Calibrating the Clock:
Using stochastic processes to measure the Rate of Evolution

Simon Tavaré
University of Southern California

Chapter 5 (pp 114-152) in
Calculating the Secrets of Life
Eds: E.S. Lander, M.S. Waterman. National Academy Press, 1995.

DNA sequences record the history of life. Although DNA replication is remarkably accurate, mutations do occur at a small but non-negligible rate, with the result that an individual's descendants begin to diverge in DNA sequence over time. By examining DNA sequences among different species or among different individuals within a single species, it is possible to reconstruct aspects of their evolutionary history. Such studies have been pursued with special interest in the human, where an unusual DNA sequence called the mitochondrial genome has been used to trace human migrations and human evolution. The author shows how mathematical tools from the theory of stochastic processes assist in calibrating the molecular clock inherent in DNA sequences.

1 Introduction

DNA sequences record the history of life: While DNA sequences are transmitted from parent to child with remarkable fidelity, mutations occur at a small but non-negligible rate with the result that an individual's descendants begin to diverge in DNA sequence over time. Some mutations are deleterious and are eliminated by natural selection, but many are thought to be selectively neutral and thus accumulate at a roughly steady rate – providing a molecular clock for measuring the time since two species or two individuals within a species shared a common ancestor. In this manner, it is possible to reconstruct an evolutionary tree and even estimate the times of key separation events.

Different biological sequences within an organism may obey different clocks. The amino acid sequence of a protein encoded by a gene changes more slowly than the DNA sequence of the underlying gene because many amino acid changes may be selectively disadvantageous (because they disrupt function). On the other hand, a significant proportion of DNA changes may be selectively neutral because they create a synonymous codon (that is, one that specifies the same amino acid). Similarly, DNA regions within genes change at a slower rate than the DNA sequences located between genes. Accordingly, evolutionary studies of distant species are often carried out by examining amino acids sequences of proteins, while evolutionary comparisons among more closely related species are better done by examining DNA sequences within or between genes.

To study evolution within a single species such as the human, it is often useful to study DNA sequences that change at especially rapid rates. The mitochondrial genome provides an ideal substrate for such studies. The mitochondrion is an organelle found in the cytoplasm of eukaryotic cells, whose primary role is to generate high-energy compounds that the cell uses to drive chemical reactions. Although the mitochondria use many proteins that are

encoded by genes in the cell nucleus, it has its own small circular chromosome that encodes a few dozen genes essential for its function.

In the human, the mitochondrial genome consists of 16,569 base pairs whose DNA sequence has been completely determined (Anderson et al. 1981). Human mitochondria are inherited only from the mother, so their genealogy is considerably simpler to follow than for genes encoded in the nucleus (which are inherited from both parents and are subject to recombination between the two homologous copies in the cell). Conveniently for evolutionary studies, mitochondrial DNA has an increased rate of nucleotide substitution compared to nuclear genes, owing to the presumed absence of certain DNA repair mechanisms. Moreover, the mitochondrial genome contains certain regions that are particularly tolerant of mutation (i.e., appear to be subject to little selective pressure; Avise (1986)) and thus show a great deal of variation. In all, the mitochondrial genome may be evolving ten times faster than the nuclear genome.

For these reasons, molecular population geneticists have carried out many studies of the DNA sequences of mitochondrial variable regions in many human populations (DiRienzo and Wilson (1991), Horai and Hayasaka (1990), Vigilant et al. (1989; 1991) and Ward et al. (1991)). Studies of mitochondrial sequences of different Native American tribes strongly suggest that there were multiple waves of colonization of North America by migrant groups from Asia, and even allow one to estimate the date of these events (Schurr et al. 1990; Ward et al. 1991). Assuming a constant evolutionary rate, the pattern of mutations between diverse human groups has been used to argue (Cann, Stoneking and Wilson, 1987) that the mitochondria of all living humans descended from a mother that lived in Africa some 200,000 years ago – the so-called Eve hypothesis. Although the precise details of the hypothesis are disputed (Maddison (1991), Nei (1992), Templeton (1992)), the general power of the methodology is well accepted. (As an aside, the reader should note that the existence of a common ancestor – Eve, so to speak – is a mathemat-

ical necessity in any branching process that satisfies very weak conditions. The biological controversies pertain to when and where Eve lived.)

