1), (9.9.2) and (9.9.3), the Hastings ratio (the probability with which we accept the new state) can be written in the form

$$h = \min\left\{1, \quad \frac{\mathbb{P}(\mathcal{D} \mid G', \mathcal{E})}{\mathbb{P}(\mathcal{D} \mid G, \mathcal{E})} \frac{\mathbb{P}(U \mid G', \mathcal{E})}{\mathbb{P}(U \mid G, \mathcal{E})} \frac{g_1(\Lambda' \mid \mathcal{E})}{g_1(\Lambda \mid \mathcal{E})} \frac{g_2(\boldsymbol{T}')}{g_2(\boldsymbol{T})} \frac{g_3(M')}{g_3(M)} \right.$$
$$\left. \times \frac{Q_1(\Lambda' \to \Lambda)}{Q_1(\Lambda \to \Lambda')} \frac{Q_2(\boldsymbol{T}' \to \boldsymbol{T} \mid \Lambda' \to \Lambda)}{Q_2(\boldsymbol{T} \to \boldsymbol{T}' \mid \Lambda \to \Lambda')} \frac{Q_3(M' \to M)}{Q_3(M \to M')} \right\},$$

the unknown term $\mathbb{P}(\mathcal{D}, \mathcal{U} \cap \mathcal{E})$ cancelling. For our choice of transition kernel Q, it can be shown that $g_1(\Lambda' \mid \mathcal{E}) = g_1(\Lambda \mid \mathcal{E})$. We also have $Q_1(\Lambda \to \Lambda') = Q_1(\Lambda' \to \Lambda)$, and we note that Q changes only two of the times associated with T or T'. Hence h reduces to

$$h = \min\left\{1, \frac{\mathbb{P}(\mathcal{D} \mid G', \mathcal{E})}{\mathbb{P}(\mathcal{D} \mid G, \mathcal{E})} \frac{\mathbb{P}(U \mid G', \mathcal{E})}{\mathbb{P}(\mathcal{U} \mid G, \mathcal{E})} \frac{g_2(\boldsymbol{T}')g_3(M')}{g_2(\boldsymbol{T})g_3(M)} \right.$$
$$\left. \times \frac{f_l(t_l)f_{l-1}(t_{l-1})}{f_l(t'_l)f_{l-1}(t'_{l-1})} \frac{Q_3(M' \to M)}{Q_3(M \to M')} \right\}, \tag{9.9.4}$$

where $f_l(\cdot)$ and $f_{l-1}(\cdot)$ are the densities of the time updating mechanism given that changes occur to the tree Λ at levels l and $l-1$.

In Section 8 we derived a number of theoretical results concerning the age of a UEP given its frequency in the sample in the limiting case $\mu \to 0$. In order to compare these results with those obtained by including the sequence information, we modified our algorithm to allow $\mu = 0$. Assuming κ is known, the mutation parameter M is now one-dimensional: $M = (g)$. The other change occurs to the conditional probability in (9.9.3), since now $\mathbb{P}(\mathcal{U} \mid G, \mathcal{E}) \propto S$, the length of the branch on which the UEP mutation must occur. This change appears in the Hastings ratio (9.9.4), where

$$\frac{\mathbb{P}(\mathcal{U} \mid G', \mathcal{E})}{\mathbb{P}(\mathcal{U} \mid G, \mathcal{E})} = \frac{S'}{S}.$$

In order to check tree moves, we can again use the infinitely-many-sites model of mutation. We compare distributions of time to the most recent common ancestor of the group of individuals carrying a specific mutation, the length of the corresponding sub-tree and the time to the mutation generated by the rejection method described in Algorithm 8.2 for the $\mu = 0$ case, and the modified version of our general MCMC scheme.

9.10 A Yakima data set

To illustrate the method we find the conditional distribution of the age of the 9 basepair mitochondrial region V deletion in a sample of Yakima described by Shields *et al.* (1993) The sample comprise $n = 42$ individuals, of whom $b = 26$ have the deletion. The data $\mathcal{D}$ comprise 360 basepairs from hypervariable region I of the control region, sequenced for all 42 individuals. The observed base frequencies are $(\pi_A, \pi_G, \pi_C, \pi_T) = (0.328, 0.113, 0.342, 0.217)$. We note that all individuals having a given control region sequence had the

same deletion status, as might be expected if the deletion arose once quite recently.

For the analysis discussed here, the output typically appeared to be non-stationary for at least 200,000 iterations of the algorithm. We generally discarded the first 25 million iterations. After this, we sampled every 5,000th iteration. Our output is typically based on 5000 samples from our stationary process. The acceptance rate was generally around 70%.

Preliminary analysis of the sequence data (without regard to presence or absence of the deletion) was performed using the approach outlined in Section 9.5. For the present mutation model, we took uninformative priors (in the form of uniform densities having wide but finite support) for the mutation rates g and w and examined the posterior distribution of $\kappa = w/g$. The posterior median was 65.9, the distribution having 25th percentile of 34.0 and 75th percentile of 160.2. The data are certainly consistent with the value of $\kappa = 100$ we used in the Nuh Chah Nulth example in Section 9.8. We therefore treat $\kappa = 100$ as fixed in the subsequent analyses; from (9.1.7) we find that $\theta = 88.17g$.

We repeated the analysis with an uninformative prior, uniform on (0, 0.1), for the single parameter g. This resulted in the posterior density for θ given in Figure 9.15. Summary statistics are shown in Table 12. Our approach also provides a way to find the maximum likelihood estimator of θ, since with a flat prior the posterior is proportional to the likelihood. From a kernel density estimate we obtained an MLE of $\hat{\theta} = 0.039$ with an estimated standard error of 0.010. This is consistent with the estimate of θ we found for the Nuu Chah Nulth data. Since the base frequencies in both data sets are similar and the mutation rates are likely to be the same, we conclude that the effective sizes of the two populations are also approximately equal. The effective population size of the Nuu Chah Nulth was estimated from anthropological data by Ward *et al.* (1991) to be about $N = 600$, a number we take for the Yakima as well.

Under the pre-data coalescent distribution, the mean time to the MRCA of a sample of $n = 42$ is $2(1 - 1/42) = 1.95$. With an effective size of $N = 600$ and a 20 year generation time, this is about 23,500 years. The posterior density of the time to the MRCA given the control region data D is shown in Figure 9.16. The posterior mean is 0.72, or about 8,600 years. Summary statistics are given in Table 13. The posterior distribution of the total tree length $L_{42} = \sum_{j=2}^{42} jT_j$ has mean 5.68.

We turn now to the deletion data. We ran our MCMC algorithm using a uniform (0, 10) prior for μ, and a uniform (0, 0.1) prior for g. The posterior density of θ is shown in Figure 9.15. Summary statistics are presented in Table 12. The distribution is qualitatively the same as that obtained by ignoring the deletion data. The posterior distribution of the deletion parameter μ has mean is 0.75 and median 0.61; the 25th percentile is 0.34 and the 75th percentile is 0.99. The posterior density of the time to the MRCA of the group carrying the deletion is shown in Figure 9.17. The summary statistics are found in Table 14.

Fig. 9.15. Posterior density of mutation rate θ

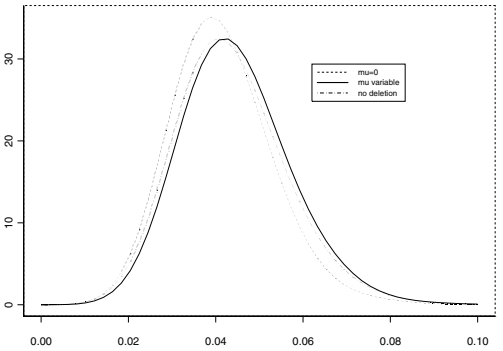


Fig. 9.16. Posterior density of TMRCA

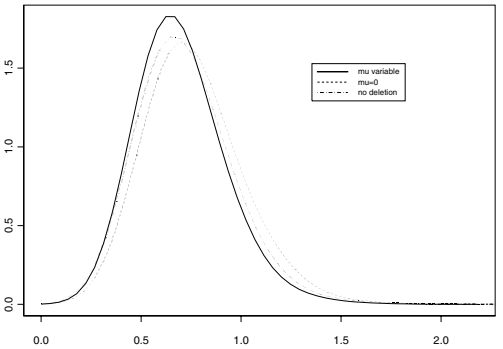


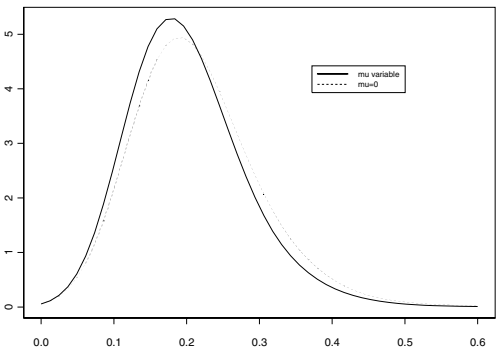
Table 12. Summary statistics for θ

θ	no deletion	μ variable	$\mu = 0$
mean	0.044	0.045	0.041
median	0.042	0.043	0.040
25th percentile	0.036	0.037	0.034
75th percentile	0.050	0.051	0.047

Table 13. Summary statistics for time to MRCA of the sample.

Time to MRCA	no deletion	μ variable	$\mu = 0$
mean	0.72 (8,600 yrs)	0.70 (8,400 yrs)	0.76 (9,200 yrs)
median	0.69 (8,300 yrs)	0.67 (8,000 yrs)	0.73 (8,800 yrs)
25th percentile	0.57 (6,800 yrs)	0.56 (6,700 yrs)	0.61 (7,300 yrs)
75th percentile	0.84 (10,100 yrs)	0.81 (9,700 yrs)	0.88 (10,600 yrs)

Fig. 9.17. Posterior density of TMRCA of deletion



The deletion arises uniformly on the branch indicated in Figure 8.1, so that the age of the mutation is the time to the MRCA of the deletion group plus a uniform fraction of the mutation branch length. The posterior distribution of the age is given in Figure 9.18, and summary statistics in Table 15.

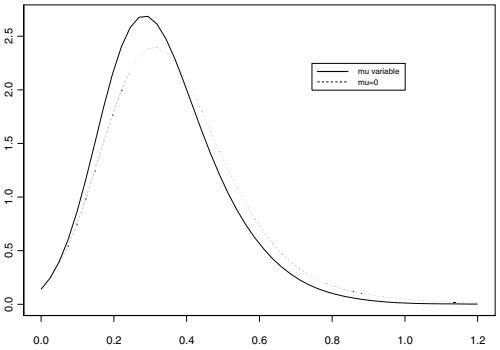
We also looked at the time to the MRCA of the entire sample when the deletion status of each sequence is included. The posterior density of this time is shown in Figure 9.16, with summary statistics given in Table 13. For these data the inclusion of deletion status has little effect on the posterior distribution.

The output from the MCMC runs can be used to assess whether the UEP assumption is reasonable. We first generated 5000 observations of the tree length L_{42} conditional on the data $\mathcal{D}$; as noted above, the sample mean is

Table 14. Summary statistics for the time to MRCA of the group carrying the deletion.

Time to MRCA	μ variable	$\mu = 0$
mean	0.20 (2400 yrs)	0.21 (2600 yrs)
median	0.19 (2300 yrs)	0.20 (2400 yrs)
25th percentile	0.15 (1800 yrs)	0.16 (1900 yrs)
75th percentile	0.24 (2900 yrs)	0.25 (3100 yrs)

Fig. 9.18. Posterior density of age of deletion



5.68. The modal posterior value of μ is 0.30, a value that we treat as a point estimate of μ. The expected number of deletions arising on the coalescent tree is then $0.30\,\mathbb{E}(L_{42}|\mathcal{D})/2$, which we estimate from the posterior mean tree length as $0.30 \times 5.68/2 = 0.85$. We can also use this value of μ and the simulated values of L_{42} to estimate the probability that exactly one mutation would occur on such a tree; we obtained an estimate of 0.36. Similarly, we estimated the probability of at least one mutation occurring as 0.57, so that the conditional probability that the mutation occurred once, given it occurred at least once, is estimated to be 0.63. Thus it is not unreasonable to assume that the deletion arose just once.

When $\mu = 0$, the posterior density of θ is shown in Figure 9.15, with summary statistics given in Table 12; there is little difference from the case where

Table 15. Summary statistics for age of the deletion.

Age of deletion	μ variable	$\mu = 0$
mean	0.34 (4100 yrs)	0.36 (4400 yrs)
median	0.31 (3700 yrs)	0.33 (4000 yrs)
25th percentile	0.23 (2800 yrs)	0.25 (3000 yrs)
75th percentile	0.41 (5000 yrs)	0.44 (5300 yrs)

μ is allowed to vary. The posterior density of the time to the MRCA is given in Figure 9.16, with summary statistics in Table 13. The mean time of 0.76 (or about 9,100 years) stands in marked contrast to the value of 2.68 (about 32,200 years) obtained from Griffiths and Marjoram (1996). The summary statistics for the posterior distribution of the time to the MRCA of the group carrying the deletion are given in Table 14. The results are qualitatively the same as the case of variable μ. The posterior density of the age of the deletion appears in Figure 9.18, with summary statistics shown in Table 15. The posterior mean is 0.36 (or about 4,400 years), compared to the value of 1.54 (or about 18,500 years) obtained from equation (8.3.4) when the sequence data are ignored. As expected, the mean age is higher than it is when μ is non-zero.

10 Recombination

In this section we study the generalization of the coalescent to the case of recombination. The basic groundwork of the subject comes from the seminal paper of Hudson (1983) and the ancestral recombination graph described by Griffiths (1991). We study the two locus model first, and then generalize to a model with arbitrary recombination rates. Later in the section we discuss methods for estimating the recombination rate, the behavior of measures of linkage disequilibrium, and uses of the coalescent for fine-scale mapping of disease genes.

