5. ANCESTRAL INFERENCE FROM DNA SEQUENCE DATA

Simon Tavaré

5.1 INTRODUCTION

After the pioneering paper of Cann et al. (1987), many authors have discussed methods for inferring ancestral history from samples of DNA sequences taken from human populations. Much of this research has focused on the evolution of mitochondrial DNA. These molecules have been exploited in evolutionary studies because of their high mutation rate; this means that DNA sequence differences can be detected between individuals who are quite closely related. In addition, mitochondria are maternally inherited, so these molecules are particularly suited to studying the female lineages in which they arise. One tantalizing problem, usually referred to as "the time to Mitochondrial Eve," is to estimate the time to the most recent common mitochondrial ancestor of the population from which the sample sequences were drawn. The papers of Templeton (1993), Ayala (1995), Wallace (1995) and Wills (1995) provide further background and discussion. More recently, DNA sequence data from the male-specific part of the Y chromosome have begun to appear, along with analyses of the time to "Y Adam." See Dorit et al. (1995), Hammer (1995), Whitfield et al. (1995), and the review of Jobling and Tyler-Smith (1995).

In this paper we describe one approach to drawing inferences about the distribution of the time to the most recent common ancestor (TMRCA) of a population, given data from a sample of DNA sequences taken from that population. In practice we do not know the ancestral history of the DNA sequences in the sample in any detail. Therefore statistical statements about TMRCA have to be based on a stochastic model for this ancestry. We use a model called the *coalescent* (Kingman 1982a; Griffiths 1980; Hudson 1983; Tajima 1983), reviewed briefly in Sections 5.2 and 5.3. The effects of deterministic fluctuations in population size are discussed in Section 5.4.

The sample consists of n individuals from the population of interest. The data $\mathcal{D}$ in the sample are n DNA sequences from a given molecular region. For example, many mitochondrial data sets contain sequences from the control region of the molecule, while Dorit et al. sequenced an intron of the ZFY locus on the Y chromosome. We assume the sequences in the sample are the same length. We can then think of the data as a matrix $X = (x_{ij})$ with n rows, where the entry x_{ij} records the DNA base (either A, C, G, or T) in individual i at site j. A *site*, then, refers to a location in the DNA. It is sometimes convenient to ignore any columns of X that have identical bases in every sequence. The remaining columns are referred to as *segregating sites*; they comprise locations in the DNA sequences where not every individual is identical. The differences observed in the sample sequences arise from the effects of mutation in their ancestry. We suppose that these differences are due to the effects of *substitutions*, the replacement of one base by another when a mutation occurs. We model the locations of the mutations in the ancestral tree of the sample in Section 5.5.

In practice it is often either difficult or uninformative to get explicit mathematical expressions for quantities of interest such as the conditional distribution of TMRCA given the data $\mathcal{D}$. Instead we use a computational approach that simulates observations from the required conditional distribution. Summary statistics such as histograms and moments can then be found from these simulated values in the usual way. In this chapter we summarize the data matrix X in terms of the random quantity S_n, the number of segregating sites in the sample. Conditional distributions of TMRCA given $S_n = k$ can be found by a rejection method, discussed briefly in Section 5.6. Applications of these simulation methods to Y-chromosome data are given in Section 5.7.

5.2 THE COALESCENT

Inferences about the TMRCA of a population are to be made on the basis of a comparison of the DNA sequences from the molecular region of interest from a sample of people in the population. Differences in these sequences come from the effects of mutation in the unknown ancestry of the sample. It follows that to study TMRCA we need a stochastic model for this ancestry. In the molecular regions of interest here (the intron in the ZFY locus on the Y chromosome or the D loop of the mitochondrion for example) there appears to be no recombination. The molecular region is passed on intact, modulo the effects of substitutions, from parent to offspring. As a result, each molecule (or "individual") has a single haploid "parent" in the previous generation (the molecule from which it was copied), that "parent" itself has a single parent in the previous generation, and so on back into time. It is this genealogical process that we have to model.

Population geneticists have modeled such genealogies in a variety of circumstances, in particular when the population size is large. Consider then a particular generation in a large random mating population of constant-size N haploid individuals, and label them $1, 2, \ldots, N$. Population genetics models are often defined by specifying the joint distribution of the numbers $\nu_1, \nu_2, \ldots, \nu_N$ of offspring born to individuals $1, 2, \ldots, N$. For example, the classical Wright-Fisher model specifies that the offspring numbers have a symmetric multinomial distribution:

$$\mathbb{P}(\nu_1 = m_1, \ldots, \nu_N = m_N) = \frac{N! N^{-N}}{m_1! \cdots m_N!} \tag{5.1}$$

where $m_1, \ldots, m_N \in \{0, 1, \ldots, N\}$ satisfy $m_1 + \cdots + m_N = N$ and the offspring numbers in different generations are independent and identically distributed. This prescription shows how to construct the model forwards in time. However, for our inference problem it is much more convenient to study not how parents have offspring, but rather how children "choose" their parents.

The Wright-Fisher model can be described by saying that each individual in a given generation chooses its parent independently of others in its generation, uniformly and at random from the N potential parents in the previous generation. Continuing this process back into the past yields a genealogical tree that relates the individuals in a given generation to their parents, grandparents, and so on.

This genealogy is hard to analyze for a given fixed value of N, but it may be approximated in a simple way when N is large. Notice that the chance that two randomly chosen individuals have distinct parents in the previous generation is $1 - N^{-1}$. It follows that the chance that these two have distinct ancestors in generations 1,2, ... , r is $(1 - N^{-1})^r$. If we measure time in units of N generations, so that $r \approx Nt$ for some $t > 0$, we see that the time W_2 during which the sample of two individuals has no common ancestor satisfies

$$\mathbb{P}(W_2 > t) = \left(1 - \frac{1}{N}\right)^{Nt} \approx e^{-t}. \tag{5.2}$$

This shows that in a large population, the time W_2 until two individuals have a common ancestor has (approximately) an exponential distribution with mean 1. What of the genealogy of a sample of size three? Looking back into the past, there will be a first time at which some members of the sample share a common ancestor. At this time, either all three will have a common ancestor, or a particular pair will. In a large population, this last possibility is overwhelmingly the most likely. Furthermore, the time W_3 (measured once more in units of N generations) has approximately an exponential distribution with mean $\binom{3}{2} = 3$, so

$$\mathbb{P}(W_3 > t) \approx e^{-3t}. \tag{5.3}$$

At the time the first pair of individuals has found a common ancestor, the sample of three individuals has two distinct ancestors. The additional time taken for these two to find their common ancestor has the distribution of W_2, independent of W_3.