Each of these applications requires a knowledge of the rate at which mutations occur in a mtDNA sequence. Estimates of this rate have been obtained by comparing a single DNA sequence from each of several species whose times of divergence are presumed known. Divergence is calculated from the number of nucleotide differences between species (using methods that correct for the possibility of multiple mutations at a site) and rate estimates are obtained by dividing the rate of sequence divergence by the divergence time. For data taken from multiple individuals in a single population, one requires a model that takes account of the population genetic aspect of the sampling: individuals in the sample are correlated by their common ancestry. In this chapter, we describe the underlying stochastic structure of this ancestry, and use the results to estimate substitution rates.

We have chosen to focus on rate estimation to give the chapter a single theme. We will not be interested *per se* in statistical aspects of tests for selective neutrality of DNA differences; rather we assume neutrality for the data sets discussed as examples. The techniques described here should be regarded as illustrative of the theoretical and practical problems that arise in sequence analysis of samples from closely related individuals. The emphasis is on exploratory methods that might be used to summarize the structure of such samples.

1.1 Overview

To illustrate the methods, we use a set of North American Indian mitochondrial sequences described in Ward et al. (1991). These authors sequenced the first 360 base pairs of the mitochondrial control region for a sample of 63 Nuu-Chah-Nulth (Nootka) Indians from Vancouver Island. The sample comprises individuals who were maternally unrelated for four generations,

chosen from 13 of the 14 tribal bands. As a consequence the sample deviates from a truly random sample, although it will be treated as such for the purposes of this chapter. An important parameter in the analysis is the effective population size of the group. This is approximated by the number of reproducing females, giving a value of about 600 for the long-term effective population size N.

The most common DNA changes seen in mitochondria are transitions (changes from one pyrimidine base to the other or one purine base to the other – i.e., $C \leftrightarrow T$ or $A \leftrightarrow G$) rather than transversions (change from a pyrimidine to a purine or vice versa) Indeed, the sequenced region shows no transversions, so that each site in the sequences has one of just two possible nucleotides. We focus on the pyrimidine (C or T) sites in the region. There are 201 such sites, in which 21 variable (or segregating) sites define 24 distinct sequences (called alleles or lineages). The details of the data, including the allele frequencies, are given in Table 1.

The parameter of particular interest here is θ, the population geneticist's stock in trade. θ is a measure of the mutation rate in the region and it figures in many important theoretical formulas in population genetics. For mitochondrial data, it is defined by

$$\theta = 2Nu$$

where N is the effective population size referred to earlier, and u is the mutation rate per gene per generation. Once θ is estimated, we can estimate u if N is known or N if u is known. In what follows, we estimate the compound parameter θ rather than its components.

In Section 2 we begin by outlining the structure of the *coalescent*, a robust description of the genealogy of samples taken from large populations. The effects of mutation are superimposed on this genealogy in several ways. The classical case, which records the allelic partition of the sample, leads to the sampling theory of the *infinitely-many-alleles* model initiated by Ewens

Table 1: Nucleotide Position in Control Region

Position	69	88	91	124	149	162	166	194	200	219	233	247	255	267	271	275	301	302	304	319	339	allele freqs.
Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
ID ref	T	C	C	C	T	C	T	T	C	C	C	C	C	C	C	T	T	T	C	T	T	
1	.	.	.	.	.	.	C	.	T	.	.	.	T	.	.	.	.	.	.	.	.	3
2&3	.	.	.	.	.	.	.	.	T	.	.	.	T	.	.	.	.	.	.	.	.	3
4	.	.	.	.	.	.	.	.	T	.	.	.	T	.	.	.	.	.	.	.	C	1
5	.	T	.	.	.	T	.	.	T	.	.	.	.	T	.	.	.	.	.	.	C	2
6	.	T	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	C	2
7	C	T	.	.	.	.	.	.	T	.	.	T	.	.	.	.	.	.	.	.	C	1
8,10,& 11	.	T	.	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	C	8
9	C	T	.	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	C	2
12 & 13	.	T	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	C	10
14	.	T	.	.	.	.	.	.	T	.	.	.	T	T	.	.	.	.	.	.	C	1
15	.	T	.	.	.	.	.	.	T	.	.	.	T	T	.	.	C	.	.	.	C	2
16	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	.	.	.	T	.	C	1
17	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	C	.	.	.	C	1
18	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	.	C	.	.	C	2
19	.	.	T	.	.	.	.	.	T	.	.	.	.	.	T	.	.	C	.	.	C	1
20	.	.	.	.	C	.	.	.	T	.	.	.	.	.	.	.	.	C	.	.	C	3
21	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	C	.	.	C	3
22	C	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	C	.	.	.	3
23	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	C	.	C	.	.	.	1
24	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	C	.	C	T	.	.	7
25	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	C	.	C	T	C	.	3
26	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	C	T	C	.	1
27	.	.	.	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	1
28	.	.	.	.	.	.	C	C	.	.	T	.	.	.	.	.	.	.	.	.	.	1

Mitochondrial data from Ward et al (1991, Figure 1). Variable pyrimidine positions in the control region. Position 69 corresponds to position 16,092 in the human reference sequence published by Anderson et al (1981). The ID numbers correspond to those given in Ward et al., Figure 1.