10.1 The two locus model

Consider two linked loci, A and B, in a population of fixed size N chromosomes; neutrality, random mating and constant population size are assumed as before. For convenience, suppose the population reproduces according to a Wright-Fisher model with recombination: independently across offspring, in the next generation

(i) with probability $1 - r$ the individual chooses a chromosome from the previous generation and inherits the genes at the A and B loci.
(ii) with probability r the individual chooses 2 chromosomes from the previous generation and inherits the gene at the A locus from one and the gene at the B locus from the other.

In this model recombination is possible only between the two loci. If we focus on either of the two loci alone, we are watching a Wright-Fisher process evolve. It follows that the genealogical tree of a sample from one of the loci is described by the coalescent. There is thus a genealogical tree for each of the two loci. The effect of recombination is to make these two trees correlated. If $r = 0$, the loci are completely linked and the trees at each locus are identical. Early results for this model were obtained by Strobeck and Morgan (1978) and Griffiths (1981).

We consider the case in which N is large and r is of order N^{-1}; this balances the effects of drift and recombination. We define the (scaled) recombination rate ρ by

$$\rho = \lim_{N\to\infty} 2Nr \tag{10.1.1}$$

The ancestral process

Just as in the earlier model, we can calculate the chance that if there are currently k ancestors of the sample then in the previous generation there are also k. To the order of approximation we need, this occurs only if there are no recombination events in the k ancestors as they choose their parents, and the k also chose distinct parents. This event has probability

$$(1-r)^k\left(1-\frac{1}{N}\right)\cdots\left(1-\frac{k-1}{N}\right),$$

which, in the light of (10.1.1) is just

$$1-\frac{k\rho}{2N}-\frac{k(k-1)}{2N}+O(N^{-2}).$$

In a similar way, we can compute the probability that the number of distinct parents chosen in the previous generation increases from k to $k+1$. To the order we need, this occurs if precisely one recombination event occurs and the other $k-1$ ancestors choose distinct parents. A straightforward calculation shows that this probability is

$$\frac{k\rho}{2N}+O(N^{-2}).$$

Finally we can compute the chance that the number of ancestors goes down by 1, from k to $k-1$. The same sort of calculation shows this is

$$\frac{k(k-1)}{2N}+O(N^{-2}).$$

All other possibilities have smaller order. Thus we conclude that the number $A_n^N(Nt)$ behaves in the limit as $N\to\infty$ like continuous time birth and death process in which the transition rates are

$$\begin{aligned} k &\to k+1 \qquad \text{at rate } k\rho/2\\ k &\to k-1 \quad \text{at rate } k(k-1)/2 \end{aligned}$$

starting from state n. Because of the quadratic death rate compared to the linear growth rate, it is clear that the process will visit the value 1 infinitely often. The first occurrence of 1 corresponds to an MRCA.

A number of properties of the ancestral process $A_n^\rho(\cdot)$ can be found simply. Let M_n denote the maximum number of ancestors of the sample before it reaches its MRCA, and let τ_n denote the time to this MRCA. Griffiths (1991) proved:

Lemma 10.1 *The expected TMRCA is given by*

$$\mathbb{E}\tau_n=\frac{2}{\rho}\int_0^1\left(\frac{1-v^{n-1}}{1-v}\right)(e^{\rho(1-v)}-1)dv, \tag{10.1.2}$$

and the distribution of M_n is given by

$$\mathbb{P}(M_n\le k)=\frac{\sum_{j=n-1}^{k-1}j!\rho^{-j}}{\sum_{j=0}^{k-1}j!\rho^{-j}},\ k\ge n. \tag{10.1.3}$$

Proof. The expected height follows from standard results for birth-and-death processes. Define

$$\rho_i = \frac{\mu_2 \cdots \mu_{i-1}}{\lambda_2 \cdots \lambda_i}, \ i \geq 2.$$

For the ancestral process, it can be checked that $\rho_i = 2\rho^{i-2}/i!$. The waiting time to reach 1 has mean given by

$$\mathbb{E}\tau_n = \sum_{r=1}^{n-1} \left(\prod_{k=2}^{r} \frac{\mu_k}{\lambda_k} \right) \sum_{j=r+1}^{\infty} \rho_j,$$

where empty products have value 1 by convention. In our setting, this reduces to

$$\begin{aligned} \mathbb{E}\tau_n &= 2 \sum_{m=2}^{n} \sum_{l \geq 0} \rho^l \frac{(m-2)!\Gamma(m+2)}{(l+m)!(m+1)!} \\ &= \frac{2}{\rho} \int_0^1 \left(\frac{1-v^{n-1}}{1-v} \right) (e^{\rho(1-v)} - 1) dv. \end{aligned}$$

To find the distribution of M_n, define $p_n(k) = \mathbb{P}(M_n \leq k)$, with $p_1(k) = 1, k \geq 1$ and $p_n(k) = 0$ if $n > k$. By considering whether a coalescence or a recombination occurs first in the ancestry of the sample, we see that

$$p_n(k) = \frac{n-1}{\rho + n - 1} p_{n-1}(k) + \frac{\rho}{\rho + n - 1} p_{n+1}(k),$$

and it may readily be checked by induction that the solution is given by (10.1.3). □

As $\rho \downarrow 0$, we see from (10.1.2) that $\mathbb{E}\tau_n \to 2\int_0^1 (1 - v^{n-1}) dv = 2(1 - 1/n)$, as expected from our study of the coalescent. As $\rho \to \infty$, $\mathbb{E}\tau_n \to \infty$ also. When $n = 2$, we have

$$\mathbb{E}\tau_2 = 2\rho^{-2}(e^\rho - 1 - \rho),$$

and as $n \to \infty$,

$$\mathbb{E}\tau_\infty = \frac{2}{\rho} \int_0^1 v^{-1}(e^{\rho v} - 1) dv.$$

This last can be interpreted as the time taken for the whole population to be traced back to its common ancestor.

It follows from (10.1.3) that $M_n/n \to 1$ in probability as $n \to \infty$, showing that the width of the graph does not exceed n by very much.

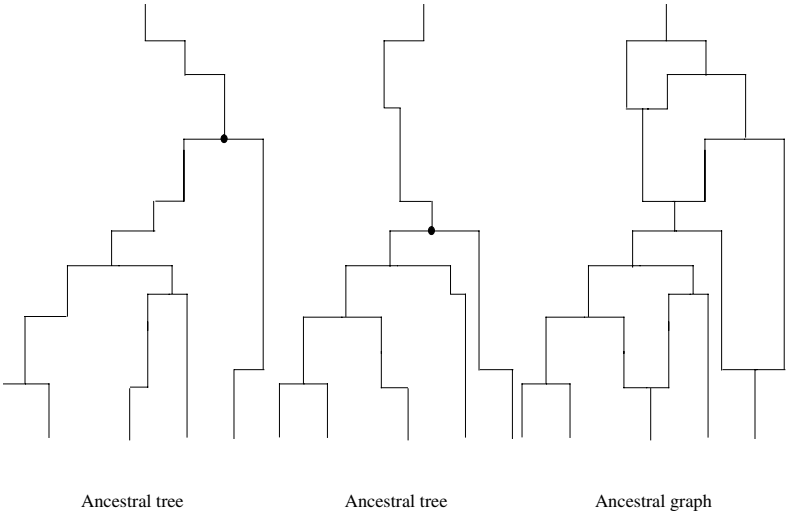
The ancestral recombination graph

We have seen that the ancestral process starts from $A_n^\rho(0) = n$, and has the property that if there are currently k ancestors then

(i) Any particular pair of branches coalesce at rate 1.
(ii) Any given branch splits into two at rate $\rho/2$.

The ancestral process $A_n^\rho(\cdot)$ is of limited use on its own; just as in the coalescent setting it is the way these individuals are related that matters. This leads to the idea of the *ancestral recombination graph* (or ARG). We construct such an ancestral recombination graph in such a way that when two edges are added at a recombination event, the genes represented by the left branch correspond to the A locus, and the right edges correspond to the B locus. In this way the ancestry of the A locus may be traced by following the left branch at each split, and the ancestry of the B locus by following the right branch. The ancestry of the A locus is a coalescent tree $\mathcal{T}_A$, and the ancestry of the B locus is a coalescent tree $\mathcal{T}_B$. These trees are dependent. Each tree has its own MRCA (which might be the same). An example of the ancestral graph, together with the two subtrees $\mathcal{T}_A$ and $\mathcal{T}_B$ is given in Figure 10.1. The MRCA at each locus marginally is denoted by a $\bullet$.

Fig. 10.1. Two locus ancestral recombination graph



Note that τ_n may now be interpreted as the height of the ARG, and M_n may be interpreted as its width. Of course, τ_n is at least as great as the time taken to find the MRCA at the A locus and at the B locus.

The structure of the ARG

In this section, we study the structure of the genealogical graph $\mathcal{G}$ in more detail. The graph includes the coalescent tree $\mathcal{T}_A$ of the A locus and the

coalescent tree $\mathcal{T}_B$ of the B locus. Denote the edge set of a graph by $\mathcal{E}(\cdot)$. It is useful to partition the edges $\mathcal{E}(\mathcal{G})$ into four disjoint sets:

$$\begin{aligned}\mathcal{A} &= \mathcal{E}(\mathcal{T}_A) \cap \mathcal{E}(\mathcal{T}_B)^c;\\ \mathcal{B} &= \mathcal{E}(\mathcal{T}_A)^c \cap \mathcal{E}(\mathcal{T}_B);\\ \mathcal{C} &= \mathcal{E}(\mathcal{T}_A) \cap \mathcal{E}(\mathcal{T}_B);\\ \mathcal{D} &= \mathcal{E}(\mathcal{G}) \cap \mathcal{E}(\mathcal{T}_A)^c \cap \mathcal{E}(\mathcal{T}_B)^c.\end{aligned}$$

Those edges in $\mathcal{A}$ represent ancestors who contribute to the genetic material of the sample at the A locus *only*, and similarly for $\mathcal{B}$ and the B locus. Edges in $\mathcal{C}$ correspond to ancestors that contribute genetic material at both loci, and those in $\mathcal{D}$ contribute no genetic material to the sample.

At any given time t, the ancestors of the sample (i.e. the edges $\mathcal{E}(\mathcal{G}_t)$ of the ancestral graph $\mathcal{G}_t$ of a cross section of $\mathcal{G}$ taken at time t) can be divided into these four types. Define

$$\begin{aligned}n_{\mathcal{A}}(t) &= |\mathcal{E}(\mathcal{G}_t) \cap \mathcal{A}|\\ n_{\mathcal{B}}(t) &= |\mathcal{E}(\mathcal{G}_t) \cap \mathcal{B}|\\ n_{\mathcal{C}}(t) &= |\mathcal{E}(\mathcal{G}_t) \cap \mathcal{C}|\\ n_{\mathcal{D}}(t) &= |\mathcal{E}(\mathcal{G}_t) \cap \mathcal{D}|,\end{aligned}$$

where $|\cdot|$ denotes the number of elements in a set. Clearly

$$n_{\mathcal{A}}(t) + n_{\mathcal{B}}(t) + n_{\mathcal{C}}(t) + n_{\mathcal{D}}(t) = |\mathcal{E}(\mathcal{G}_t)| \equiv A_n^{\rho}(t),$$

where $A_n^{\rho}(t)$ is the ancestral process of the ARG. Furthermore,

$$n_{\mathcal{A}}(t) + n_{\mathcal{C}}(t) = |\mathcal{E}(\mathcal{T}_{\mathcal{A}}(t))| \equiv A_n(t), \tag{10.1.4}$$

and

$$n_{\mathcal{B}}(t) + n_{\mathcal{C}}(t) = |\mathcal{E}(\mathcal{T}_{\mathcal{B}}(t))| \equiv B_n(t), \tag{10.1.5}$$

where $A_n(\cdot)$ and $B_n(\cdot)$ are the marginal ancestral processes for the A and B loci respectively.

Of interest is the evolution of the process

$$\boldsymbol{m}(t) = (n_{\mathcal{A}}(t), n_{\mathcal{B}}(t), n_{\mathcal{C}}(t), n_{\mathcal{D}}(t)), \quad t \geq 0.$$

One way to think of the process $\boldsymbol{m}$ is to label edges as (1,0), (0,1), (1,1), or (0,0) according as the edge is in $\mathcal{A}, \mathcal{B}, \mathcal{C}$, or $\mathcal{D}$ respectively. When a coalescence occurs to two edges of type (α, β) and (γ, δ) the resultant ancestor is of type $(\max(\alpha, \gamma), \max(\beta, \delta))$, and if a recombination occurs to an edge of type (α, β), the two new edges are of type $(\alpha, 0)$ and $(0, \beta)$.

Ethier and Griffiths (1990a) show that the process is Markovian. If the current state is (a, b, c, d), the next state and its transition rate are given by

$$\begin{array}{lll}
(a+1, b+1, c-1, d) & & c\rho/2 \\
(a-1, b-1, c+1, d) & & ab \\
(a-1, b, c, d) & & ac + a(a-1)/2 \\
(a, b-1, c, d) & \text{at rate} & bc + b(b-1)/2 \\
(a, b, c-1, d) & & c(c-1)/2 \\
(a, b, c, d+1) & & (a+b+d)\rho/2 \\
(a, b, c, d-1) & & d(a+b+c) + d(d-1)/2.
\end{array}$$

To see this, consider first the transition $(a, b, c, d) \to (a+1, b+1, c-1, d)$: this occurs if a recombination event occurs on an edge of type (1,1). This results in loss of a (1,1) edge, and the addition of one (1,0) edge and one (0,1) edge. The rate of such changes is $c\rho/2$. Considering the change $(a, b, c, d) \to (a-1, b, c, d)$ for example, we see that this results from a coalescence of a (1,0) edge and a (1,1) edge, or the coalescence of two (1,0) edges. Both possibilities result in the net loss of a (1,0) edge. The first type of change occurs at rate ac and the second sort at rate $a(a-1)/2$. In a similar way the other transitions and their rates may be verified. The overall transition rate is the sum of these rates; if $a + b + c + d = n$, this rate is given by $d_n \equiv c\rho/2 + n(n-1)/2$.