Thus in a large population we can give a simple description of the genealogy of a sample of n individuals. This stochastic process, known as the *coalescent*, describes the genealogical tree of the sample as time goes back into the past. With time measured in units of N generations, the time W_j during which the sample has j distinct ancestors has an exponential distribution with parameter $\binom{j}{2} = j(j-1)/2$, the times $W_n, W_{n-1}, \ldots, W_2$ being independent for different j. W_j should be thought of as the length of each of the j branches of the genealogical tree when the sample has j distinct ancestors. This tree is bifurcating; at the time W_n, two of the n ancestors are chosen at random and their branches are joined, giving $n - 1$ ancestors for the sample. At the time $W_n + W_{n-1}$, two of these $n-1$ ancestors are chosen at random and their branches are joined, resulting in $n-2$ distinct ancestors in the sample. This process continues until the time

$$T_n = W_n + \cdots + W_2, \tag{5.4}$$

when all the individuals in the sample have been traced back to their most recent common ancestor (MRCA). A sample path of this process appears in Figure 5.1.

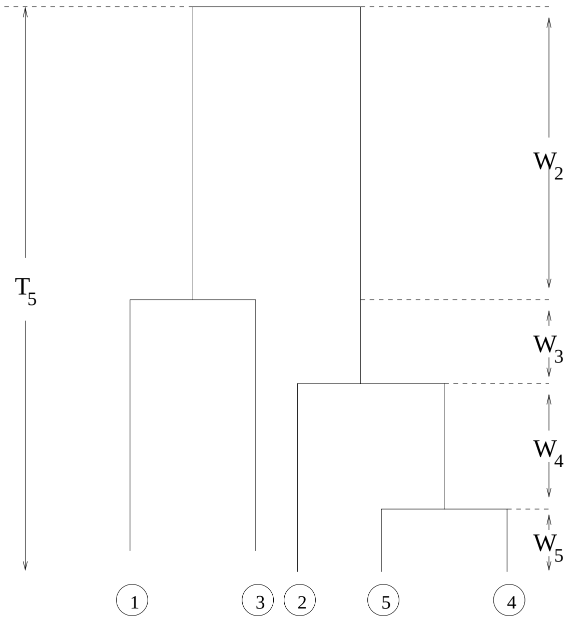

Figure 5.1. A sample path of the coalescent for a sample of size $n = 5$.

The previous discussion was based on the Wright-Fisher model. Remarkably, the same approximation applies to a very wide class of discrete exchangeable reproduction models. Kingman (1982a; 1982b) showed how the coalescent arises as the limiting approximation (as the population size $N \to \infty$) to these underlying discrete genealogies. In this approximation, time is measured in units of $\sigma^{-2}N$ generations, where $\sigma^2 \in (0, \infty)$ is the limiting variance of the number ν_1 of offspring born to a typical individual. For ease of exposition, we assume $\sigma^2 = 1$ (as it is for the Wright-Fisher model) in what follows.

The mean time to the MRCA, and so the mean height of the ancestral tree, can be found from (5.4) as

$$\begin{aligned}\mathbb{E}T_n &= \mathbb{E}(W_n + \cdots + W_2) && (5.5)\\ &= \mathbb{E}W_n + \cdots + \mathbb{E}W_2 && (5.6)\\ &= \frac{2}{n(n-1)} + \cdots + \frac{2}{2(2-1)} && (5.7)\\ &= 2\left(1 - \frac{1}{n}\right) && (5.8)\end{aligned}$$

in coalescent units. The variance of T_n can be computed easily because the W_j are independent and exponentially distributed. We obtain

$$\mathrm{Var}(T_n) = \sum_{j=2}^{n} \frac{4}{j^2(j-1)^2}. \qquad (5.9)$$

In large samples this variance is about 1.16, most of which comes from the time W_2 when the sample has just two ancestors. Times are often converted from the coalescent time T_n to years T_n^y via

$$T_n^y = T_n \times N \times G, \tag{5.10}$$

where G is the number of years in a generation.

5.2.1 The ancestral process

In the sequel we make use of the Markov chain $\{A_n(t), t \geq 0\}$ that counts the number of distinct ancestors of the sample of size n at times $t \geq 0$. In Markov chain parlance, this is a *death process*: it starts from $A_n(0) = n$, waits an exponential amount of time W_j in state j, and then moves to state $j-1$ and so forth. Eventually the process is absorbed in the state 1, at the time T_n. The probability distribution $g_{nj}(t)$ of $A_n(t)$ was found by Griffiths (1979).

$$\begin{aligned} g_{nj}(t) &= \mathbb{P}(A_n(t) = j) \\ &= \sum_{k=j}^{n} (-1)^{k-j} e^{-k(k-1)t/2} \frac{(2k-1)\, j_{(k-1)} n_{[k]}}{j!(k-j)! n_{(k)}}, \end{aligned} \tag{5.11}$$

where we have used the notation

$$a_{(n)} = a(a+1)\cdots(a+n-1); \quad a_{(0)} = 1; \tag{5.12}$$

$$a_{[n]} = a(a-1)\cdots(a-n+1); \quad a_{[0]} = 1. \tag{5.13}$$

Because $\{T_n \leq t\} = \{A_n(t) = 1\}$, the distribution function of T_n follows immediately from (5.11):

$$\mathbb{P}(T_n \leq t) = g_{n1}(t), \quad t \geq 0. \tag{5.14}$$

While this provides an explicit formula for the distribution of T_n, it is harder to find explicit results for other quantities of interest such as the distribution of the total length L_n of the tree, defined by

$$L_n = nW_n + (n-1)W_{n-1} + \cdots + 2W_2. \tag{5.15}$$

Instead we can resort to a Monte Carlo approach, in which observations having the required distribution are simulated. These simulated values can then be used to estimate the probability density of the underlying random variable, together with any required statistics such as percentiles, mean and variance. A convenient introduction to stochastic simulation can be found in Ripley (1987). To illustrate the ideas, we give an algorithm for simulating the times $W_n, W_{n-1}, \ldots, W_2$.