(1972). The Ewens sampling formula is then described, followed by a brief digression into the simulation structure of mutations in the coalescent, both in top-down and bottom-up form. Next, the *infinitely-many-sites* model is introduced as a simple description of the detailed structure of the segregating sites in the sample. Finally, we return to classical population genetics theory, albeit from a coalescent point of view, to discuss the structure of K-allele models. This in turn develops into the study of the *finitely-many-sites* models, which play a crucial role in the study of sequence variability when back substitutions are prevalent.

In Section 3 we digress to present a mathematical vignette in the area of random combinatorial structures. The Ewens sampling formula was derived as a means to analyze allozyme frequency data that became prevalent in the late 1960s. Current population genetic data is more sequence oriented, and requires more detailed models for its analysis. Nonetheless, the combinatorial structure of the Ewens sampling formula has recently emerged as a useful approximation to the component counting process of a wide range of combinatorial objects, among them random permutations, random mapping functions, and factorization of polynomials over a finite field. We show how a result of central importance in the development of statistical inference for molecular data has a new lease on life in an area of discrete mathematics.

Finally, Section 4 briefly discusses some of the outstanding problems in the area, with particular emphasis on likelihood methods for coalescent processes. Some aspects of the mathematical theory, for example measure-valued diffusions, are also mentioned, together with applications to other, more complicated, genetic mechanisms.

2 The coalescent and mutation

The genealogy of a sample of n genes (i.e., stretches of DNA sequence) drawn at random from a large population of approximately constant size

may be described in terms of independent exponential random variables $T_n, T_{n-1}, \ldots, T_2$ as follows. The time T_n during which the sample has n distinct ancestors has an exponential distribution with parameter $\binom{n}{2}$, at which time two of the lines are chosen at random to coalesce, giving the sample $n-1$ distinct ancestors. The time T_{n-1} during which the sample has $n-1$ such ancestors is exponentially distributed with parameter $\binom{n-1}{2}$, at which point two more ancestors are chosen at random to coalesce. This process of coalescing continues until the sample has two distinct ancestors. From that point it takes an exponential amount of time T_2, with parameter $\binom{2}{2} = 1$, to trace back to the sample's common ancestor. For our purposes, the time scale is measured in units of N generations, where N is the (effective) size of the population from which the sample was drawn. This structure, made explicit by Kingman (1982a,b), arises as an approximation for large N to many models of reproduction, among them the Wright–Fisher and Moran models. A sample path of a coalescent with $n = 5$ is shown in Figure 1.

Insert Figure 1 about here

From the description of the genealogy, it is clear that the time τ_n back to the common ancestor has mean

$$\mathbb{E}\tau_n = \sum_{j=2}^{n} \mathbb{E}T_j = \sum_{j=2}^{n} \frac{2}{j(j-1)} = 2\left(1 - \frac{1}{n}\right),$$

or approximately $2N$ generations for large sample sizes. Further aspects of the structure of the ancestral process may be found in Tavaré (1984). Rather than focus further on such issues, we describe how the genealogy may be used to study the genetic composition of the sample.

To this end, assume that in the population from which the sample was drawn there is a probability u that any gene mutates in a given generation,

mutation acting independently for different individuals. In looking back r generations through the ancestry of a randomly chosen gene, the number of mutations along that line is a binomial random variable with parameters r and u. If we measure time in units of N generations, so that $r = \lfloor Nt \rfloor$, and assume that $2Nu \to \theta$ as $N \to \infty$, then the Poisson approximation to the binomial distribution shows that the number of mutations in time t has in the limit a Poisson distribution with mean $\theta t/2$. This argument can be extended to show that the mutations that arise on different branches of the coalescent tree follow independent Poisson processes, each of rate $\theta/2$. For example, the total number of mutations μ_n that occur in the history of our sample back to its common ancestor has a mixed Poisson distribution – given $T_n, \ldots, T_2$, μ_n has a Poisson distribution with mean $\frac{1}{2}\theta \sum_{j=2}^{n} jT_j$. The mean and variance of the number of mutations are given by Watterson (1975):