There is a reduced version of the Markov chain $\boldsymbol{m}(\cdot)$ that records only the first three coordinates:

$$\boldsymbol{n}(t) = (n_{\mathcal{A}}(t), n_{\mathcal{B}}(t), n_{\mathcal{C}}(t)), \quad t \geq 0.$$

Examining the transition rates of $\boldsymbol{m}(\cdot)$ given above shows that $\boldsymbol{n}(\cdot)$ is also Markovian, and from a state of the form (a, b, c) its transitions are to

$$(a_1, b_1, c_1) = \begin{cases}
(a+1, b+1, c-1) & & r_1 = c\rho/2 \\
(a-1, b-1, c+1) & & r_2 = ab \\
(a-1, b, c) & \text{at rate} & r_3 = ac + a(a-1)/2 \\
(a, b-1, c) & & r_4 = bc + b(b-1)/2 \\
(a, b, c-1) & & r_5 = c(c-1)/2
\end{cases} \tag{10.1.6}$$

Note that recombination takes place only on the edges in $\mathcal{C}$. The rate of change from a state (a, b, c) with $n = a + b + c$ is given by

$$d_n \equiv \frac{c\rho}{2} + \frac{n(n-1)}{2}. \tag{10.1.7}$$

Since the values of both $n_{\mathcal{A}}(t) + n_{\mathcal{C}}(t)$ and $n_{\mathcal{B}}(t) + n_{\mathcal{C}}(t)$ cannot increase as t increases, and eventually both must have the value 1, we see that the reduced process has absorbing states at $(1, 0, 0)$, $(0, 1, 0)$ and $\{(1, 1, 0), (0, 0, 1)\}$. It starts from $\boldsymbol{n}(0) = (0, 0, n)$. We might also consider the case in which only some of the genes, say $c < n$, are typed at both loci, while a are typed only at the A locus and the remaining $b = n - a - c$ are typed only at the B locus. In this case, $\boldsymbol{n}(0) = (a, b, c)$.

10.2 The correlation between tree lengths

In this section, we derive a recursion satisfied by the covariance of the tree lengths L^A and L^B of the marginal trees $\mathcal{T}_A$ and $\mathcal{T}_B$ respectively. The development here follows that of Pluzhnikov (1997).

For an initial configuration $\boldsymbol{n}(0) = (a, b, c)$ define $F(a, b, c; \rho)$ to be the covariance between L^A and L^B. Thus $F(0, 0, n; \rho)$ is the covariance of the marginal tree lengths for a sample of size n typed at both loci. We watch the Markov chain $\boldsymbol{n}(\cdot)$ only at the points it changes state. The resulting jump chain is denoted by $\boldsymbol{N}(\cdot)$. Let Z be a random variable that gives the outcome of a one-step jump of the chain $\boldsymbol{N}(\cdot)$ starting from (a, b, c), and let $Z = z_1$ correspond to the move to $(a+1, b+1, c-1)$, $Z = z_2$ correspond to the move to $(a-1, b-1, c+1)$ and so on, in the order given in (10.1.6). The jump probabilities are

$$p_i = \mathbb{P}(Z = z_i) = r_i/d_n, \quad i = 1, \ldots, 5. \tag{10.2.1}$$

Pluzhnikov (1997) established the following representation, which follows immediately from the properties of the coalescent trees $\mathcal{T}_A$ and $\mathcal{T}_B$ and the ARG.

Lemma 10.2 *Conditional on $Z = (a_1, b_1, c_1)$, we have*

$$L^A = X_A + T_A \tag{10.2.2}$$

where

(i) $X_A \sim L_{a_1+c_1}$, where L_m denotes the length of an m-coalescent tree;
(ii) $T_A \sim n_1 T$, where T is exponential(d_n) and $n_1 = a + c$;
(iii) X_A and T_A are independent.

Furthermore, a similar representation holds for L^B given Z:

$$L^B = X_B + T_B \sim L_{b_1+c_1} + n_2 T, \tag{10.2.3}$$

where $n_2 = b + c$. In addition, X_B and T_A are independent, as are X_A and T_B.

This leads to the main result of this section, derived originally in somewhat different form by Kaplan and Hudson (1985).

Theorem 10.3 *For any $\rho \in [0, \infty)$, the covariance* $\mathrm{Cov}(L^A, L^B) := F(a, b, s; \rho)$ *satisfies the linear system*

$$\begin{aligned} d_n F(a, b, c; \rho) = {} & r_1 F(a+1, b+1, c-1; \rho) + r_2 F(a-1, b-1, c+1; \rho) \\ & + r_3 F(a-1, b, c; \rho) + r_4 F(a, b-1, c; \rho) \\ & + r_5 F(a, b, c-1; \rho) + R_n \end{aligned} \tag{10.2.4}$$

where $n = a + b + c, n_1 = a + c, n_2 = b + c, d_n = (n(n-1) + c\rho)/2$, *the* r_i *are given in (10.1.6), and* $R_n = 2c(c-1)/((n_1-1)(n_2-1))$. *The system (10.2.4) has a unique solution satisfying the boundary conditions*

$$F(a, b, c; \rho) = 0 \text{ whenever } n_1 < 2, \text{ or } n_2 < 2, \text{ or } a < 0, \text{ or } b < 0, \text{ or } c < 0. \tag{10.2.5}$$

Proof. The proof uses the formula for conditional covariances, namely

$$\mathrm{Cov}(X, Y) = \mathbb{E}(\mathrm{Cov}(X, Y \mid Z)) + \mathrm{Cov}(\mathbb{E}(X \mid Z), \mathbb{E}(Y \mid Z)),$$

with $X = L^A, Y = L^B$ and Z as defined above. Clearly,

$$\mathbb{E}(\mathrm{Cov}(X, Y \mid Z)) = \sum_{i=1}^{5} p_i \mathrm{Cov}(X, Y \mid Z = z_i),$$

where the p_i are defined in (10.2.1). Now

$$\begin{aligned} \mathrm{Cov}(X, Y \mid Z = z_1) &= \mathrm{Cov}(X_A + T_A, X_B + T_B) \\ &= \mathrm{Cov}(X_A, X_B) + \mathrm{Cov}(T_A, T_B) \\ &= F(a+1, b+1, c-1, ; \rho) + n_1 n_2 \mathrm{Var}(T) \\ &= F(a+1, b+1, c-1, ; \rho) + n_1 n_2 d_n^{-2} \end{aligned} \tag{10.2.6}$$

Using similar arguments gives

$$\begin{aligned} \mathbb{E}(\mathrm{Cov}(X, Y \mid Z)) = r_1 F(a+1, b+1, c-1; \rho) + r_2 F(a-1, b-1, c+1; \rho) \\ r_3 F(a-1, b, c; \rho) + r_4 F(a, b-1, c; \rho) \\ + r_5 F(a, b, c-1; \rho) + n_1 n_2 d_n^{-2}. \end{aligned} \tag{10.2.7}$$

Next, recall that

$$\mathrm{Cov}(\mathbb{E}(Y \mid Z), \mathbb{E}(Y \mid Z)) = \mathbb{E}[(\mathbb{E}(X \mid Z) - \mathbb{E}(X))(\mathbb{E}(Y \mid Z) - \mathbb{E}(Y))].$$

Using basic properties of the regular coalescent, we can derive the distributions of $f(Z) = \mathbb{E}(X \mid Z) - \mathbb{E}(X)$ and $g(Z) = \mathbb{E}(Y \mid Z) - \mathbb{E}(Y)$; these are given in Table 16. Hence we find that

$$\begin{aligned} \mathrm{Cov}(\mathbb{E}(X \mid Z), \mathbb{E}(Y \mid Z)) &= \mathbb{E}(f(Z)g(Z)) \\ &= \sum_{i=1}^{5} p_i f(z_i) g(z_i) \\ &= -\frac{n_1 n_2}{d_n^2} + \frac{2c(c-1)}{d_n (n_1 - 1)(n_2 - 1)} \end{aligned} \tag{10.2.8}$$

Adding (10.2.7) and (10.2.8) yields (10.2.4).

Table 16. The probability distribution of $f(Z)$ and $g(Z)$

Z	$f(Z)$	$g(Z)$	$\mathbb{P}(Z = z_i)$
$(a+1, b+1, c-1)$	n_1/d_n	n_2/d_n	p_1
$(a-1, b-1, c+1)$	n_1/d_n	n_2/d_n	p_2
$(a-1, b, c)$	$n_1/d_n - 2/(n_1-1)$	n_2/d_n	p_3
$(a, b-1, c)$	n_1/d_n	$n_2/d_n - 2/(n_2-1)$	p_4
$(a, b, c-1)$	$n_1/d_n - 2/(n_1-1)$	$n_2/d_n - 2/(n_2-1)$	p_5

The boundary conditions follow from the restriction that the ancestral process for each locus be considered no further back than its MRCA. □

Equations like (10.2.4) can be solved by observing that if the degree of $F(a, b, c)$ is defined as $a + b + 2c$, then the degree on the right is at most the degree on the left; knowing lower degree terms allows the higher degree terms to be found by solving a lower triangular system of equations. Ethier and Griffiths (1990) developed an efficient computational method for solving such systems. The solution is known explicitly in very few cases, among them Griffiths' (1981) result

$$F(0, 0, 2; \rho) = \frac{4(\rho + 18)}{\rho^2 + 13\rho + 18}. \tag{10.2.9}$$

Some other examples

The equation in (10.2.4) can be written in the form

$$F(a, b, c; \rho) = \mathcal{L}F + g(a, b, c; \rho) \tag{10.2.10}$$

where in (10.2.4) we had $g(a, b, c; \rho) = d_n^{-1} R_n$. The same type of equation arises in studying many properties of the ARG. We mention two of them, derived by Griffiths (1991).

Define the time $W_n = \max(T_A, T_B)$ by which the sample of size n has a common ancestor at both the A and B loci. This is the time taken to reach the states $\{(1, 1, 0), (0, 0, 1)\}$ starting from $(0, 0, n)$. Starting from a configuration of (a, b, c) with $a + b + c = n$, the expected waiting time $f(a, b, c; \rho)$ satisfies (10.2.10) with

$$g(a, b, c; \rho) = d_n^{-1},$$

and boundary conditions

$$f(1, 0, 0; \rho) = 0,\ f(0, 1, 0; \rho) = 0,\ f(1, 1, 0; \rho) = 0,\ f(0, 0, 1; \rho) = 0. \tag{10.2.11}$$

We are interested in $\mathbb{E}W_n = f(0, 0, n; \rho)$. When $n = 50$, representative times are $\mathbb{E}W_{50} = 1.96\,(\rho = 0),\ = 2.14\,(\rho = 0.5),\ = 2.36\,(\rho = 2.0),\ = 2.50\,(\rho = 10),\ = 2.52\,(\rho = \infty)$.

Hudson and Kaplan (1985) studied the number of recombination events R_n^0 that occur in the history of the sample up to time W_n to ancestors of the sample having material belonging to both marginal trees. Define $f^0(a,b,c\rho)$ to be the expected number of transitions of the form $(a',b',c') \to (a'+1, b'+1, c'-1)$ until reaching the state $\{(1,1,0),(0,0,1)\}$, starting from (a,b,c). By considering the type of the first transition, we see that f^0 satisfies an equation of the form (10.2.10), with

$$g(a,b,c;\rho) = \frac{c\rho}{n(n-1)+c\rho},$$

and boundary conditions (10.2.11). The quantity we want is $\mathbb{E}R_n^0 = f^0(0,0,n;\rho)$. When $n = 50$, representative values are $\mathbb{E}R_{50}^0 = 0.00\,(\rho = 0)$, $= 2.13\,(\rho = 0.5)$, $= 7.51\,(\rho = 2.0)$, $= 25.6\,(\rho = 10)$.

In contrast, the expected number of recombination events $\mathbb{E}R_n$ in the entire history back to the grand MRCA can be found from the random walk which makes transitions according to

$$\begin{aligned} m &\to m+1 \qquad \text{with probability } \rho/(\rho+m-1),\ m \geq 0 \\ m &\to m-1 \quad \text{with probability } (m-1)/(\rho+m-1),\ m \geq 1. \end{aligned}$$

R_n is the number of times the random walk makes a move of the form $m' \to m'+1$ before reaching value 1. Standard random walk theory shows that

$$\mathbb{E}R_n = \rho \int_0^1 \frac{1-(1-v)^{n-1}}{v} e^{\rho v} dv. \tag{10.2.12}$$

When $n = 50$, representative times are $\mathbb{E}R_{50} = 0.00\,(\rho = 0)$, $= 2.52\,(\rho = 0.5)$, $= 16.2\,(\rho = 2.0)$, $= 24,900\,(\rho = 10)$. A comparison with the values of $\mathbb{E}R_n^0$ shows that $\mathbb{E}R_n$ and $\mathbb{E}R_n^0$ may differ dramatically.