Algorithm 1 *Algorithm to generate $W_n, \ldots, W_2$ for constant population size.* U denotes a random variable with the uniform distribution on (0,1), generated independently at each use.

```
1. Set $t = 0$, $j = n$.

2. Generate $s = -2\log(U)/j(j-1)$.
```

3. Set $w_j = s$, $t = t + s$.
4. Set $j = j - 1$. If $j \geq 2$, go to 2. Else return $T_n = t$, $W_n = w_n, \ldots, W_2 = w_2$.

Step 2 generates an observation having the exponential distribution with mean $2/j(j-1)$, just as needed for W_j. The value t returned in Step 4 has the distribution of $T_n = W_n + \cdots + W_2$. The algorithm can be modified to generate observations having the distribution of L_n; simply set $l = 0$ at Step 1, $l = l + js$ at Step 3, and return $L_n = l$ at the end of Step 4. Later in this chapter, we exploit this simulation approach in cases where exact results are unobtainable.

Genealogical methods based on variations of the coalescent, using both theoretical and simulation approaches, have proved very powerful for understanding the structure of complex stochastic models in population genetics and as a useful guide to intuition in understanding the evolution of many population genetic phenomena. The recent reviews of Hudson (1991); Hudson (1992) and Donnelly and Tavaré (1995) describe some of these developments.

5.3 THE BIVARIATE ANCESTRAL PROCESS

In order to study TMRCA for a population given sequence data from a sample, we need to understand the *joint* behavior of the genealogy of both population and sample. We make use of the process $\{(A_m(t), A_n(t)), t \geq 0\}$ that counts the number of ancestors in a population of size m and a random sample of size n taken from it. (It is convenient to refer to the set of m individuals as the population, and the subset of size n as the sample. This avoids ambiguity and terms like "sample" and "subsample" or "supersample" and "sample".) This bivariate process is Markovian and it makes transitions from a state of the form (i, j) whenever two individuals in the current population of size i share a common ancestor. If this coalescence event involves the ancestors of two individuals in the sample, then the new state becomes $(i-1, j-1)$. In any other case, it is $(i-1, j)$. From (i, j) we move to

$$(i-1, j) \qquad \text{at rate} \quad (i(i-1) - j(j-1))/2 \tag{5.16}$$

$$(i-1, j-1) \qquad \text{at rate} \quad j(j-1)/2. \tag{5.17}$$

A sample path of the bivariate process for $m = 9$, $n = 5$ is given in Figure 5.2.

The distribution of $(A_m(t), A_n(t))$ was found by Saunders et al. (1984) as

$$\mathbb{P}(A_m(t) = l, \quad A_n(t) = k) = g_{ml}(t) q(n, k \mid m, l), \tag{5.18}$$

where $g_{ml}(t)$ is given in (5.11), and $q(n, k \mid m, l)$ is given by

$$q(n, k \mid m, l) = \tag{5.19}$$
$$\frac{(m-n)!(m-l)!n!(n-1)!l!(l-1)!(m+k-1)!}{(n-k)!(l-k)!m!(m-1)!k!(k-1)!(l+n-1)!(m+k-l-n)!}.$$

The quantity $q(n, k \mid m, l)$ is the probability that the sample of size n taken at any time $t > 0$ has k distinct ancestors given that the population of size m has l ancestors at that time.

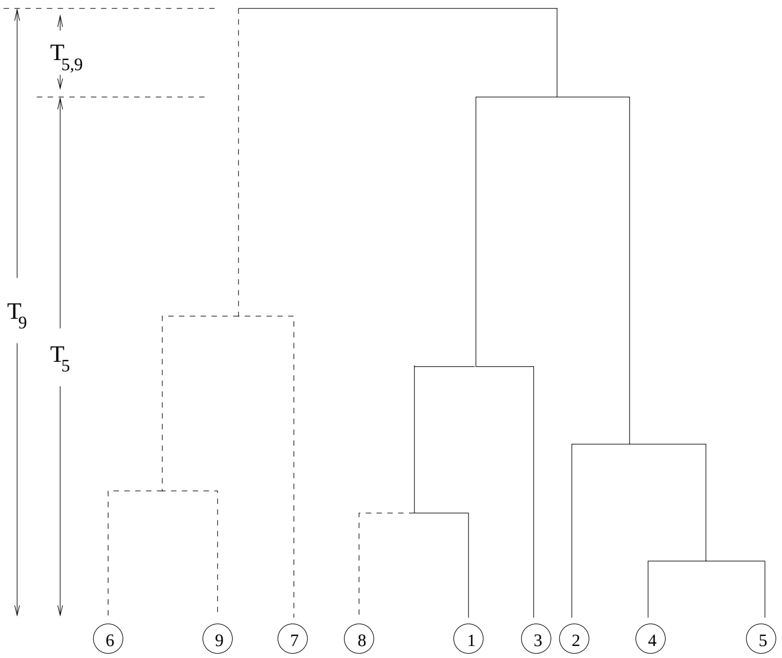

Figure 5.2. A sample path of the bivariate coalescent for a population of size $m = 9$ and a sample of size $n = 5$. The sample individuals are labeled 1,2, . . . , 5.

The sample path in Figure 5.2 shows that at the time T_n when the sample reaches its MRCA, the number $A_m(T_n)$ of distinct ancestors of the population is random. The distribution of the number $A_m(T_n)$ is known. In particular, Watterson (1982) showed that the probability that the population of size m and a sample of size n share a common ancestor is

$$\mathbb{P}(A_m(T_n) = 1) = \frac{(n-1)(m+1)}{(n+1)(m-1)}. \tag{5.20}$$

If the sample is at all large, there is an appreciable chance that the sample and the subsample will share their MRCA, and that the time to the MRCA is thus the same for both sample and subsample. On the other hand, if they do not share a common ancestor then the *extra* time T_{nm} required to reach the MRCA of the sample is stochastically larger than W_2, the time taken for two individuals to be traced back to their common ancestor.