$$\mathbb{E}\mu_n = \frac{\theta}{2} \sum_{j=2}^{n} j\mathbb{E}T_j = \theta \sum_{j=1}^{n-1} \frac{1}{j}, \tag{1}$$

and

$$\operatorname{Var} \mu_n = \theta \sum_{j=1}^{n-1} \frac{1}{j} + \theta^2 \sum_{j=1}^{n-1} \frac{1}{j^2}. \tag{2}$$

We are now in a position to describe the effect that mutation has on the individuals in the sample.

2.1 The Ewens Sampling Formula

Motivated by the realization that mutations in DNA sequences could lead to an essentially infinite number of alleles at the given locus, Kimura and Crow (1964) advocated modeling the effects of mutation as an *infinitely-many-alleles model*. In this process, a gene inherits the type of its ancestor if no mutation occurs, and inherits a type not currently (or previously) existing in the population if a mutation does occur. In such a process the genes in the sample are thought of as *unlabeled*, so that the experimenter knows whether

two genes are different, but records nothing further about the identity of alleles. In this case the natural statistic to record about the sample is its *configuration* $\boldsymbol{C}_n \equiv (C_1, C_2 \ldots, C_n)$, where

$$C_j = \text{ number of alleles represented } j \text{ times.}$$

Of course, $C_1 + 2C_2 + \cdots + nC_n = n$, and the number of alleles in the sample is

$$K_n \equiv C_1 + C_2 + \cdots + C_n. \tag{3}$$

The sampling distribution of $\boldsymbol{C}_n$ was found by Ewens (1972):

$$\mathbb{P}(\boldsymbol{C}_n = \boldsymbol{a}) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!}, \tag{4}$$

for $\boldsymbol{a} = (a_1, \ldots, a_n)$ satisfying $a_j \geq 0$ for $j = 1, \ldots, n$ and $\sum_{j=1}^{n} j a_j = n$, and where

$$\theta_{(n)} \equiv \theta(\theta + 1) \cdots (\theta + n - 1).$$

From (4) it follows that

$$\mathbb{P}(K_n = k) = \frac{|S_n^k| \theta^k}{\theta_{(n)}}, \ k = 1, \ldots, n \tag{5}$$

and

$$\mathbb{E}K_n = \sum_{j=0}^{n-1} \frac{\theta}{\theta + j}, \tag{6}$$

S_n^k being the Stirling number of the first kind. From (5) and (4) it follows that K_n is sufficient for θ, so that the information in the sample relevant for estimating θ is contained just in K_n. This allows us (Ewens, 1972, 1979) to calculate the maximum likelihood (and moment) estimator of θ as the solution $\hat{\theta}$ of the equation

$$k = \sum_{j=0}^{n-1} \frac{\hat{\theta}}{\hat{\theta} + j} \tag{7}$$

where k is the number of alleles observed in the sample. In large samples, the estimator $\hat{\theta}$ has variance given approximately by

$$\mathrm{Var}(\hat{\theta}) \approx \theta \left(\sum_{k=2}^{n} \frac{k-1}{(\theta + k - 1)^2} \right)^{-1} \tag{8}$$

For the pyrimidine sequence data described in Section 1.1, there are $k = 24$ alleles. Solving equation (7) for $\hat{\theta}$ gives $\hat{\theta} = 10.62$, with a variance of 9.89. An approximate 95% confidence interval for θ is therefore 10.62 ± 6.29. This example serves to underline the variability inherent in estimating θ from this model. The pyrimidine region comprises 201 sites, so that the *per site* substitution rate is estimated to be 0.053 ± 0.031.

The goodness of fit of the model to the data may be assessed by using the sufficiency of K_n for θ: given K_n, the conditional distribution of the allele frequencies is independent of θ. Ewens (1972, 1979) gives further details on this point. To describe alternative goodness-of-fit methods, we return briefly to the probabilistic structure of mutation in the colaescent.

2.2 Forwards and backwards in the tree

Hudson (1991) decribes many situations in which simulation of genealogical trees is useful. In its simplest form, the idea is to construct (a simulation of) a coalescent tree, with times and branching order, and then superimpose the effects of mutation on this tree using the Poisson nature of the mutation process. In this section we make use of two equivalent descriptions of the effects of mutation in the coalescent tree in which the mutations and coalescent events evolve simultaneously.