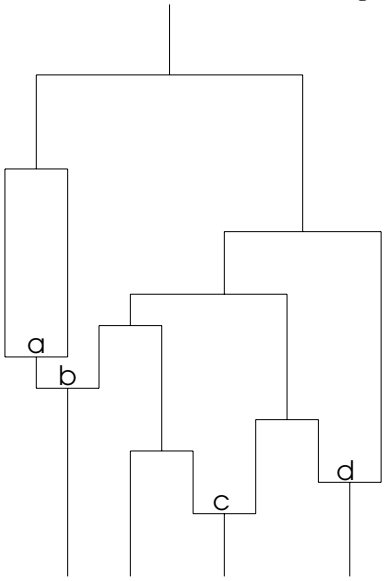
10.3 The continuous recombination model

We now consider a more general class of model in which each chromosome is represented by the unit interval $[0,1]$. This (and the figures in this section) comes from Griffiths and Marjoram (1997). If a recombination occurs, a position Z for the break point is chosen (independently from other break points) according to a given distribution, and the recombined chromosome is formed from the lengths $[0,Z]$ and $[Z,1]$ from the first and second parental chromosomes. Other details are as for the 2-locus model. There are several interesting potential choices for the break point distribution Z: Z is constant at 0.5, giving rise to the two-locus model studied earlier; Z is discrete, taking values $\frac{1}{m}, \ldots, \frac{m-1}{m}$, giving rise to a m-locus model; and Z has a continuous distribution on $[0,1]$, where breaks are possible at any point in $[0,1]$; a particular choice might be the uniform distribution on $[0,1]$.

As for the 2-locus model we are lead to the concept of ancestral graphs, but now the position at which a recombination occurs is also relevent. Figure 10.2

illustrates an ancestral graph for a sample of $n = 4$ individuals. Positions $Z_1, Z_2, \ldots$ where recombination breaks occur are labeled on the graph. The process $A_n^\rho(t)$ which records the number of ancestors of a sample of size n has identical transition rates as the corresponding process for the 2-locus model.

Fig. 10.2. Ancestral recombination graph.

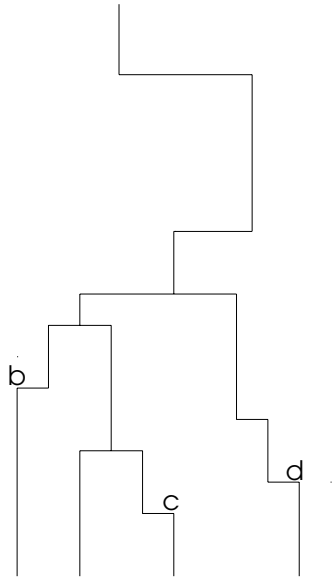


Whereas in the 2-locus model there were two ancestral trees corresponding to the ancestral graph, one for each locus, we now find that each point $x \in [0, 1]$ has an ancestral tree $\mathcal{T}(x)$ associated with it, and marginally each of these trees is described by the coalescent. To obtain $\mathcal{T}(x)$ we trace from the leaves of the ARG upward toward the MRCA. If there is a recombination vertex with label z, we take the left path if $x \leq z$, or right path if $x > z$. The MRCA in $\mathcal{T}(x)$ may occur in the graph before the grand MRCA. Figure 10.2 shows an example of $\mathcal{T}(x)$ when $x > b$ and $x < c, d$.

Since there are a finite number of recombination events in the graph, there are only a finite number of trees in $\{\mathcal{T}(x); x \in [0, 1]\}$. There are potentially 2^R if R recombination events have occurred, but some trees may be identical, or may not exist, depending on the ordering of the recombination break points. Of course (just as before) different trees share edges in the graph, and so are not independently distributed.

Figure 10.4 shows all possible trees corresponding to the ancestral graph in Figure 10.2. Trees 1 and 9 are identical; the other trees are all distinct. If $b > a$ then all trees exist as marginal trees in the graph, otherwise if $b < a$

Fig. 10.3. Marginal tree $\mathcal{T}(x)$, when $x > b$ and $x < c, d$.

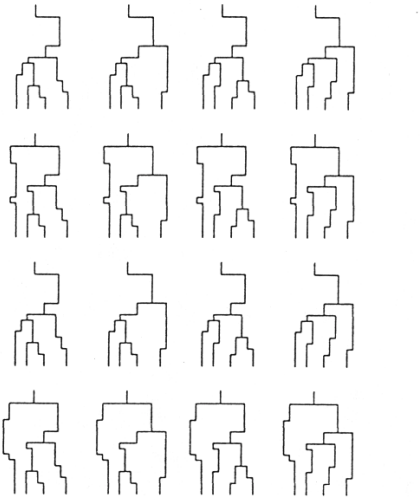


trees in Figure 10.4 with the right edge at vertex a do not exist as marginal trees.

Just as for the two-locus ARG, ancestor individuals may now only have part of their gametic material in common with the sample. It is also possible that some ancestors in the graph contain no material in common. A point x on an ancestor represented by an edge e in the graph has ancestral material in common with the sample if and only if e is included in $\mathcal{T}(x)$. Thus the subset of $[0, 1]$ over which that ancestor has ancestral material in common with the sample is $\mathcal{P}_e = \{x; \mathcal{T}(x) \ni e, x \in [0, 1]\}$. $\mathcal{P}_e$ is a union of a finite number of intervals, whose endpoints are a subset of the positions where recombination breaks have occurred. If e and f are two edges, and $e \vee f$ denotes a coalesced edge from e and f, then $\mathcal{P}_{e\vee f} = \mathcal{P}_e \bigcup \mathcal{P}_f$. If a recombination break occurs at z, to an edge e, then the left and right hand edges from e in the graph are $\mathcal{P}_e \bigcap [0, z]$ and $\mathcal{P}_e \bigcap [z, 1]$.

In the ancestral graph each ancestor can be labeled by which sample genes, and subsets of material it is ancestral to. The sample is represented as $\bigotimes_{i=1}^{n}(i, [0, 1])$ and at any given time the ancestors of the sample can be thought of as a partition of this set. An illustration of this is given in Figure 10.5, adapted from Nordborg and Tavaré (2002). The figure shows the regions of various ancestral segments that are ancestral to the members of the sample.

Fig. 10.4. All possible marginal trees for the graph in Figure 10.2.



10.4 Mutation in the ARG

Mutation is superimposed on the ARG just as it was in the single locus case: Mutations arise at rate $\theta/2$ independently on different branches of the tree, and their effects are modeled by the mutation operator Γ. In the coalescent model with recombination, it often makes no sense to consider mutations that arise on lineages that are lost in the history of the sample due to recombination. Instead, we consider just those mutations which occurred on lineages having meterial in common with the sample. In the m-locus model, there are now m marginal trees, denoted by $\mathcal{T}_1, \ldots, \mathcal{T}_m$. In we denote by $M_n^{(i)}$ the number of mutations occurring on the ith subtree back to its common ancestor, then the total number of mutations is

$$M_n = \sum_{i=1}^{m} M_n^{(i)}. \tag{10.4.1}$$

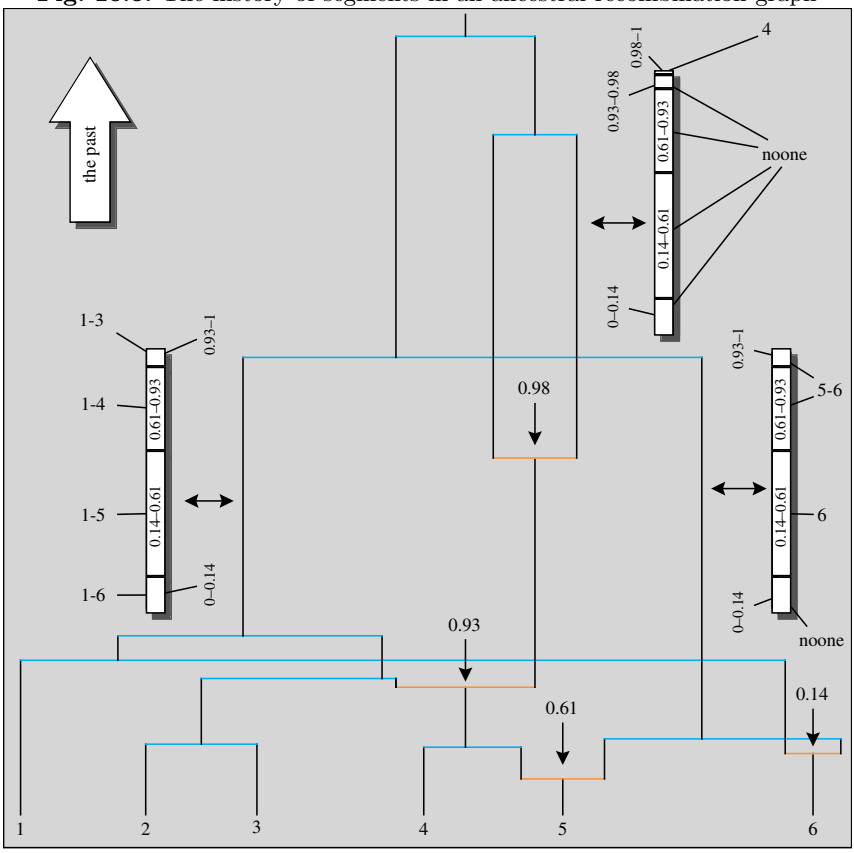
If the mutation rate at each locus is the same, then the overall mutation rate is $\Theta = m\theta$, so that

$$\mathbb{E}M_n = \sum_{i=1}^{m} \mathbb{E}M_n^{(i)} = \Theta \sum_{j=1}^{n-1} \frac{1}{j}. \tag{10.4.2}$$

Furthermore

$$\mathrm{Var}(M_n) = \sum_{i=1}^{m} \mathrm{Var}(M_n^{(i)} + 2 \sum_{i=1}^{m} \sum_{k=i+1}^{m} \mathrm{Cov}(M_n^{(i)}, M_n^{(k)}).$$

Fig. 10.5. The history of segments in an ancestral recombination graph



To evaluate the second term Σ_2, note that conditional on the two marginal subtrees, the mutation processes on those trees are independent. Denoting the tree length at the ith locus by $L_n^{(i)}$, this leads to Hudson's (1983) observation that

$$\mathrm{Cov}(M_n^{(i)}, M_n^{(k)}) = \frac{\theta^2}{4}\mathrm{Cov}(L_n^{(i)}, L_n^{(k)}).$$

In Theorem 10.3 we found the covariance $F_n(\rho) \equiv F(0, 0, n; \rho)$ of the tree lengths in a two locus model with recombination parameter ρ. We can use this to find the covariances in the m-locus model in which the recombination rate ρ between any two adjacent loci is assumed to be the same. The overall recombination rate is $R = (m-1)\rho$, and for $1 \le i < k \le m$, the covariance between $L_n^{(i)}$ and $L_n^{(j)}$ is given by $F_n((k-i)\rho)$. Hence

$$\Sigma_2 = \frac{\theta^2}{2} \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} F_n((k-i)\rho) = \frac{\theta^2}{2} \sum_{k=1}^{m-1} (m-k) F_n(k\rho).$$

Combining these results, we see that

$$\begin{aligned} \mathrm{Var}(M_n) &= m \left(\theta \sum_{j=1}^{n-1} \frac{1}{j} + \theta^2 \sum_{j=1}^{n-1} \frac{1}{j^2} \right) + \frac{\theta^2}{2} \sum_{k=1}^{m-1} (m-k) F_n(k\rho) \\ &= \Theta \sum_{j=1}^{n-1} \frac{1}{j} + \frac{\Theta}{m} \sum_{j=1}^{n-1} \frac{1}{j^2} + \frac{\Theta^2}{2m} \sum_{k=1}^{m-1} \left(1 - \frac{k}{m}\right) F_n\left(\frac{kR}{m-1}\right). \end{aligned}$$

Hudson considered the limiting case in which $m \to \infty$ while Θ and R are held fixed. This results in

$$\begin{aligned} \mathrm{Var}(M_n) &= \Theta \sum_{j=1}^{n-1} \frac{1}{j} + \frac{\Theta^2}{2} \int_0^1 (1-w) F_n(Rw) dw \qquad (10.4.3) \\ &= \Theta \sum_{j=1}^{n-1} \frac{1}{j} + \frac{1}{2} \frac{\Theta^2}{R^2} \int_0^R (R-w) F_n(w) dw. \end{aligned}$$

10.5 Simulating samples

We consider first the two-locus case. Suppose that there is an overall mutation rate of θ_A at the A locus, and θ_B at the B locus, and let $\theta = \theta_A + \theta_B$. We begin by describing the sequence of mutation, recombination, and coalescence events that occur in the history of the sample back to the MRCA.

Since mutations occur according to independent Poisson processes of rate $\theta/2$ along each lineage, we see that if there are currently m edges in the ancestral graph then the next event on the way back to the MRCA will be a mutation with probability $m\theta/(m(m-1) + m\theta + m\rho) = \theta/(m-1+\rho+\theta)$, a recombination with probability $\rho/(m-1+\theta+\rho)$, and a coalescence with probability $(m-1)/(m-1+\rho+\theta)$. With these events, we may associate a random walk $\{T_k, k \geq 0\}$ which makes transitions according to

$$\begin{array}{lll} m \to m+1 & \text{with probability} & \rho/(\theta+\rho+m-1), \\ m \to m & \text{with probability} & \theta/(\theta+\rho+m-1), \\ m \to m-1 & \text{with probability} & (m-1)/(\theta+\rho+m-1), \end{array}$$

for $m \geq 1$. To describe a sample of size n, the process starts from $T_0 = n$, and ends at the MRCA when $T_\cdot = 1$.

The effects of each mutation can be modeled in many different ways, for example allowing different combinations of infinitely-many-alleles, infinitely-many-sites, and finitely-many-sites at each locus. In the constant population size model, we can exploit the Markov chain $\{T_k, k \geq 0\}$ to provide an urn

model that can be used to simulate samples efficiently, particularly when the recombination rate ρ is not too large. First we have to generate the sequence of mutation, recombination, and coalescence events *back to the MRCA, starting at the sample*, and then superimpose the effects of each type of event starting at the MRCA and going down to the sample. Here is how this works.

Algorithm 10.1 To simulate from two-locus model.