5.4 VARIABLE POPULATION SIZE

In order to apply coalescent methods to human population data, we need to account for the effects of variations in population size through time. Fortunately, this is straightforward in the case of deterministic fluctuations. To keep the presentation simple, we concentrate on the approximation to the Wright-Fisher model once more. The effect of variable population size is to change the joint distribution of the times W_j (Kingman 1982b; Griffiths and

Tavaré 1994b; Donnelly and Tavaré 1995). In particular, these times are no longer independent. We assume that the population size at the time of sampling is N, and again measure time in units of N generations. We write $Np(t)$ for the population size a (coalescent) time t ago and define $\lambda(t) = 1/p(t)$. The conditional distribution of the time W_j for which there are exactly j ancestors of the sample, given that the time for which there are more than j ancestors is s, is

$$\mathbb{P}(W_j > t | W_n + \cdots + W_{j+1} = s) = \exp\left(-\binom{j}{2}\int_s^{s+t} \lambda(v)\,dv\right). \tag{5.21}$$

We assume that $\int_0^\infty \lambda(v)\,dv = \infty$ to ensure that any pair of individuals (and thus the sample) can be traced back to a common ancestor.

The process $\{A_n^v(t), t \geq 0\}$ that counts the number of ancestors at time t of a sample of size n taken at time 0 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

$$A_n^v(t) = A_n(\Lambda(t)), \quad t \geq 0, \tag{5.22}$$

where $A_n(\cdot)$ is the corresponding ancestral process for the constant-population-size case, and

$$\Lambda(t) = \int_0^t \lambda(s)\,ds. \tag{5.23}$$

Thus 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,

$$\mathbb{P}(A_n^v(t) = j) = g_{nj}(\Lambda(t)), \quad j = 1, \ldots, n \tag{5.24}$$

where $g_{nj}(t)$ is given in (5.11), and so

$$\mathbb{P}(T_n \leq t) = \mathbb{P}(A_n^v(t) = 1) = g_{n1}(\Lambda(t)), \quad t \geq 0. \tag{5.25}$$

Once more, simulation provides a valuable way to study properties of genealogy when the population size varies. The representation (5.21) gives a direct way to simulate the times $W_n, W_{n-1}, \ldots, W_2$.

Algorithm 2 *Algorithm to generate $W_n, \ldots, W_2$with variable population size.* U denotes a random variable with a uniform distribution on (0,1), generated independently at each use.

1. Set $t = 0$, $j = n$.
2. Generate
$$w_j^* = \frac{-2\log(U)}{j(j-1)}.$$
3. Solve for s the equation
$$\Lambda(t+s) - \Lambda(t) = w_j^*. \tag{5.26}$$

4. Set $w_j = s$, $t = t + s$.

5. Set $j = j - 1$. If $j \geq 2$, go to 2. Else return $T_n = t$, $W_n = w_n, \ldots, W_2 = w_2$.

As noted after Algorithm 1, w_j^* generated in step 2 has an exponential distribution with mean $2/j(j-1)$. If the population size is constant, then $\Lambda(t) = t$, and Algorithm 2 reduces to Algorithm 1. Observations having the distribution of the tree length L_n can be generated as described after Algorithm 1.

5.4.1 The bivariate process revisited

The analysis of the bivariate ancestral process with variable population size follows immediately from the representation

$$(A_m^v(t), A_n^v(t)) = (A_m(\Lambda(t)), A_n(\Lambda(t))), \quad t \geq 0. \tag{5.27}$$

From this follows the fact that

$$\mathbb{P}(A_m^v(t) = l, A_n^v(t) = k) = g_{ml}(\Lambda(t))q(n, k \mid m, l), \tag{5.28}$$

where $q(n, k \mid m, l)$ is given in (5.19). Note that the combinatorics of the bivariate process remain as they were in the constant-population-size case; only the waiting times between the jumps of the process change. In particular, the probability that the population and the sample share their MRCA is still given by (5.20).

Distributions in the bivariate process can also be simulated easily. Algorithm 3 gives a method for simulating values of the height $T_n = W_n + \cdots + W_2$ of the coalescent tree of the sample of size n, the number $A_m(T_n)$ of ancestors of the sample at the time the subsample reaches its MRCA, and the time T_{nm} from then until the population reaches its MRCA. This extra time may, of course, be 0.

Algorithm 3 *Algorithm for bivariate ancestral process.* U denotes uniform (0,1) random variable, independently generated at each use.

1. Set $a_m = m$, $a_n = n$, $t = 0$.

2. Set $w = -2\log(U)/(a_m(a_m - 1))$.

3. Solve for s the equation $\Lambda(t + s) - \Lambda(t) = w$.

4. Set $t = t + s$.

5. Set

$$p = \frac{a_n(a_n - 1)}{a_m(a_m - 1)} \tag{5.29}$$

and $a_m = a_m - 1$.

6. With probability p, set $a_n = a_n - 1$. If $a_n > 1$, go to 2.

7. Set $t_n = t$, $a^* = a_m$. If $a^* = 1$, set $t_{mn} = 0$, and stop. Otherwise, use Algorithm 2 starting from $t = t_n$, $n = a^*$ to generate an observation t_{nm} on the total height of a coalescent tree of a^* individuals, then stop.

The values of t_n, a^*, and t_{nm} returned by a single pass through Algorithm 3 have the joint distribution of the height T_n, the number of ancestors $A_m(T_n)$ of the population at the time the subsample finds its MRCA, and the additional time required to get to the MRCA of the population. Note that $t_m = t_n + t_{nm}$ has the distribution of T_m. More detailed information about the genealogical trees could also be recorded, but this is all we need later on.

5.5 MUTATIONS IN THE GENEALOGICAL TREE

To model the effect of mutations in the genealogy of the sample, we assume that the times at which mutations occur form a Poisson process of constant rate $\theta/2$, independently in each branch of the tree. A branch of length w therefore has a Poisson number of mutations with mean $w\theta/2$. The parameter θ is defined by

$$\theta = 2N\mu, \tag{5.30}$$

where μ is the mutation rate per gene per generation in the underlying discrete model. When the mutation rate μ is of the order of the reciprocal of the population size N, the genealogy and the genetics compete on equal terms; both features are included in the coalescent approximation.