2.2.1 Top-down

The first of these methods is a very useful 'top-down' scheme exploited by Ethier and Griffiths (1987) in the context of the infinitely-many-sites model.

We start at the common ancestor of sample, and think of the genetic process running down to the sample. Just after the first split, we have a sample of two individuals, each of identical genetic type. Attach to each individual a pair of independent exponential alarm clocks, one of rate $\theta/2$, the second of rate $1/2$, and suppose the clocks are independent for different individuals. The θ-clocks will determine mutations, the other clocks split times. Now watch the clocks until the first one rings: if a θ-clock rings, a mutation occurs in that gene, whereas if one of the other clock rings, a split occurs in which that gene is copied, now making a sample of three individuals. Using the standard 'competing exponentials' argument, the probability that a mutation occurs first is

$$\frac{\theta/2 + \theta/2}{\theta/2 + \theta/2 + 1/2 + 1/2} = \frac{\theta}{\theta + 1},$$

whereas a split occurs first with probability $1/(\theta + 1)$. Furthermore, given that a mutation occurs first, the gene in which it occurs is chosen uniformly and at random, and given that a split occurs first, the gene that is copied is chosen uniformly and at random.

Once an event occurs, the process repeats itself in a similar way. Suppose, then, that there are currently m genes in the sample. Attach independent mutation clocks of rate $\theta/2$ and independent split clocks of rate $(m-1)/2$ to each of the m genes, and wait for one to ring. The probability that a mutation clock rings first is $\theta/(\theta + m - 1)$, and, given that a mutation occurs first, the gene that mutates is chosen uniformly and at random. Similarly, the probability that a split occurs first is $(m-1)/(\theta + m - 1)$, with the splitting gene being chosen at random from the m possibilities.

The only wrinkle left is to describe the rule that tells us when to stop generating splits or mutations. In order to have the right distribution for the numbers of mutations when the sample has n ancestors, we must run until the first split after n, discard the last observation and then stop.

This simple scheme can be used effectively to simulate observations from

extremely complex mutation mechanisms using only Bernoulli random variables, and a way of generating and storing the effects of each of the mutations. Some examples are given in the following sections.

2.2.2 Bottom up

The second scheme, which proves very useful for deriving recurrence relations for the distribution of allele configurations, is the 'bottom-up' method. In this case, the idea is to use the exponenetial alarm clocks from the bottom of the tree (that is, beginning at the sample) and run up to the common ancestor at the top. If we look up from the sample of size n toward the root, the probability that we will encounter a mutation before a coalescence is $\theta/(\theta+n-1)$, and the probability that a coalescence occurs first is $(n-1)/(\theta+n-1)$. The probability distribution of the configuration at the tips may then be related to the distribution of the configuration at the mutation or coalescence time.

To illustrate how this works, consider the infinitely-many-alleles mutation structure. Suppose that the current configuration consists of counts $\boldsymbol{a} = (a_1, \ldots, a_n)$ with $a_n = 0$, and let $P_n(\boldsymbol{a})$ denote the probability of this configuration. If the first event in the past is a coalescence, then the configuration of $n-1$ genes must have been

$$(a_1, \ldots, a_j + 1, a_{j+1} - 1, \ldots, a_{n-1})$$

for some $j = 1, 2, \ldots, n-2$, and a gene in class j must be chosen to have an offspring. Since this last event has probability $\frac{j(a_j+1)}{n-1}$, the contribution to $P_n(\boldsymbol{a})$ from such terms is

$$\frac{n-1}{\theta+n-1}\left[\sum_{j=1}^{n-2}\frac{j(a_j+1)}{n-1}P_{n-1}(a_1, \ldots, a_j+1, a_{j+1}-1, \ldots, a_{n-1})\right]. \qquad (9)$$