(i) Simulate the random walk T_k starting from n until it reaches 1 at step τ. For $k = 1, \ldots, \tau$, write $U_k = T_{\tau-k+1} - T_{\tau-k}$.

(ii) Start by generating the type of the MRCA. For example, for a stationary sample choose the type of this individual according to the stationary distribution of the mutation process. If mutation is independent at each locus this is the product of the stationary distributions of each mutation process.

(iii) We now use the sequence $U_1, U_2, \ldots U_\tau$ (in that order) to generate the sample. For $k = 1, 2, \ldots, \tau$:

- If $U_k = -1$ then a recombination event has occurred. Choose two individuals at random without replacement from the current individuals, and recombine them. The first individual chosen contributes the A locus allele, the second the B locus allele.
- If $U_k = 0$, a mutation has occurred. Choose an individual at random and generate a mutation. With probability θ_A/θ the mutation occurs at the A locus, in which case a transition is made according to the mutation distribution $\Gamma^A(x, \cdot)$ if the type is currently x, and similarly for the B locus.
- If $U_k = 1$, then a coalescence has occurred. Choose an individual at random and duplicate its type.

(iv) After τ steps of the process, the sample has size n and the distribution of the sample is just what we wanted.

It can be seen that the efficiency of this algorithm depends on the expected value of τ. When either ρ or θ is large, $\mathbb{E}\tau$ can be very large, making the simulation quite slow.

This method extends directly to simulations of samples from the general ARG. Once the locations of the recombination events have been simulated according to Algorithm 10.1, we can choose recombination break points according to any prescribed distribution on [0,1]. Essentially any mutation mechanism can be modeled too. For example, for the infinitely-many-sites model we can suppose that mutations occur according to a continuous distribution on (0,1), and that the label of a mutation is just the position at which it occurs. In the case of variable population size this method does not work, and the ancestral recombination graph needs to be simulated first, and then mutations are superimposed from the MRCAs. Hudson (1991) is a useful reference.

10.6 Linkage disequilibrium and haplotype sharing

Because the genealogical trees at different linked positions in a segment are not independent of one another, neither will be the allelic states of these loci – there will be *linkage disequilibrium* (LD) between the loci. LD is usually quantified by using various measures of association between pairs of loci. Consider two such loci, each of which has two possible alleles, and denote the relative frequency of the A_iB_j haplotype by $p(A_i, B_j)$, and let $p(A_i), p(B_j)$ denote the relative frequency of each allele. Among the pairwise measures of LD are

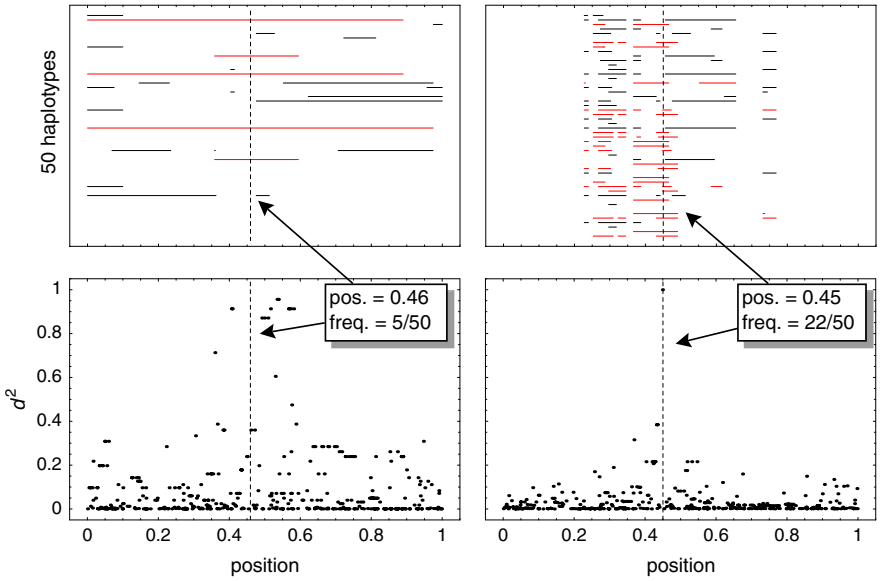
- D', the value of $D = p(A_1, B_1) - p(A_1)p(B_1)$, normalized to have values between -1 and 1 regardless of allele frequencies;
- r^2, the correlation in allelic state between the two loci as they occur in haplotypes;
- $d^2 = (p(B_1 \mid A_2) - p(B_1 \mid A_1))^2$, which measures the association between the alleles at (marker) locus B and the alleles at (disease) locus A.

These and other measures of LD are discussed further in Guo (1997), Hudson (2001) and Pritchard and Przeworski (2001).

Because of the history of recombination and mutation in a sample, pairwise LD is expected to be extremely variable. This is illustrated in Figure 10.6, adapted from Nordborg and Tavaré (2002). The horizontal axis, which represents chromosomal position, corresponds to roughly 100 kb. The plots illustrate the haplotype sharing and LD with respect to particular focal mutations. In the left column, a relatively low-frequency mutation (5/50=10%) was chosen as focus, and in the right column, a relatively high-frequency one (22/50=44%). The chromosomal position of these mutations are indicated by the vertical lines. The top row of plots shows the extent of haplotype sharing with respect to the MRCA of the focal mutation among the 50 haplotypes. The horizontal lines indicate segments that descend from the MRCA of the focal mutation. Light lines indicates that the current haplotype also carries the focal mutation, dark lines that it does not. Note that the light segments necessarily overlap the position of the focal mutation. For clarity, segments that do not descend from the MRCA of the focal mutation are not shown at all, and haplotypes that do not carry segments descended from the MRCA of the focal mutation are therefore invisible. The second row of plots shows the behavior of LD as measured by d^2 for different choices of markers. In each plot, the horizontal position of a dot represents the chromosomal position of the marker, and the vertical position the value of the measure (on a zero-to-one scale).

Because of interest in mapping disease susceptibility genes, the extent of LD across the human genome has been much debated. What is clear is that while there is a relationship between LD measures and distance, the inherent variability in LD makes this relationship hard to infer. In particular, it is difficult to compare studies that use different measures of pairwise LD as these measures can differ dramatically in their estimates of the range of LD.

Fig. 10.6. Decay of haplotype sharing



For reviews of these issues in relation to mapping, see for example Clayton (2000), Weiss and Clark (2002), Nordborg and Tavaré (2002) and Ardlie et al. (2002).

Estimating the recombination fraction

There is a sizable literature on estimation of the scaled recombination rate ρ, among them methods that use summary statistics of the data such as Hudson (1987), Hey and Wakeley (1997), and Wakeley (1997). Griffiths and Marjoram (1996) and Fearnhead and Donnelly (2001) exploit the importance sampling approach developed in Section 6 for the infinitely-many-sites model, while Nielsen (2000) and Kuhner et al. (2000) use MCMC methods, the latter specifically for DNA sequence data. Wall (2000) has performed an extensive comparison of these approaches. One conclusion is that (reliable) estimation of pairwise recombination fractions is extremely difficult. See Fearnhead and Donnelly (2002) for another approach, and Morris et al. (2002) and the references contained therein for approaches to mapping disease genes using the coalescent.

11 ABC: Approximate Bayesian Computation

Several of the previous sections have described methods for simulating observations from a posterior distribution. One key ingredient in these methods is the likelihood function; we have until now assumed this could be computed numerically, for example using the peeling algorithm described in Section 9.4. In this section we describe some methods that can be used when likelihoods are hard or impossible to compute.

In this section, data $\mathcal{D}$ are generated from a model $\mathcal{M}$ determined by parameters θ. We denote the prior for θ by $\pi(\theta)$. The posterior distribution of interest is $f(\theta \mid \mathcal{D})$ given by

$$f(\theta \mid \mathcal{D}) = \mathbb{P}(\mathcal{D} \mid \theta)\pi(\theta)/\mathbb{P}(\mathcal{D}),$$

where $\mathbb{P}(\mathcal{D}) = \int \mathbb{P}(\mathcal{D} \mid \theta)\pi(\theta)\, d\theta$ is the normalizing constant.

11.1 Rejection methods

We have already seen examples of the rejection method for discrete data:

Algorithm 11.1

1. Generate θ from $\pi(\cdot)$
2. Accept θ with probability $h = \mathbb{P}(\mathcal{D} \mid \theta)$, and return to 1.

It is easy to see that accepted observations have distribution $f(\theta \mid \mathcal{D})$, as shown for example in Ripley (1987). As we saw in Section 7.3, the computations can often be speeded up if there is constant c such that $\mathbb{P}(\mathcal{D} \mid \theta) \leq c$ for all θ. h can then then be replaced by h/c.

There are many variations on this theme. Of particular relevance here is the case in which the likelihood $\mathbb{P}(\mathcal{D} \mid \theta)$ cannot be computed explicitly. One approach is then the following:

Algorithm 11.2

1. Generate θ from $\pi(\cdot)$
2. Simulate $\mathcal{D}'$ from model $\mathcal{M}$ with parameter θ
3. Accept θ if $\mathcal{D}' = \mathcal{D}$, and return to 1.

The success of this approach depends on the fact that the underlying stochastic process $\mathcal{M}$ is easy to simulate for a given set of parameters. We note also that this approach can be useful when explicit computation of the likelihood is possible but time consuming.

The practicality of algorithms like these depends crucially on the size of $\mathbb{P}(\mathcal{D})$, because the probability of accepting an observation is proportional to

$\mathbb{P}(\mathcal{D})$. In cases where the acceptance rate is too small, one might resort to approximate methods such as the following:

Algorithm 11.3

1. Generate θ from $\pi(\cdot)$
2. Simulate $\mathcal{D}'$ from model $\mathcal{M}$ with parameter θ
3. Calculate a measure of distance $\rho(\mathcal{D}, \mathcal{D}')$ between $\mathcal{D}'$ and $\mathcal{D}$
4. Accept θ if $\rho \leq \epsilon$, and return to 1.

This approach requires selection of a suitable metric ρ as well as a choice of ϵ. As $\epsilon \to \infty$, it generates observations from the prior, and as $\epsilon \to 0$, it generates observations from the required density $f(\theta \mid \mathcal{D})$. The choice of ϵ reflects the interplay between computability and accuracy. For a given ρ and ϵ accepted observations are independent and identically distributed from $f(\theta \mid \rho(\mathcal{D}, \mathcal{D}') \leq \epsilon)$.

11.2 Inference in the fossil record

In this section, we give an application of Algorithm 11.3 to a problem concerning estimation of the time to the most recent common ancestor of primates. Our inference is based not on molecular data but on a sampling of the fossil record itself.

The problem

In Table 17 the number of primate species found as fossils in a series of stratigraphic intervals is given. Tavaré *et al.* (2002) developed a statistical method for estimating the temporal gap between the base of the stratigraphic interval in which the oldest fossil was found and the initial point of divergence of the species in the sample. The bias in the estimators and approximate confidence intervals for the parameters were found by using a parametric bootstrap approach. Estimates of the divergence time of primates (more accurately, the time of the haplorhine-strepsirrhine split) based on molecular sequence data give a time of about 90 million years. A literal interpretation of the fossil record suggests a divergence time of about 60 million years. One reason for the present studies is to reconcile these two estimates. A more detailed account of the problem is given in Soligo *et al.* (2002).

A model for speciation and sampling

We adopt the same framework as in Tavaré *et al.* (2002). We model speciation with a non-homogeneous Markov birth-and-death process. To model evolution from the last common ancestor of all living and fossil species included in the

Table 17. Data for the primate fossil record. References can be found in the supplemental material in Tavaré *et al.* (2002).

Epoch	k	T_k	Observed number of species (D_k)
Late Pleistocene	1	0.15	19
Middle Pleistocene	2	0.9	28
Early Pleistocene	3	1.8	22
Late Pliocene	4	3.6	47
Early Pliocene	5	5.3	11
Late Miocene	6	11.2	38
Middle Miocene	7	16.4	46
Early Miocene	8	23.8	36
Late Oligocene	9	28.5	4
Early Oligocene	10	33.7	20
Late Eocene	11	37.0	32
Middle Eocene	12	49.0	103
Early Eocene	13	54.8	68
Pre-Eocene	14		0

analysis, we start with two species at time 0. Species go extinct at rate λ, and so have exponential lifetimes with mean $1/\lambda$, time being measured in millions of years. A species that goes extinct at time u is replaced by an average of $m(u)$ new species. We denote by Z_t the number of species alive at time t. The expected number of species extant at time t is given by

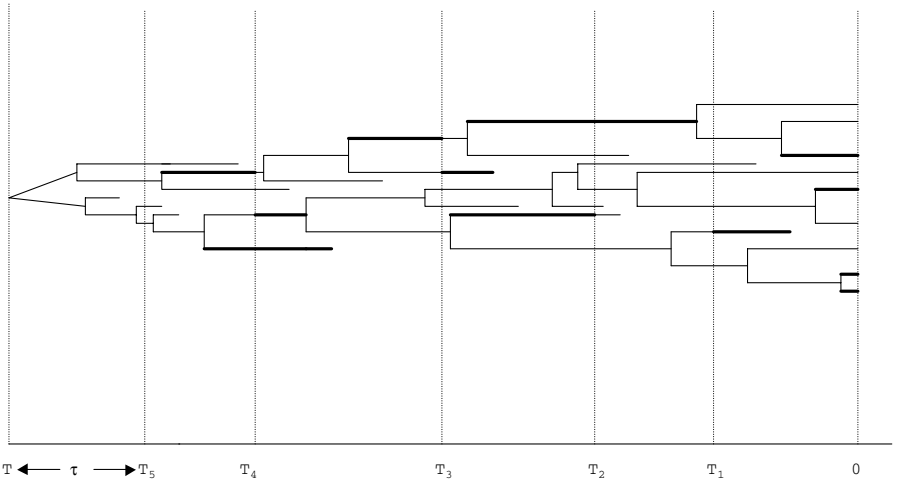
$$\mathbb{E}Z_t = 2\exp\left\{\lambda\int_0^t (m(u)-1)du\right\}; \tag{11.2.1}$$

cf. Harris (1963), Chapter 5. Furthermore, if $B(s,t]$ denotes the number of species born in the interval $(s,t]$, then

$$\mathbb{E}B[s,t) = \lambda\int_s^t m(u)\,\mathbb{E}Z_u\,du, \quad s<t. \tag{11.2.2}$$

We divide time into k stratigraphic intervals, following this sequence (see Table 17 and Figure 11.1). The base of the first (youngest) stratigraphic interval is at T_1 mya and the base of the k^{th} is at T_k million years ago (mya). The earliest known fossil is found in this interval. The founding species originate at time $T := T_k + \tau$ mya, and we define a $(k+1)^{st}$ stratigraphic interval that has its base at $T_{k+1} := T$ mya and ends T_k mya. Note that no fossils have

Fig. 11.1. An illustration of the stochastic model of fossil finds. Bases of 5 stratigraphic intervals at $T_1, , \ldots, T_5$ mya are shown along the x-axis. The temporal gap between the base of the final interval and the point at which the two founding species originate is denoted by τ. Thick lines indicate species found in the fossil record. Time 0 is the present day.



been found in this interval. We wish to approximate the posterior distribution of the time τ and other parameters of the model, using as data the number of different species found in the fossil record in the first, second, ..., k^{th} intervals. We model the number of species alive u mya by the value Z_{T-u} of the Markov branching process described earlier.