To model the evolution of DNA-sequence data we have to describe how a sequence is changed when a mutation occurs in it. We here use the infinitely-many-sites model of Watterson (1975). Because we are ignoring the effects of recombination (it is not thought to occur in the data at hand), each sequence may be thought of as a completely linked sequence of DNA sites. Whenever a mutation occurs, it occurs at a site that has not had a mutation before.

Observing that each mutation in the coalescent tree introduces a new segregating site into the sample, the number of segregating sites S_n in the sample of n chromosomes is precisely the number of mutations that arise in its genealogical tree. This in turn has a Poisson distribution with mean $\theta L_n/2$, where L_n is the total length of the tree, defined in (5.15) by $L_n = nW_n + (n-1)W_{n-1} + \cdots + 2W_2$. That is,

$$\mathbb{P}(S_n = k \mid L_n = l) = \mathrm{Po}\,(k, \theta l/2), \tag{5.31}$$

where $\mathrm{Po}\,(k, \mu)$ is the Poisson probability

$$\mathrm{Po}\,(k, \mu) = e^{-\mu}\frac{\mu^k}{k!}, \quad k = 0, 1, \ldots . \tag{5.32}$$

5.6 CONDITIONING ON THE DATA

Our aim is to find the distribution of the time to the MRCA of a population of size m given data from a sample of n individuals. In this chapter, we summarize the data by taking $\mathcal{D}$ to be the number of segregating sites in the DNA sequences. Prior to sampling, the required probability density is that of T_m, defined as

$$T_m = W_m + \cdots + W_2. \tag{5.33}$$

We call this the *predata* density of TMRCA. The *postdata* density function is that of T_m given $\mathcal{D} = \{S_n = k\}$, which, using Bayes' formula, satisfies

$$f_{T_m}(t \mid \mathcal{D}) \propto f_{T_m}(t)\mathbb{P}(S_n = k \mid T_m = t). \tag{5.34}$$

As in (13) of Tavaré et al. (1996), this can be expressed as follows:

$$\begin{aligned} f_{T_m}(t)\mathbb{P}(S_n = k \mid T_m = t) &= \int_0^\infty f_{T_m,L_n}(t,l)\mathbb{P}(S_n = k \mid T_m = t, L_n = l)\,dl \\ &= \int_0^\infty f_{T_m,L_n}(t,l)\mathbb{P}(S_n = k \mid L_n = l)\,dl \\ &= \int_0^\infty f_{T_m,L_n}(t,l)\mathrm{Po}(k, l\theta/2)\,dl. \end{aligned} \tag{5.35}$$

In (5.35), $f_{T_m,L_n}(t,l)$ is the joint probability density of T_m and L_n in the bivariate coalescent process. Supposing for the moment that an observation (t,l) could be generated from the joint density of T_m and L_n, we see from (5.35) that the *rejection method* can be used to generate from the conditional distribution in (5.35). The idea is to keep the observation t with probability $u = \mathrm{Po}\,(k, l\theta/2)$, and reject it otherwise. The rejection step can be improved by noting that t can be accepted with probability u/c for any constant $c > u$. Because $\mathrm{Po}\,(k, l\theta/2) \le \mathrm{Po}\,(k,k)$, we may take $c = \mathrm{Po}\,(k,k)$, and hence we accept t with probability u given by

$$u = \frac{\mathrm{Po}\,(k, l\theta/2)}{\mathrm{Po}\,(k,k)}. \tag{5.36}$$

Ripley (1987)(pg. 60) gives a description of the rejection method.

To generate an observation from the joint distribution of T_m and L_n, we can use Algorithm 3 directly. In addition, it generates observations from the conditional distribution of the number $A_m(T_n)$ of ancestors of the population of size m given $\mathcal{D}$. In summary, we have the following algorithm.

Algorithm 4 *Rejection algorithm for* $f_{T_m}(t \mid S_n = k)$. U denotes a uniform (0,1) random variable, independently generated at each use.

1. Set $a_m = m$, $a_n = n$, $t = 0$, $l = 0$.
2. Set $w = -2\log(U)/a_m(a_m - 1)$.
3. Solve for s the equation $\Lambda(t+s) - \Lambda(t) = w$.
4. Set $t = t + s$, $l = l + a_n s$.
5. Set
$$p = \frac{a_n(a_n - 1)}{a_m(a_m - 1)}, \tag{5.37}$$
and $a_m = a_m - 1$.

6. With probability p, set $a_n = a_n - 1$. If $a_n > 1$, go to 2.
7. Set $u = \text{Po}\,(k, l\theta/2)/\text{Po}\,(k, k)$. Accept (t, a_m) with probability u, else go to 1.
8. Set $t_n = t$, $a^* = a_m$. If $a^* = 1$, set $t_{mn} = 0$, and stop. Otherwise, use Algorithm 2 starting from $t = t_n, n = a^*$ to generate an observation t_{nm} on the total height of a coalescent tree of a^* individuals, then stop.

The values of t_n, a^*, t_{nm} generated by a single run through Algorithm 4 have the joint distribution of the sample tree height T_n, the number of ancestors $A_m(T_n)$ of the sample at the time the sample finds its MRCA, and the additional time T_{nm} required to get to the MRCA of the population conditional on having observed k segregating sites in the sample. The value of $t_m = t_n + t_{nm}$ has the distribution of T_m given $S_n = k$.

5.7 APPLICATIONS

Whitfield et al. (1995) sequenced a region of 15,680 base pairs from the Y chromosome of $n = 5$ individuals. They observed just three segregating sites and estimated the coalescence time of the sample to be between 37,000 and 49,000 years. Their analysis was not based on a population genetics model. Tavaré et al. (1996) use coalescent methods to reanalyze these data, using a number of plausible scenarios about variability in the effective population size N and the underlying mutation rate μ. Whitfield et al. (1995) estimated the mutation rate in the region to be $\mu = 3.52 \times 10^{-4}$ substitutions per generation, based on a generation time of $G = 20$ years. Using an estimate of $N = 4,900$, the value used by Hammer (1995), Tavaré et al. (1996) found a 95% credible region for TMRCA of the sample of 30,000–183,000 years.