If on the other hand the first event in the past was a mutation, then the

configuration must have been either

$$(a_1 - 1, a_2, \ldots, a_{j-1} - 1, a_j + 1, \ldots, a_{n-1}, 0)$$

and the mutation occurred to a gene in a j class, $j = 3, \ldots, n-1$ (probability $j(a_j + 1)/n$), or

$$(a_1 - 2, a_2 + 1, a_3, \ldots, a_{n-1}, 0)$$

and the mutation occurred to a gene in the 2 class (probability $2(a_2 + 1)/n$), or

$$(a_1, \ldots, a_{n-1}, 0)$$

and the mutation occurred to a singleton gene (probability a_1/n). Finally, the configuration could have been

$$(a_1 - 1, a_2, \ldots, a_{n-2}, a_{n-1} - 1, 1)$$

and the mutation occurred in the n class (probability 1). Combining all these possibilities, and adding the term in (9), gives

$$\begin{aligned} P_n(\boldsymbol{a}) \;=\; & \frac{\theta}{\theta + n - 1}\Bigg[\quad P_n(a_1 - 1, a_2, \ldots, a_{n-2}, a_{n-1} - 1, 1) \\ & + \frac{a_1}{n} P_n(a_1, \ldots, a_{n-1}, 0) \\ & + \frac{2(a_2 + 1)}{n} P_n(a_1 - 2, a_2 + 1, a_3, \ldots, a_{n-1}, 0) \\ & + \sum_{j=3}^{n-1} \frac{j(a_j + 1)}{n} P_n(a_1 - 1, a_2, \ldots, a_{j-1} - 1, a_j + 1, \ldots, a_{n-1}, 0)\Bigg] \\ & + \frac{n-1}{\theta + n - 1}\left[\sum_{j=1}^{n-2} \frac{j(a_j + 1)}{n-1} P_{n-1}(a_1, \ldots, a_j + 1, a_{j+1} - 1, \ldots, a_{n-1})\right]. \end{aligned} \tag{10}$$

The only case not covered by equation (10) is the one in which $\boldsymbol{a} = (0, \ldots, 0, 1)$. In this case the previous event had to be a coalescence, and so

$$P_n(0, \ldots, 0, 1) = \frac{n-1}{\theta + n - 1} P_{n-1}(0, \ldots, 0, 1). \tag{11}$$

The persistent reader will be able to verify that $P_n(\boldsymbol{a})$ given by the Ewens sampling formula (4) does indeed satisfy equations (10) and (11).

2.3 The infinitely-many-sites model

The infinitely-many-sites model of Kimura (1969) and Watterson (1975) is the simplest description of the evolution of a population of DNA sequences. The sites in the sequences are completely linked, and each mutation that occurs in the ancestral tree of the sample introduces a new *segregating site* into the sample. In this process, each new mutation occurs at a site not previously segregating – new mutations arise just once. It follows that at each segregating site, the sample may be classified as type 0 (ancestral) or type 1 (mutant). Of course, in practice we do not know which is which. The sequences in the sample may now be described by strings of 0s and 1s. If distinct sequences are treated as alleles, then the sampling theory is reduced to that covered by the Ewens sampling formula.

The number S_n of segregating sites is an important summary statistic for the sample. Since each new mutation produces a segregating site, it follows that $S_n = \mu_n$, the number of mutations in the ancestral tree. The mean and variance of S_n are therefore given by (1) and (2) respectively.

The number of segregating sites has been studied extensively for many variants of the infinitely-many-sites process, including the effects of selection and recombination for example. Hudson (1991) gives an accessible summary of this work. When there is no recombination, the fundamental results have been established by Watterson (1975), Ethier and Griffiths (1987) and Griffiths (1989).

Watterson (1975) parlayed the moments of S_n into an unbiased estimator $\tilde{\theta}$ of θ, viz.

$$\tilde{\theta} = \frac{S_n}{\sum_{j=1}^{n-1} \frac{1}{j}}, \tag{12}$$

with variance

$$\operatorname{Var} \tilde{\theta} = \frac{\operatorname{Var} S_n}{\left(\sum_{j=1}^{n-1} \frac{1}{j}\right)^2}.$$

Note that $\tilde{\theta}$ does not depend on knowing which type at a site is ancestral,

and does not make full use of the data. For the pyrimidine data, there are 21 segregating sites, giving an approximate 95% confidence interval for θ of 4.46 ± 3.10. This should be compared to the estimate of 10.62 ± 6.29 obtained from the Ewens sampling formula.

Now think of the data as an $n \times s$ matrix of 0s and 1s, s being the number of segregating sites in the sample. When 0 is known to be ancestral in each site, Griffiths (1987) established that the data are consistent with the infinitely-many-sites model as long as in any set of three rows of the matrix, at most 1 of the patterns

$$\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

occurs. This is equivalent to the pairwise compatibility condition for binary characters established by Estabrook, Johnson and McMorris (1976) and McMorris (1977): two sites are compatible if two or fewer of the patterns $01, 10, 11$ occur. When the ancestral state is unknown an analogous result holds: two sites are compatible if at most three of the patterns $00, 01, 10, 11$ occur.