The number N_j of distinct species living in the jth stratigraphic interval having base T_j mya is the sum of those that were extant at the beginning of the interval, Z_{T-T_j}, plus those that originated in the interval, $B[T-T_j, T-T_{j-1})$. It follows from (11.2.1) and (11.2.2) that the expected number of distinct species that can be sampled in the jth stratigraphic interval is

$$\mathbb{E}N_j = \mathbb{E}Z_{T-T_{j-1}} + \lambda \int_{T-T_j}^{T-T_{j-1}} \mathbb{E}Z_u \, du, \quad j = 1, \ldots, k+1. \quad (11.2.3)$$

We assume that, conditional on the number of distinct species N_j that lived in the jth stratigraphic interval, the number of species D_j actually found in the fossil record in this interval is a binomial random variable with parameters N_j and α_j, $\quad j = 1, 2, \ldots, k$. Furthermore, the D_j are assumed to be conditionally independent given the N_j. The parameter α_j gives the probability of

sampling a species in the jth stratigraphic interval. A typical data set is given in Table 17.

A Bayesian approach

We write $\mathcal{D} = (D_1, \ldots, D_{k+1})$ for the counts observed in the $k+1$ stratigraphic intervals, and we write θ for the vector of parameters of the process, one of which is τ, the temporal gap. The likelihood can be written in the form

$$\mathbb{P}(\mathcal{D} \mid \theta) = \mathbb{E} \prod_{j=1}^{k+1} \binom{N_j}{D_j} \alpha_j^{D_j} (1-\alpha_j)^{N_j - D_j}, \tag{11.2.4}$$

where the expectation is over trajectories of the speciation process Z that run for time T with parameter θ, and such that both initial branches have offspring species surviving to time T. By convention the term under the expectation sign is 0 if any $D_j > N_j$.

While the acceptance probability is difficult to compute, the stochastic process itself can be simulated easily, and Algorithm 11.3 comes into play. One crucial aspect of this method is the choice of ρ in Algorithm 11.3. The counts $D_1, \ldots, D_{k+1}$ can be represented as the total number of fossils found,

$$D_+ = D_1 + \cdots + D_{k+1},$$

and a vector of proportions

$$(Q_1, \ldots, Q_{k+1}) := \left(\frac{D_1}{D_+}, \ldots, \frac{D_{k+1}}{D_+} \right).$$

We can therefore measure the distance between $\mathcal{D}$ and a simulated data set $\mathcal{D}'$ by

$$\rho(\mathcal{D}, \mathcal{D}') = \left| \frac{D'_+}{D_+} - 1 \right| + \frac{1}{2} \sum_{j=1}^{k+1} |Q_j - Q'_j|. \tag{11.2.5}$$

The first term measures the relative error in the total number of fossils found in a simulated data set and the actual number, while the second term is the total variation distance between the two vectors of proportions.

Results

Tavaré *et al.* (2002) modelled the mean diversification via the logistic function, for which

$$\mathbb{E}Z_t = 2/\{\gamma + (1-\gamma)e^{-\rho t}\}. \tag{11.2.6}$$

This form is quite flexible; for example, $\gamma = 0$ corresponds to exponential growth. They equated the expected number of species known at the present time with the observed number, and also specified the time at which the mean

diversification reached 90% of its current value. These two equations serve to determine the form of the speciation curve. They also assumed a mean species lifetime of 2.5 my (although their results were little changed by assuming a 2 my or 3 my lifetime). They modelled the sampling fractions α_j in the form

$$\alpha_j = \alpha p_j, \quad j = 1, 2, \ldots, k+1, \tag{11.2.7}$$

where the p_j are known proportions, and α is a scale parameter to be estimated from the data. The particular values of the p_j they used are given in Table 18. The average value is $\bar{p} = 0.73$.

Table 18. Sampling proportions p_j

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14
p_j	1.0	1.0	1.0	1.0	0.5	0.5	1.0	0.5	0.1	0.5	1.0	1.0	1.0	0.1

Using the data from Table 17, they estimated a temporal gap of 26.7 my with an approximate 95% confidence interval of 17.2 my to 34.8 my. As the oldest known fossil primate is 54.8 my old, this is equivalent to an age of 81.5 my for the last common ancestor of living primates. The average sampling fraction $\bar{\alpha}$, defined as

$$\bar{\alpha} = \alpha \bar{p} \tag{11.2.8}$$

was estimated to be 5.7% with an upper 95% confidence limit of 7.4%.

For comparison with the earlier approach, we treat both ρ and γ as fixed parameters, so that the parameter θ is given by $\theta = (\tau, \alpha)$. The prior distributions were chosen as

$$\tau \sim U(0, 100)$$
$$\alpha \sim U(0, 0.3)$$

the notation $U(a, b)$ denotes the uniform density on (a, b). In Tavaré *et al.* (2002), we used fixed values of $\rho = 0.2995, \gamma = 0.0085$. From 500 accepted observations with $\epsilon = 0.1$, we obtain the summaries in Figure 11.2 and Table 19. A median value of 27.6 my for the posterior value of the temporal gap τ is very close to that estimated in the previous analysis (Tavaré *et al.* (2002)) and is equivalent to an age of 82.4 my for the last common ancestor of living primates. The 2.5% and 97.5% points of the posterior of τ are estimated to be 15.4 my and 57.9 my, and the 95% point of the posterior for $\bar{\alpha}$ is 10%; these values are all broadly consistent with the previously published analysis. The posterior distribution of the number of present-day species serves as a

Fig. 11.2. Left panel: posterior for τ. Right panel: posterior for $\bar{\alpha}$.

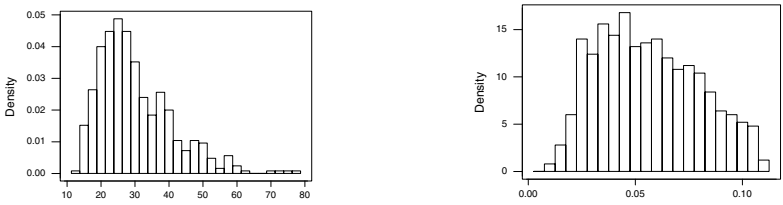


Table 19. Summary statistics for τ, $\bar{\alpha}$ and N_0 when ρ and γ are fixed.

	τ	$\bar{\alpha}(\%)$	N_0
25th percentile	22.4	3.7	180
median	27.6	5.4	253
mean	30.1	5.7	294
75th percentile	36.6	7.5	357

goodness-of-fit assessment. The observed number of extant primates, 235, is clearly a typical value under the posterior.

The analysis here can be compared to a full MCMC approach. The results are essentially indistinguishable; see Plagnol and Tavaré (2003) for further details. One advantage of approximate Bayesian approaches are their flexibility. A number of other scenarios, such as different species diversity curves and sampling schemes, can be examined quickly. For further details, see Will *et al.* (2003).

11.3 Using summary statistics

In Section 7 we found the posterior distribution conditional on a summary statistic rather than the full sequence data. The motivating idea behind this is that if the set of statistics $S = (S_1, \ldots, S_p)$ is sufficient for θ, in that $\mathbb{P}(\mathcal{D} \mid S, \theta)$ is independent of θ, then $f(\theta \mid \mathcal{D}) = f(\theta \mid S)$. The normalizing constant is now $\mathbb{P}(S)$ which is typically larger than $\mathbb{P}(\mathcal{D})$, resulting in more acceptances.

In practice it is be hard, if not impossible, to identity a suitable set of sufficient statistics, and we might then resort to a more heuristic approach that uses knowledge of the particular problem at hand to suggest summary statistics that capture information about θ. With these statistics in hand,

we have the following approximate Bayesian computation scheme for data $\mathcal{D}$ summarized by S:

Algorithm 11.4

1. Generate θ from $\pi(\cdot)$
2. Simulate $\mathcal{D}'$ from model $\mathcal{M}$ with parameter θ, and compute the corresponding statistics S'
3. Calculate the distance $\rho(S, S')$ between S and S'
4. Accept θ if $\rho \leq \epsilon$, and return to 1.

Examples of this algorithm approach appear frequently in the population genetics literature, including Fu and Li (1997), Weiss and von Haeseler (1998), Pritchard *et al.* (1999) and Wall (2000). Beaumont *et al.* (2002) describes a novel generalization of the rejection method in which all observations generated in steps 1 and 2 of Algorithm 11.4 are used in a local-linear regression framework to improve the simulation output. They also describe a number of other examples of this approach.

11.4 MCMC methods

There are several advantages to these rejection methods: they are usually easy to code, they generate independent observations (and so can use embarrassingly parallel computation), and they readily provide estimates of Bayes factors, which can be used for model comparison. On the other hand, for complex probability models sampling from the prior does not make good use of accepted observations, so these methods can be prohibitively slow. Here we describe an MCMC approach to problems in which the likelihood cannot be readily computed.

As we saw in Section 9, the Metropolis-Hastings Algorithm for generating observations from $f(\theta \mid \mathcal{D})$ uses output from a Markov chain. It can be described as follows:

Algorithm 11.5

1. If at θ, propose a move to θ' according to a transition kernel $q(\theta \to \theta')$
2. Calculate
$$h = \min\left(1, \frac{\mathbb{P}(\mathcal{D} \mid \theta')\pi(\theta')q(\theta' \to \theta)}{\mathbb{P}(\mathcal{D} \mid \theta)\pi(\theta)q(\theta \to \theta')}\right)$$
3. Move to θ' with probability h, else stay at θ; go to 1.

In Marjoram *et al.* (2003) we describe an MCMC approach that is the natural analog of Algorithm 11.4, in that no likelihoods are used (or estimated) in its implementation. It is based on the following steps:

Algorithm 11.6

1. If at θ propose a move to θ' according to a transition kernel $q(\theta \to \theta')$
2. Generate $\mathcal{D}'$ using model $\mathcal{M}$ with parameters θ'
3. If $\rho(S', S) \leq \epsilon$, go to 4, and otherwise stay at θ and return to 1,
4. Calculate
$$h = h(\theta, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')}\right)$$
5. Move to θ' with probability h, else stay at θ; go to 1.

The stationary distribution of the chain is indeed $f(\theta \mid \rho(S', S) \leq \epsilon)$. Applications of this approach to inference about mutation rates are given in Marjoram *et al.* (2003) and Plagnol and Tavaré (2003). The method usually has to be implemented by including part of the underlying coalescent tree and the mutation process as part of the MCMC update (making it part of θ, as it were).

The method seems to allow some flexibility in studying problems where existing methods don't work well in practice, such as analyzing complex models of mutation and analyzing restriction fragment length polymorphism data. There is a need for research on implementable methods for identifying approximately sufficient statistics, and for the development of more sophisticated MCMC methods that do not use likelihoods. Such approaches will be necessary when addressing problems involving high-dimensional parameters.

11.5 The genealogy of a branching process

Thus far the genealogical processes used in these notes have been evolving in continuous time. In this section, we describe informally a method for generating the genealogical history of a sample of individuals evolving according to a discrete-time branching process.

The conventional way to describe the evolution of a Galton-Watson process is as a series of population sizes $Z_0, Z_1, Z_2, \ldots$ at times 0,1,2, ... The number of individuals Z_{m+1} is a random sum:

$$Z_{m+1} = \sum_{j=1}^{Z_m} \xi_{mj},$$

where $\xi_{mj}, j \geq 1$ are identically distributed random variables having a distribution that may depend on m. A more detailed description of the process gives the number of families F_{mk} born into generation m that have k members, $k = 0, 1, 2, \ldots$. Given Z_{m-1}, the joint distribution of $F_{mj}, j \geq 0$ is multinomial with sample size Z_{m-1} and $q_{m-1,k}, k \geq 0$; here,

$$q_{m-1,k} = \mathbb{P}(\xi_{m-1,1} = k).$$

To simulate the genealogy of a random sample of size n from generation g of the process we proceed as follows; cf. Weiss and von Haeseler (1997). Starting from Z_0 individuals, generate the family size statistics $F_{1k}, k \geq 0$. These determine Z_1, after which the family sizes $F_{2k}, k \geq 0$ can be generated. Continuing in this way we finally generate the family sizes $F_{gk}, k \geq 0$. This done, a random subtree with n leaves can be generated backwards from generation g as follows. Randomly choose n individuals without replacement, recording which family they belong to (there being F_{gk} families of size k). Count the number A of families represented, each one corresponding to a distinct ancestor in generation $g-1$. Next, sample A individuals from generation $g-1$ and record which families they belong to (there now being $F_{g-1,k}$ families of size k), and so on. Iterating this scheme back through generations $g-1, g-2, \ldots, 1$ produces a genealogical tree having the required distribution.