Here we examine two aspects in more detail: we estimate TMRCA for the population, and we estimate the chance that the sample and the population share their most recent common ancestor, given the data of 3 segregating sites from the sample of 5 individuals. For illustration, we use a model of deterministic fluctuation in population size of the form

$$\lambda(t) = \begin{cases} \alpha^{-1}, & t > V, \\ \alpha^{-t/V}, & 0 \le t \le V. \end{cases} \qquad (5.38)$$

This corresponds to a model in which the population has constant relative size $\alpha \in (0, 1)$ prior to time V ago, and exponential growth to relative size 1 at the time of sampling. We take a value of 50,000 years for V, and $\alpha = 10^{-4}$. For comparison with our earlier work, we suppose that the effective size in the constant phase is 4900, so that $N = 4.9 \times 10^7$.

To apply Algorithm 4 we need a value to use for m. In practice, it is difficult to detect a difference between the distributions of, say, T_{200} and T_{500}, and we choose for this illustration a value of $m = 200$. We use Algorithms 3 and 4 to simulate 10,000 observations from the predata distribution of (T_5, T_{200}) and the corresponding postdata distribution, given that in the sample of size 5 there are 3 segregating sites.

For this demographic model, the predata distribution of T_5 has a mean of 199,000 years and 95% of the distribution of lies in the interval (76,000–464,000) years. (Here and in what follows, all ages are rounded to the nearest 1000 years. We use the shorthand I95 to denote the interval such that 2.5% of the mass of a distribution is to the left of the left endpoint, 2.5% to the right of the right endpoint.) The predata distribution of T_{200}, representing the TMRCA of the whole population, has a mean of 238,000 years, and an I95 of (113,000–504,000). In the 10,000 simulations, the observed fraction of times that the sample and the population had the same MRCA was 0.671, in good agreement with the theoretical value of 0.673 from (5.20).

The postdata distribution of T_5 has a mean of 108,000 years, and an I95 of (61,000–194,000) years. Note that the postdata distribution suggests a much shorter time for the TMRCA of the *sample*, and the postdata distribution is much more concentrated than the predata distribution. The estimated densities are plotted in Figure 5.3.

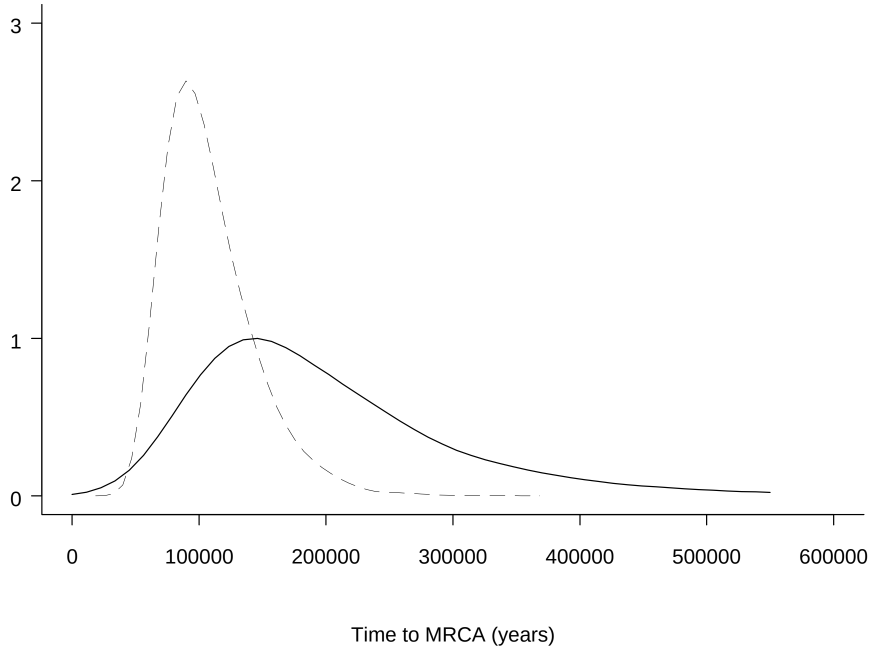

Figure 5.3. Density functions for the pre-data (solid lines) and post-data distribution of T_5. x-axis is in years.

The proportion of the 10,000 runs that resulted in the population and the sample having a common MRCA was 0.233, markedly smaller than the predata fraction. On the basis of this, we anticipate that the additional time to the MRCA of the population will be much larger than the corresponding increment in the predata distribution. This is indeed the case; the former has a mean of 101,000 years, the latter 39,000 years. The mean time to the population MRCA is 209,000 years, with an I95 of (99,000–465,000) years. This interval is somewhat shorter than the I95 for the predata distribution. The estimated densities are plotted in Figure 5.4.

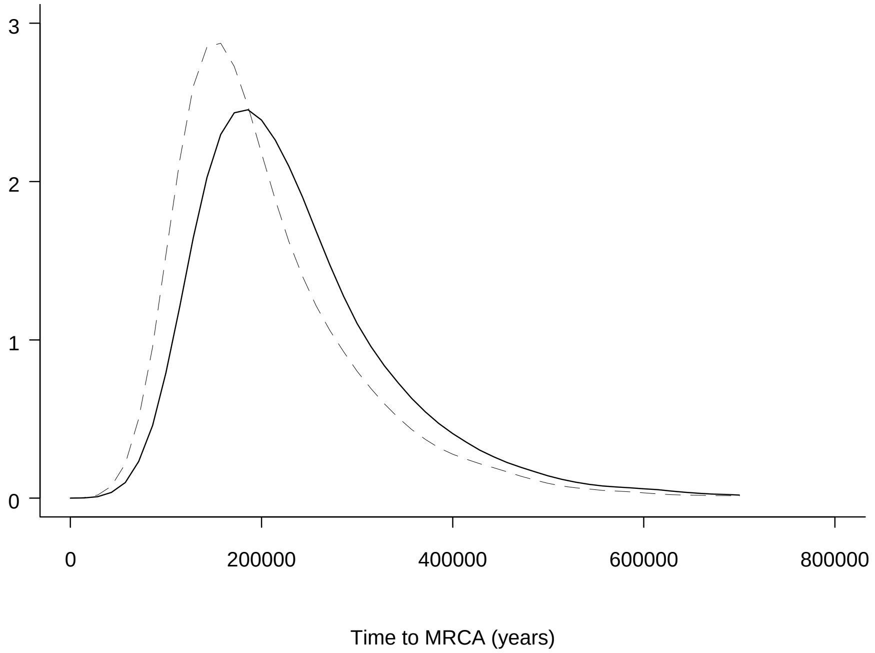

Figure 5.4. Density functions for the predata (solid lines) and postdata distribution of T_{200}. x-axis is in years.