This translates into a simple test of whether a given set of binary site data is consistent with the infinitely-many-sites model. If in all pairs of columns at most three of the patterns 00, 01, 10, 11 occur, then there is at least one labelling of the sites that is consistent. McMorris (1977) proved that consistent data remain consistent when the most frequent type is taken as ancestral.

In practice, back mutations and recombination make most molecular data inconsistent with this model. However, it is worthwhile to look for maximal subsets of sites which are consistent, as this provides a way to identify regions of the sequence with simple structure. For the pyrimidine data described in Table 1, the maximal consistent set has 14 sites, those in positions 2-8, 11-12,

14-16, 20-21. The remaining 7 sites have some inconsistencies, attributable to back substitutions for example.

Of the $2^{14} = 16,384$ possible relabelings of the consistent set, just 16 are consistent. Each of these labelings is associated with a genealogical tree that describes the relationships between the mutations in the coalescent. The precise definition of the (equivalence class of) trees is given in Ethier and Griffiths (1987), and Griffiths (1989). The tree is equivalent to those built using compatibility methods for binary characters; see Felsenstein (1982, pp. 389-393) for a detailed discussion and references. The nodes in the tree represent the mutations that have generated the segregating sites, and the tips represent the sequences. A convenient algorithm for finding these trees is provided by Griffiths (1987), who also shows (Griffiths, 1989) how the probability of a tree with a given ancestral labeling may be computed under the infinitely-many-sites model. Griffiths' program PTREE can then be used to construct true maximum likelihood estimators of the parameter θ. It can also be used to compare 'likelihoods' of the different ancestral labelings. At the time of writing, there seem to be no useful computational methods for computing likelihoods for data sets of the size described here, in the case where the ancestral labeling is unknown.

Our analysis of the mitochondrial data set has shown that while parts of the region are consistent with a simple evolutionary model, there are sites which are behaving in a more complicated way. In the next section, we describe a finitely-many-sites model that is useful for modelling regions in which back mutations have occurred.

2.4 K-allele models

We turn first to the 'somewhat old-fashioned' K-allele model. In this process, we assume that there are K possible alleles at the locus in question. When a mutation occurs to an allele of type i, there is a probability m_{ij} that the

resulting allele is of type j. To allow for different rates of substitution for different alleles, we can have $m_{ii} > 0$, and we write $M = (m_{ij})$. The effects of mutation along a given line are now modeled by a continuous time Markov chain whose transition matrix $P(t) \equiv (p_{ij}(t))$ gives the probabilities that a gene of type i has been replaced by a descendent gene of type j a time t later. Indeed,

$$P(t) = \exp\left(\frac{\theta t}{2}(M - I)\right),$$

where I is the $K \times K$ identity matrix, so that the generator of the mutation process is

$$Q \equiv (q_{ij}) = \frac{\theta}{2}(M - I). \tag{13}$$

It is worth pointing out that a given Q matrix can be represented in more than one way in the form (13), so that θ, for example, is not identifiable without further assumptions. However, the rates $q_{ij}, j \neq i$ are identifiable. If Q has a stationary distribution $\pi = (\pi_1, \ldots, \pi_K)$ satisfying $\pi Q = 0, \sum_{j=1}^{K} \pi_j = 1$, and if the common ancestor of the sample has distribution π then the distribution of a gene at any point in the tree is also π; the process is then stationary.

From the data analyst's perspective, the sample of n genes can be sorted into a vector $\boldsymbol{N} \equiv (N_1, \ldots, N_K)$ of counts, there being N_j alleles of type j in the sample. Surprisingly, the stationary distribution of $\boldsymbol{N}$ is known explicitly only for the special case

$$q_{ij} = \frac{1}{2}\epsilon_j > 0, j \neq i.$$

This is equivalent to the *independent mutations* case in which $\theta = \epsilon_1 + \cdots + \epsilon_K$, $\pi_i = \epsilon_i/\theta$, and $m_{ij} = \pi_j$ for all i and j. In this case, Wright's Formula (cf. Wright, 1968) can be used to show that

$$\mathbb{P}(\boldsymbol{N} = \boldsymbol{n}) = \binom{\theta + n - 1}{n}^{-1} \prod_{i=1}^{K} \binom{n_i + \theta\pi_i - 1}{n_i}, \tag{14}$$

for $\boldsymbol{n} = (n_1, \ldots, n_K), n_j \geq 0$ for $j = 1, \ldots, K$, and $n_1 + \cdots + n_K = n$.