Versions of this scheme have been used to study the polymerase chain reaction by Weiss and von Haeseler (1997), and to estimate the time to loss of mismatch repair in a colon tumor by Tsao *et al.* (2000) and Tavaré (2004). In both examples, the effects of a mutation process are superimposed on the genealogy, thereby generating sample data. Because the simulated genealogies are relatively quick to produce, they can be used for statistical inference such as implementations of Algorithm 11.4.

Finally we note that the simulation scheme can be used in much more general settings. For example, the distribution $q_{mj}, j \geq 0$ can depend on the history of the process in generations $0, 1, \ldots, m$; this covers cases of density dependent reproduction. This approach can also be applied to multitype branching processes.

12 Afterwords

This section concludes the lecture notes by giving some pointers to topics that were mentioned in the Saint Flour lectures, but have not been written up for the printed version.

12.1 The effects of selection

The previous sections have focussed on neutral genes, in which the effects of mutation could be superimposed on the underlying coalescent genealogy. When selection is acting at some loci, this separation is no longer possible and the analysis is rather more complicated.

Two basic approaches have emerged. In the first the evolution of the selected loci is modelled forward in time, and then the neutral loci are studied by coalescent methods (cf. Kaplan *et al.* (1988, 1989). In the second a genealogical process known as the *ancestral selection graph*, the analog of the neutral coalescent, is developed by Neuhauser and Krone (1997) and Krone and Neuhauser (1997). See Neuhauser (2001), Neuhauser and Tavaré (2002) and Nordborg (2001) for reviews. Methods for simulating selected genealogies are an important current area of research; see Slatkin (2001), Slade (2000, 2001) and Fearnhead (2001) for some examples. Such simulations can be used to explore the consequences of different selection mechanisms on the pattern of variation observed in data. Methods for detecting selection in sequence data are reviewed in Kreitman (2000). Methods for inference and estimation using coalescent methods are an active area of research. For an introduction to models with spatial structure, see Nordborg (2001) for example.

12.2 The combinatorics connection

Mathematical population genetics in the guise of the Ewens Sampling Formula (3.5.3) and Poisson approximation intersect in an area of probabilistic combinatorics. This leads directly to an extremely powerful and flexible method for studying the asymptotic behavior of decomposable combinatorial structures such as permutations, polynomials over a finite field, and random mappings. The joint distribution of counts of components of different sizes can be represented as the distribution of independent random variables conditional on a weighted sum; recall (3.5.4). Consequences of this representation are exploited in Arratia and Tavaré (1994). Connections with prime factorization are outlined in the expository article of Arratia *et al.* (1997). The book of Arratia, Barbour and Tavaré (2003) provides a detailed account of the theory, which places the Ewens Sampling Formula in much the same position as the Normal distribution in the central limit theorem: informally, many decomposable combinatorial models behave asymptotically like the Ewens Sampling Formula, and the closeness of the approximation can be measured in the total variation metric. A preprint of the book can be found at

http://www-hto.usc.edu/books/tavare/ABT/index.html

Pitman's lecture notes, *Combinatorial Stochastic Processes*, from the 2002 Saint Flour lectures contains related material. A draft may be obtained from

http://stat-www.berkeley.edu/users/pitman/bibliog.html

12.3 Bugs and features

Errors and typos will be reported at

http://www-hto.usc.edu/papers/abstracts/coalescent.html

I also intend to make a set of exercises available there.

I leave it to Søren Kierkegaard (1813-1855) to summarize why coalescents are interesting and why these notes end here:

> *Life can only be understood going backwards, but it must be lived going forwards.*

References

1. S. Anderson, A. T. Bankier, B. G. Barrell, M. H. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. Smith, R. Staden, and I. G. Young. Sequence and organization of the human mitochondrial genome. *Nature*, 290:457–465, 1981.
2. K. G. Ardlie, L. Kruglyak, and M. Seielstad. Patterns of linkage disequilibrium in the human genome. *Nature Rev. Genet.*, 3:299–309, 2002.
3. R. Arratia, A. D. Barbour, and S. Tavaré. Random combinatorial structures and prime factorizations. *Notices of the AMS*, 44:903–910, 1997.
4. R. Arratia, A. D. Barbour, and S. Tavaré. *Logarithmic Combinatorial Structures: a Probabilistic Approach*. European Mathematical Society Publishing House, 2003. In press.
5. R. Arratia and S. Tavaré. Independent process approximations for random combinatorial structures. *Adv. Math.*, 104:90–154, 1994.
6. A. D. Barbour, L. Holst, and S. Janson. *Poisson Approximation*. Oxford University Press, Oxford, 1992.
7. M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162:2025–2035, 2002.
8. N. G. Best, M. K. Cowles, and S. K. Vines. *CODA Manual version 0.30.* MRC Biostatistics Unit,, Cambridge, UK, 1995.
9. T. A. Brown. *Genomes*. John Wiley & Sons, New York, New York, 1999.
10. P. Buneman. The recovery of trees from measures of dissimilarity. In D. G. Kendall and P. Tautu, editors, *Mathematics in the Archaeological and Historical Sciences*, pages 387–395. Edinburgh University Press, 1971.
11. C. Cannings. The latent roots of certain Markov chains arising in genetics: A new approach. I. Haploid models. *Adv. Appl. Prob.*, 6:260–290, 1974.
12. D. Clayton. Linkage disequilibrium mapping of disease susceptibility genes in human populations. *International Statistical Review*, 68:23–43, 2000.
13. J. F. Crow and M. Kimura. *An Introduction to Population Genetics Theory*. Harper and Row, New York, 1970.

14. P. Donnelly and T. G. Kurtz. The asymptotic behavior of an urn model arising in poulation genetics. *Stochast. Process. Applic.*, 64:1–16, 1996.
15. P. Donnelly and S. Tavaré. Coalescents and genealogical structure under neutrality. *Ann. Rev. Genet.*, 29:401–421, 1995.
16. P. Donnelly, S. Tavaré, D. J. Balding, and R. C. Griffiths. Estimating the age of the common ancestor of men from the ZFY intron. *Science*, 257:1357–1359, 1996.
17. R. L. Dorit, H. Akashi, and W. Gilbert. Estimating the age of the common ancestor of men from the ZFY intron. *Science*, 272:1361–1362, 1996.
18. S. N. Ethier and T. G. Kurtz. *Markov Processes. Characterization and Convergence.* John Wiley & Sons, Inc., New York, 1986.
19. S.N. Ethier and R.C. Griffiths. The infinitely–many–sites model as a measure valued diffusion. *Ann. Probab.*, 15:515–545, 1987.
20. S.N. Ethier and R.C. Griffiths. On the two-locus sampling distribution. *J. Math. Biol.*, 29:131–159, 1990.
21. W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoret. Popn. Biol.*, 3:87–112, 1972.
22. W. J. Ewens. *Mathematical Population Genetics.* Springer-Verlag, Berlin, Heidelberg, New York, 1979.
23. W. J. Ewens. Population genetics theory - the past and the future. In S. Lessard, editor, *Mathematical and statistical developments of evolutionary theory.* Kluwer Academic Publishers, 1990.
24. W. J. Ewens and S. Tavaré. The Ewens Sampling Formula. In S. Kotz, C. B. Read, and D. L. Banks, editors, *Encyclopedia of Statistical Science,* Volume 2, pages 230–234. Wiley, New York, 1998.
25. P. Fearnhead. Perfect simulation from population genetic models with selection. *Theoret. Popul. Biol.*, 59:263–279, 2001.
26. P. Fearnhead and P. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159:1299–1318, 2001.
27. P. Fearnhead and P. Donnelly. Approximate likelihood methods for estimating local recombination rates. *J. Royal Statist. Soc. B*, 64:657–680, 2002.
28. J. Felsenstein. The rate of loss of multiple alleles in finite haploid populations. *Theoret. Popn. Biol.*, 2:391–403, 1971.
29. J. Felsenstein. Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from data on discreate charecters. *Syst. Zool.*, 22:240–249, 1973.
30. J. Felsenstein. Evolutionary trees from DNA sequence data: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.
31. J. Felsenstein, M. Kuhner, J. Yamato, and P. Beerli. Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In F. Seillier-Moiseiwitsch, editor, *Statistics in Molecular Biology and Genetics*, pages 163–185. Institute of Mathematical Statistics and American Mathematical Society, Hayward, California, 1999.
32. R. A. Fisher. On the dominance ratio. *Proc. Roy. Soc. Edin.*, 42:321–431, 1922.
33. G. E. Forsythe and R. A. Leibler. Matrix inversion by the Monte Carlo method. *Math. Comp.*, 26:127–129, 1950.
34. Y.-X. Fu. Estimating the age of the common ancestor of a DNA sample using the number of segregating sites. *Genetics*, 144:829–838, 1996.

35. Y.-X. Fu and W.-H. Li. Estimating the age of the common ancestor of men from the ZFY intron. *Science*, 257:1356–1357, 1996.
36. Y.-X. Fu and W.-H. Li. Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol.*, 14:195–199, 1997.
37. Y.-X. Fu and W.-H. Li. Coalescing into the 21st century: An overview and prospects of coalescent theory. *Theoret. Popul. Biol.*, 56:1–10, 1999.
38. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.
39. K. Gladstien. The characteristic values and vectors for a class of stochastic matrices arising in genetics. *SIAM J. Appl. Math.*, 34:630–642, 1978.
40. R. C. Griffiths. Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theor. Pop. Biol.*, 17:37–50, 1980.
41. R. C. Griffiths. Neutral two-locus multiple allele models with recombination. *Theoret. Popn. Biol.*, 19:169–186, 1981.
42. R. C. Griffiths. Genealogical-tree probabilities in the infinitely-many-site model. *J. Math. Biol.*, 27:667–680, 1989.
43. R. C. Griffiths. The two-locus ancestral graph. In I.V. Basawa and R.L. Taylor, editors, *Selected Proceedings of the Symposium on Applied Probability, Sheffield, 1989*, pages 100–117. Institute of Mathematical Statistics, Hayward, CA, 1991b.
44. R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *J. Comp. Biol.*, 3:479–502, 1996.
45. R. C. Griffiths and P. Marjoram. An ancestral recombination graph. In P. Donnelly and S. Tavaré, editors, *Progress in Population Genetics and Human Evolution*, pages 100–117. Springer Verlag, New York, 1997.
46. R. C. Griffiths and S. Tavaré. Ancestral inference in population genetics. *Statist. Sci.*, 9:307–319, 1994.
47. R. C. Griffiths and S. Tavaré. Simulating probability distributions in the coalescent. *Theor. Popn. Biol.*, 46:131–159, 1994.
48. R. C. Griffiths and S. Tavaré. Unrooted genealogical tree probabilities in the infinitely-many- sites model. *Math. Biosci.*, 127:77–98, 1995.
49. R. C. Griffiths and S. Tavaré. Monte Carlo inference methods in population genetics. *Mathl. Comput. Modelling*, 23:141–158, 1996.
50. R. C. Griffiths and S. Tavaré. Computational methods for the coalescent. In P. Donnelly and S. Tavaré, editors, *Progress in Population Genetics and Human Evolution*, pages 165–182. Springer Verlag, New York, 1997.
51. R. C. Griffiths and S. Tavaré. The age of a mutation in a general coalescent tree. *Stochastic Models*, 14:273–295, 1998.
52. R. C. Griffiths and S. Tavaré. The ages of mutations in gene trees. *Ann. Appl. Prob.*, 9:567–590, 1999.
53. R. C. Griffiths and S. Tavaré. The genealogy of a neutral mutation,. In P. J. Green, N. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*. Oxford University Press,, 2003. in press.
54. S.-W. Guo. Linkage disequilibrium measures for fine-scale mapping: A comparison. *Hum. Hered.*, 47:301–314, 1997.
55. D. Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21:19–28, 1991.
56. D. Gusfield. *Algorithms on Strings, Trees and Sequences. Computer Science and Computational Biology*. Cambridge University Press, Cambridge, 1997.

57. T. E. Harris. *The Theory of Branching Processes*. Springer Verlag, Berlin, 1963.
58. D. L. Hartl and E. W. Jones. *Genetics. Analysis of Genes and Genomes*. Jones and Bartlett, Sudbury, MA., Fifth edition, 2001.
59. W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.
60. J. Hey and J. Wakeley. A coalescent estimator of the population recombination fraction. *Genetics*, 145:833–846, 1997.
61. R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoret. Popn. Biol.*, 23:183–201, 1983.
62. R. R. Hudson. Estimating the recombination parameter of a finite population model without selection. *Genet. Res. Camb.*, 50:245–250, 1987.
63. R. R. Hudson. Gene genealogies and the coalescent process. In D. Futuyma and J. Antonovics, editors, *Oxford Surveys in Evolutionary Biology,* Volume 7, volume 7, pages 1–44. Oxford University Press, 1991.
64. R. R. Hudson. The how and why of generating gene genealogies. In N. Takahata and A. G. Clark, editors, *Mechanisms of Molecular Evolution*, pages 23–36. Sinauer, 1992.
65. R. R. Hudson. Linkage disequilibrium and recombination. In D. J. Balding, M. J. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, pages 309–324. John Wiley & Sons, Inc., Chichester, U.K., 2001.
66. R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147–164, 1985.
67. N. L. Kaplan, T. Darden, and R. R. Hudson. The coalescent process in models with selection. *Genetics*, 120:819–829, 1988.
68. N. L. Kaplan and R. R. Hudson. The use of sample genealogies for studying a selectively neutral m-loci model with recombination. *Theoret. Popn. Biol.*, 28:382–396, 1985.
69. N. L. Kaplan, R. R. Hudson, and C. H. Langley. The "hitch-hiking" effect revisited. *Genetics*, 123:887–899, 1989.
70. S. Karlin and J. McGregor. Addendum to a paper of W. Ewens. *Theoret. Popn. Biol.*, 3:113–116, 1972.
71. S. Karlin and H. M. Taylor. *A Course in Stochastic Processes*, volume 2. Wiley, New York, NY, 1980.
72. M. Kimura and T. Ohta. The age of a neutral mutant persisting in a finite population. *Genetics*, 75:199–212, 1973.
73. J. F. C. Kingman. *Mathematics of Genetic Diversity*, volume 34 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1980.
74. J. F. C. Kingman. The coalescent. *Stoch. Proc. Applns.*, 13:235–248, 1982.
75. J. F. C. Kingman. Exchangeability and the evolution of large populations. In G. Koch and F. Spizzichino, editors, *Exchangeability in Probability and Statistics*, pages 97–112. North-Holland Publishing Company, 1982.
76. J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Prob.*, 19A:27–43, 1982.
77. J. F. C. Kingman. Origins of the coalescent: 1974-1982. *Genetics*, 156:1461–1463, 2000.
78. M. Kreitman. Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics Hum. Genet.*, 1:539–559, 2000.