Two points from this analysis should be emphasized. First, the postdata TMRCA of the *sample* can be a serious underestimate of the corresponding time for the *population*. Second, the extra information contained in this small sample does rather little to refine our predata assessment of TMRCA for the population; as can be seen in Figure 5.4, the two densities are very similar.

Finally, what is the effect of the population expansion? For a constant-size model with $N = 4900$ and the same value of μ, the I95 of the postdata distribution of T_{200} is (59,000–410,000) years. The population expansion has shifted this interval to the right by about 54,000 years, essentially the time to the start of the constant phase. The intuition behind this is clear. By the beginning of the expansion phase at time V, a sample of moderate size m is likely to have almost m distinct ancestors. The time to the MRCA is therefore approximately V plus the corresponding time in a population of constant size.

5.8 DISCUSSION

This chapter gives a feel for one approach to an intriguing problem in the historical sciences: estimation of the time to the most recent common ancestor of a population of individuals given DNA-sequence data on a sample from the population. The approach described here, based on a stochastic model for the ancestral relationships between individuals, is intended to be illustrative of the field.

We based our inferences on a summary statistic (the number of segregating sites) of the sample of DNA sequences. Related methods for the full data (once more under the infinitely-many-sites mutation model) have also been developed. The theory of the reduced genealogical trees that represent the complete sequence information appears in Griffiths (1989) and Griffiths and Tavaré (1995). Computer-intensive inference methods are described in Griffiths and Tavaré (1994a).

There are many open problems that remain to be solved. Careful analysis of samples of molecular sequences should take account of the role of demography: nonrandom mating, population subdivision, and fluctuations in population size. With the availability of more data, more refined mutation models could also be exploited. The assumption of random sampling is implicit in most analyses, but nonrandom samples are more likely the rule. Finally, we have assumed that mutation rates and population-size fluctuations are known. Methods that allow for variability in these parameters, and further discussion of many related issues, appear in Donnelly et al. (1996) and Tavaré et al. (1996).

ACKNOWLEDGMENTS S.T. was supported in part by NSF grant BIR 95-04393. The article was completed while the author was Gastprofessor in Applied Mathematics at the University of Zürich.

REFERENCES

Ayala, F. J. 1995. Association Affairs: The myth of Eve: Molecular biology and human origins. *Science*, **270**, 1930–1936. {*81, 95*}

Cann, R., Stoneking, M., and Wilson, A. C. 1987. Mitochondrial DNA and human evolution. *Nature*, **325**, 31–36. {*81, 95*}

Donnelly, P. and Tavaré, S. 1995. Coalescents and genealogical structure under neutrality. *Annual Review of Genetics*, **29**, 401–421. {*86, 87, 95*}

Donnelly, P., Tavaré, S., Balding, D. J., and Griffiths, R. C. 1996. On the time since Adam. *Science*, **272**, 1357–1359. {*95*}

Dorit, R. L., Akashi, H., and Gilbert, W. 1995. Absence of polymorphism at the ZFY locus on the human Y chromosome. *Science*, **268**, 1183–1185. {*81, 95*}

Griffiths, R. C. 1979. Exact sampling distributions from the infinite neutral alleles model. *Advances in Applied Probability*, **11**, 326–354. {*85, 95*}

Griffiths, R. C. 1980. Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoretical Population Biology*, **17**, 37–50. {*81, 95*}

Griffiths, R. C. 1989. Genealogical-tree probabilities in the infinitely-many-sites model. *Journal of Mathematical Biology*, **27**, 667–68. {*95*}

Griffiths, R. C. and Tavaré, S. 1994a. Ancestral inference in population genetics. *Statistical Science*, **9**, 307–319. {*95*}

Griffiths, R. C. and Tavaré, S. 1994b. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London Series B, Biological sciences*, **344**, 403–410. {*87, 95*}

Griffiths, R. C. and Tavaré, S. 1995. Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Mathematical Biosciences*, **127**, 77–98. {*95*}

Hammer, M. F. 1995. A recent common ancestry for human Y chromosomes. *Nature*, **378**, 376–378. {*81, 92, 95*}

Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, **23**, 183–201. {*81, 95*}

Hudson, R. R. 1991. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**, 1–44. {*86, 95*}

Hudson, R. R. 1992. The how and why of generating gene genealogies. *Pages 23–36 of:* Takahata, Naoyuki and Clark, Andrew G. (eds), *Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology*. Sunderland, MA: Sinauer Associates, Inc. {*86, 96*}

Jobling, M. A. and Tyler-Smith, C. 1995. Fathers and sons: the Y chromosome and human evolution. *Trends in Genetics*, **11**, 449–456. {*81, 96*}

Kingman, J. F. C. 1982a. Exchangeability and the evolution of large populations. *Pages 97–112 of:* Koch, G. and Spizzichino, F. (eds), *Exchangeability in probability and statistics: proceedings of the International Conference on Exchangeability in Probability and Statistics, Rome, 6th–9th April, 1981, in honour of Professor Bruno de Finetti*. Amsterdam, The Netherlands: North-Holland Publishing Co. {*81, 84, 96*}

Kingman, J. F. C. 1982b. On the genealogy of large populations. *Journal of Applied Probability*, **19A**, 27–43. {*84, 87, 96*}

Ripley, Brian D. 1987. *Stochastic Simulation*. Wiley series in probability and mathematical statistics. New York, NY: John Wiley and Sons. ISBN 0-471-81884-4. Pages xi + 237. {*85, 91, 96*}

Saunders, I. W., Tavaré, S., and Watterson, G. A. 1984. On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability*, **16**, 471–491. {*96*}

Tajima, F. 1983. Evolutionary relationships of DNA sequences in finite population. *Genetics*, **105**, 437–460. {*81, 96*}

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. 1996. Inferring coalescence times from DNA sequence data. *Genetics*. Submitted. {*92, 95, 96*}

Templeton, A. R. 1993. The "Eve" hypothesis: a genetic critique and reanalysis. *American Anthropologist*, **95**, 51–72. {*81, 96*}

Wallace, D. C. 1995. Mitochondrial DNA variation in human evolution, degenerative disease, and aging. *American Journal of Human Genetics*, **57**, 201–223. {*81, 96*}

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276. {*90, 96*}

Watterson, G. A. 1982. Mutant substitutions at linked nucleotide sites. *Advances in Applied Probability*, **14**, 206–224. {*87, 96*}

Whitfield, L. S., Sulston, J. E., and Goodfellow, P. N. 1995. Sequence variation of the human Y chromosome. *Nature*, **378**, 379–380. {*81, 92, 96*}

Wills, C. 1995. When did Eve live? An evolutionary detective story. *Evolution*, **49**, 593–607. {*81, 96*}

PART II: CELL BIOLOGY

Mark A. Lewis

Organisms and their constituent cells have an immense number of tasks to perform, from information processing to growth and cell division to signaling to maintaining a constant internal environment in the face of external variation (homeostasis). While the disciplines of biochemistry, biophysics, genetics, and molecular biology are central to any understanding of how these tasks are performed, biological experience cannot always predict the consequences of complex physiological interactions. In this realm of unpredictability, mathematical of models can have a crucial role in our understanding of physiology. Mathematics provides a quantitative means with which to formulate precise hypotheses about physiology. The purpose of a model is not to "simulate" the biology, but to judiciously simplify the level of detail to where only the most important elements controlling the physiology remain. Analyses of the model then serve to show which hypotheses are consistent with the more complex experimental detail.

The chapters in this section discuss mechanisms governing physiological phenomena. Subjects include signal transduction via control of calcium and second-messenger dynamics (Chapter 6 by H. G. Othmer), control of cellular replication (Chapter 7 by J. J. Tyson, K. Chen and B. Novak), periodic diseases of the blood (Chapter 8 by M. C. Mackey), oscillatory responses of pupil nervous system to stimuli (Chapter 9 by J. Milton and J. Foss), and electrical bursting behaviors of cells (Chapter 10 by A. Sherman). Although each chapter deals with distinct phenomenon, mathematics allow us to look beneath the specific details to underlying themes that pertain to many of the chapters.

One theme common to each section is biological oscillations, whether in the intracellular calcium levels (H. G. Othmer), the control of cell cycling (J. J. Tyson, K. Chen, and B. Novak), blood counts in diseased individuals (M. C. Mackey), the diameter of the pupil responding to a light source (J. Milton and J. Foss) or the onset of active spiking interspersed with silent states in electrical behavior of cells (A. Sherman). Mechanisms governing each of the above oscillatory systems clearly depend upon specific biological milieu. The mathematics used in their study, however, have some common origins. For example, it may interest the reader to note that Mackey's discussion of periodic diseases of the blood (Chapter 8) and Milton and Foss' discussion of the pupil-light reflex (Chapter 9) use similar nonlinear delay differential equation (DDE) models. Although the processes differ in the biological specifics, oscillations arise in both cases from negative-feedback loops with delays, making DDEs suitable for their study. Both show that the results of long delays can be dramatic, leading not only to oscillations, but also to a variety of very complex temporal patterns. These patterns are not simply mathematical niceties; they are observed biologically either under disease conditions (Mackey) or under laboratory conditions (Milton and Foss).

The theme of excitable dynamics arises in Othmer's discussion of calcium dynamics (Chapter 6), Tyson et al.'s discussion of cellular control (Chapter 7) and Sherman's discussion of electrophysiology (Chapter 10). Models in each of these chapters possess a nonlinear threshold, so that small (subthreshold) stimuli have little effect, but larger (suprathreshold) stimuli cause a dramatic, large-scale response. The classic example of an excitable system arises in the firing of a nerve: the stimulus to a nerve must exceed a threshold before the nerve fires. Mathematical details of this behavior were pioneered in the early 1950s by Hodgkin and Huxley, with an ODE model for the nerve. This elegant work had a profound impact on electrophysiology that is still evident today (see Chapter 10 by A. Sherman) and earned Hodgkin and Huxley a Nobel prize in physiology. Many models in this section more closely resemble later simplifications of Hodgkin and Huxley's (1952) work by FitzHugh (1961) and Nagumo (1962), who showed how the essence of the nerve could be described by a pair of coupled ODEs with excitable dynamics. The themes of oscillations and excitability are connected. It turns out that small modifications of excitable dynamics can give rise to oscillatory dynamics. In the modified excitable system, for example, the nerve never returns completely to a rest state, but fires repeatedly. This kind of oscillation mechanism, exemplified in Chapters 6, 7 and 10, differs from the delayed-negative-feedback loop in Chapters 8 and 9.

Although this section examines a diverse body of work, readers may find other themes that unify. They are also encouraged to ask questions of each chapter:

- **How is the mathematics useful?** Are precise hypotheses formulated? Does analysis of the model yield specific predictions? If so, how do these compare with experiment? How would one proceed without mathematics?
- **Where do the equations come from?** Each chapter derives equations using specific modeling principles. How do the authors make the transition from scientific hypotheses to mathematical formulae?
- **Does the model work?** A key element of each chapter is the author's ability to tailor a mathematical model to focus on the specific question at hand. What is missing from each model? How do we know whether it is important or not?

REFERENCES

FitzHugh, R. 1961. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical Journal*, **1**, 445–466. {*98, 299, 305*}

Hodgkin, A. L. and Huxley, A. F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, **117**, 500–544. {*98*}

Nagumo, J. S., Arimoto, S., and Yoshizawa, S. 1962. An active pulse transmission line simulating nerve axon. *Proceedings of the Institute of Radio Engineers*, **50**, 2061–2071. {*98, 299, 307*}