79. S. M. Krone and C. Neuhauser. Ancestral processes with selection. *Theoret. Poul. Biol.*, 51:210–237, 1997.
80. M. Kuhner, J. Yamato, and J. Felsenstein. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, 140:1421–1430, 1995.
81. M. Kuhner, J. Yamato, and J. Felsenstein. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, 149:429–434, 1998.
82. M. Kuhner, J. Yamato, and J. Felsenstein. Maximum likelihood estimation of recombination rates from population data. *Genetics*, 156:1393–1401, 2000.
83. B. Larget and D. L. Simon. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.*, 16:750–759, 1999.
84. W.-H. Li and Y.-X. Fu. Coalescent theory and its applications in population genetics. In M. E. Halloran and S. Geisser, editors, *Statistics in Genetics*, pages 45–79. Springer Verlag, 1999.
85. J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, New York, 2001.
86. R. Lundstrom. *Stochastic models and statistical methods for DNA sequence data.* PhD thesis, Mathematics Department, University of Utah, 1990.
87. R. S. Lundstrom, S. Tavaré, and R. H. Ward. Estimating mutation rates from molecular data using the coalescent. *Proc. Natl. Acad. Sci. USA*, 89:5961–5965, 1992.
88. R. S. Lundstrom, S. Tavaré, and R. H. Ward. Modeling the evolution of the human mitochondrial genome. *Math. Biosci.*, 112:319–335, 1992.
89. P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 000:000–000, 2003.
90. L. Markovtsova. *Markov chain Monte Carlo methods in population genetics.* PhD thesis, Mathematics Department, University of Southern California, 2000.
91. L. Markovtsova, P. Marjoram, and S. Tavaré. The effects of rate variation on ancestral inference in the coalescent. *Genetics*, 156:1427–1436, 2000b.
92. B. Mau, M. A. Newton, and B. Larget. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, 55:1–12, 1999.
93. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.
94. M. Möhle. Robustness results for the coalescent. *J. Appl. Prob.*, 35:438–447, 1998.
95. M. Möhle. Total variation distances and rates of convergence for ancestral coalescent processes in exchangeable population models. *Adv. Appl. Prob.*, 32:983–993, 2000.
96. M. Möhle. The coalescent in population models with time-inhomogeneous environment. *Stoch. Proc. and Applns.*, 97:199–227, 2002.
97. M. Möhle and S. Sagitov. A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.*, 29:1547–1562, 2001.
98. A. P. Morris, J. C. Whittaker, and D. J. Balding. Fine scale mapping of disease loci via shattered coalescent modelling of genealogies. *Amer. J. Hum. Genet.*, 70:686–707, 2002.
99. C. Neuhauser. Mathematical models in population genetics. In D. J. Balding, M. J. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, pages 153–177. John Wiley and Sons, Inc., New York, New York., 2001.

100. C. Neuhauser and S. M. Krone. The genealogy of samples in models with selection. *Genetics*, 145:519–534, 1997.
101. C. Neuhauser and S. Tavaré. The coalescent. In S. Brenner and J. Miller., editors, *Encyclopedia of Genetics,* Volume 1, pages 392–397. Academic Press, New York, 2001.
102. R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154:931–942, 2000.
103. M. Nordborg. Coalescent theory. In D. J. Balding, M. J. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, pages 179–208. John Wiley and Sons, Inc., New York, New York., 2001.
104. N. Nordborg and S. Tavaré. Linkage disequilibrium: what history has to tell us. *Trends in Genetics*, 18:83–90, 2002.
105. N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. Fodor, and D. R. Cox. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294:1719–1723, 2001.
106. J. W. Pitman. Coalescents with multiple collisions. *Ann. Probab.*, 27:1870–1902, 1999.
107. V. Plagnol and S. Tavaré. Approximate Bayesian computation and MCMC. In H. Niederreiter, editor, *Monte Carlo and Quasi-Monte Carlo Methods 2002.* Springer-Verlag., 2004. In press.
108. A. Pluzhnikov. *Statistical inference in population genetics.* PhD thesis, Statistics Department, University of Chicago, 1997.
109. J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: Models and data. *Amer. J. Hum. Genet.*, 69:1–14, 2001.
110. J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Mol. Biol. Evol.*, 16:1791–1798, 1999.
111. W. B. Provine. *The Origins of Theoretical Population Genetics.* University of Chicago Press, second edition, 2001.
112. B. D. Ripley. *Stochastic Simulation.* Wiley, New York, 1987.
113. S. Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Prob.*, 36:1116–1125, 1999.
114. I. W. Saunders, S. Tavaré, and G. A. Watterson. On the genealogy of nested subsamples from a haploid population. *Adv. Appl. Prob.*, 16:471–491, 1984.
115. S. Sawyer, D. Dykhuizen, and D. Hartl. Confidence interval for the number of selectively neutral amino acid polymorphisms. *Proc. Natl. Acad. Sci. USA*, 84:6225–6228, 1987.
116. J. Schweinsberg. Coalescents with simultaneous multiple colllisions. *Electron. J. Prob.*, 5:1–50, 2000.
117. G. F. Shields, A. M. Schmeichen, B. L. Frazier, A. Redd, M. I. Vovoeda, J. K. Reed, and R. H. Ward. mtDNA sequences suggest a recent evolutionary divergence for Beringian and Northern North American populations. *Am. J. Hum. Genet.*, 53:549–562, 1993.
118. P. F. Slade. Most recent common ancestor probability distributions in gene genealogies under selection. *Theor. Popul. Biol.*, 58:291–305, 2000.
119. P. F. Slade. Simulation of 'hitch-hiking' genealogies. *J. Math. Biol.*, 42:41–70, 2001.

120. M. Slatkin. Simulating genealogies of selected alleles in a population of variable size. *Genet. Res.*, 78:49–57, 2001.
121. M. Slatkin and B. Rannala. Estimating the age of alleles by use of intraallelic variability. *Am. J. Hum. Genet.*, 60:447–458, 1997.
122. M. Slatkin and B. Rannala. Estimating allele age. *Annu. Rev. Genomics Hum. Genet.*, 1:225–249, 2000.
123. C. Soligo, O. Will, S. Tavaré, C. R. Marshall, and R. D. Martin. New light on the dates of primate origins and divergence. In M. J. Ravosa and M. Dagosto, editors, *Primate Origins and Adaptations*. Kluwer Academic/Plenum Publishers, New York, 2003.
124. J. Claiborne Stephens, Julie A. Schneider, Debra A. Tanguay, Julie Choi, Tara Acharya, Scott E. Stanley, Ruhong Jiang, Chad J. Messer, Anne Chew, Jin-Hua Han, Jicheng Duan, Janet L. Carr, Min Seob Lee, Beena Koshy, A. Madan Kumar, Ge Zhang, William R. Newell, Andreas Windemuth, Chuanbo Xu, Theodore S. Kalbfleisch, Sandra L. Shaner, Kevin Arnold, Vincent Schulz, Connie M. Drysdale, Krishnan Nandabalan, Richard S. Judson, Gualberto Ruanwo, and Gerald F. Vovis. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, 293:489–493, 2001.
125. M. Stephens. Times on trees and the age of an allele. *Theor. Popul. Biol.*, 57:109–119, 2000.
126. M. Stephens. Inference under the coalescent. In D. J. Balding, M. J. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, pages 213–238. John Wiley and Sons, Inc., New York, New York., 2001.
127. M. Stephens and P. Donnelly. Inference in molecular population genetics. *J. Roy. Statist. Soc. B*, 62:605–655, 2000.
128. F. M. Stewart. Variability in the amount of heterozygosity maintained by neutral mutations. *Theoret. Popn. Biol.*, 9:188–201, 1976.
129. C. Strobeck and K. Morgan. The effect of intragenic recombination on the number of alleles in a finite population. *Genetics*, 88:828–844, 1978.
130. F. Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105:437–460, 1983.
131. H. Tang, D. O. Siegmund, P. Shen, P. J. Oefner, and M. W. Feldman. Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics*, 161:447–459, 2002.
132. S. Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoret. Popn. Biol.*, 26:119–164, 1984.
133. S. Tavaré. Calibrating the clock: using stochastic processes to measure the rate of evolution. In E. S. Lander and M. S. Waterman, editors, *Calculating the secrets of life*, pages 114–152. National Academy Press, Washington DC, 1993.
134. S. Tavaré. Ancestral inference from DNA sequence data. In H. G. Othmer, F. R. Adler, M. A. Lewis, and J. Dallon, editors, *Case Studies in Mathematical Modeling: Ecology, Physiology, and Cell Biology*, pages 81–96. Prentice-Hall, 1997.
135. S. Tavaré. Ancestral inference for branching processes. In P. Haccou and P. Jagers, editors, *Branching Processes in Biology: Variation, Growth, Extinction*. Cambridge University Press., 2004. In press.
136. S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times for molecular sequence data. *Genetics*, 145:505–518, 1997.

137. S. Tavaré and W. J. Ewens. Multivariate Ewens distribution. In N. S Johnson, S. Kotz, and N. Balakrishnan, editors, *Discrete Multivariate Distributions*, chapter 41, pages 232–246. Wiley, New York, 1997.
138. S. Tavaré, C. R. Marshall, O. Will, C. Soligo, and R. D. Martin. Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature*, 416:726–729, 2002.
139. J. L. Thorne, H. Kishino, and J. Felsenstein. Inching towards reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.*, 34:3–16, 1992.
140. J. Tsao, Y. Yatabe, R. Salovaara, H. J. Järvinen, J. Mecklin, L. A. Altonen, S. Tavaré, and D. Shibata. Genetic reconstruction of individual colorectal tumor histories. *Proc. Natl. Acad. Sci. USA*, 97:1236–1241, 2000.
141. J. Wakeley. Using the variance of pairwise differences to estimate the recombination rate. *Genet. Res. Camb.*, 69:45–58, 1997.
142. J. D. Wall. A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.*, 17:156–163, 2000.
143. R. H. Ward, B. L. Frazier, K. Dew, and S. Pääbo. Extensive mitochondrial diversity within a single Amerindian tribe. *Proc. Natl. Acad. Sci. USA*, 88:8720–8724, 1991.
144. M. S. Waterman. *Introduction to Computational Biology. Maps, Sequences and Genomes*. Chapman & Hall, London, 1995.
145. G. A. Watterson. Models for the logarithmic species abundance distributions. *Theoret. Popn. Biol.*, 6:217–250, 1974.
146. G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoret. Popn. Biol.*, 7:256–276, 1975.
147. G. A. Watterson. Heterosis or neutrality? *Genetics*, 85:789–814, 1977.
148. G. A. Watterson. Reversibility and the age of an allele II. Two-allele models, with selection and mutation. *Theoret. Popul. Biol.*, 12:179–196, 1977.
149. G. A. Watterson. The homozygosity test of neutrality. *Genetics*, 88:405–417, 1978.
150. G. A. Watterson. Motoo Kimura's use of diffusion theory in population genetics. *Theoret. Popul. Biol.*, 49:154–188, 1996.
151. G. Weiss and A. von Haeseler. Estimating the age of the common ancestor of men from the ZFY intron. *Science*, 257:1359–1360, 1996.
152. G. Weiss and A. von Haeseler. A coalescent approach to the polymerase chain reaction. *Nucleic Acids Research*, 25:3082–3087, 1997.
153. G. Weiss and A. von Haeseler. Inference of population history using a likelihood approach. *Genetics*, 149:1539–1546, 1998.
154. K. M. Weiss and A. G. Clark. Linkage disequilibrium and the mapping of complex human traits. *TIG*, 18:19–24, 2002.
155. L. S. Whitfield, J. E. Sulston, and P. N. Goodfellow. Sequence variation of the human Y chromosome. *Nature*, 378:379–380, 1995.
156. O. Will, V. Plagnol, C. Soligo, R. D. Martin, and S. Tavaré. Statistical inference in the primate fossil record: a Bayesian approach. In preparation, 2003.
157. I. J. Wilson and D. J. Balding. Genealogical inference from microsatellite data. *Genetics*, 150:499–510, 1998.
158. C. Wiuf and P. Donnelly. Conditional genealogies and the age of a neutral mutant. *Theor. Popul. Biol.*, 56:183–201, 1999.
159. S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.

160. S. Wright. *Evolution and the Genetics of Populations.*, volume 4, Variability within and among natural populations. University of Chicago Press, Chicago, 1978.
161. Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Mol. Biol. Evol.*, 14:717–724, 1997.