In the next section, we use this result for the case $K = 2$. If

$$Q = \frac{1}{2}\begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} \tag{15}$$

then equation (14) specializes to

$$g(l) \equiv \mathbb{P}(N_1 = l) = \binom{\alpha+\beta+n-1}{n}^{-1}\binom{l+\beta-1}{l}\binom{n-l+\alpha-1}{n-l}, \tag{16}$$

for $l = 0, \ldots, n$.

Because the sampling formula for general Q is not known explicitly, it is useful to have a way to compute it. Perhaps the simplest is an application of the 'bottom-up' method described in Section 2.2.2. Define $q(\boldsymbol{n}) = \mathbb{P}(\boldsymbol{N} = \boldsymbol{n})$, and set $\boldsymbol{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)$, the 1 occurring in position j. Look up the tree to the first event that occurred. This is either a mutation (with probability $\theta/(\theta+n-1)$) or a coalescence (with probability $(n-1)/(\theta+n-1)$). By considering the configuration of the sample at this event, we see that $q(\boldsymbol{n})$ satisfies the recursion

$$\begin{aligned} q(\boldsymbol{n}) &= \frac{\theta}{\theta+n-1}\left[\sum_i \frac{n_i}{n} m_{ii}\, q(\boldsymbol{n}) + \sum_i \sum_{j\neq i} \frac{n_i+1}{n} m_{ij}\, q(\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j)\right] \\ &\quad + \frac{n-1}{\theta+n-1}\sum_j \frac{n_j-1}{n-1} q(\boldsymbol{n}-\boldsymbol{e}_j), \end{aligned} \tag{17}$$

where $q(\boldsymbol{n}) \equiv 0$ if any $n_i < 0$, and $q(\boldsymbol{e}_i) = \pi_i^*$. The process is stationary if $\pi_i^* = \pi_i$ for all i. We exploit this recursion more fully in Section 4.1

2.5 The finitely-many-sites models

We now have the machinery necessary to describe the finitely-many-sites model for molecular sequence data involving n sequences, each of s sites. The sites are thought of as completely linked, and each site is typically one of either 2 or 4 possibilities. At its grossest level, the finitely-many-sites

model is 'just' a K-allele model in which $K = 2^s$ or 4^s. From an inference point of view, however, there are far too many parameters in such a model, and some simplification is required. The simplest null model of sequence evolution is the case in which mutations still occur at rate $\theta/2$ *per gene*, but when a mutation occurs, a site is chosen at random to mutate and the base at that site changes according to a mutation matrix M. A slightly more general model might allow site j to mutate with probability p_j, once more according to M. For a two-type classification of each site, the first model has 2 parameters to be estimated, and the second $s + 1$. These schemes can be modified to allow for other correlation structures between the sites at the expense of more complicated methods of analysis.

Motivated by our sequence data, we concentrate on the two-state case, and discuss methods for estimating the parameters of the simplest null model. At a single site, the model behaves exactly like a 2-allele process with

$$Q = \frac{\theta}{2s}(M - I),$$

because the *per site* substitution rate is θ/s. This has the structure of (15), with $\alpha = m_{12}\theta/s$ and $\beta = m_{21}\theta/s$. The distribution of sites is exchangeable (since, conditional on the coalescent tree, mutations are laid down independently at each site; this is a simple example of a marked Poisson process argument) and in particular have identical distributions. They are *not* of course independent because of correlations induced by the common ancestry in the coalescent. However, some simple properties of the sequences are easy to calculate. In particular, the number S_n of segregating sites has mean

$$\mathbb{E}S_n = s\mathbb{P}(\text{ site is segregating}) = s(1 - g(0) - g(n)), \tag{18}$$

where $g(\cdot)$ is given by (16).

The equation (18) provides a simple heuristic method for estimating the parameters of the process. First, the equilibrium base frequencies $\pi_1 =$

$\beta/(\alpha+\beta)$, and $\pi_2 = \alpha/(\alpha+\beta)$ are estimated from the sequence data. This done, the expected fraction of sites that are *not* segregating is, from (16) and (18)

$$s^{-1}\,\mathbb{E}(s - S_n) = \frac{\Gamma(\theta_s)}{\Gamma(\theta_s + n)}\left[\frac{\Gamma(\theta_s\pi_1 + n)}{\Gamma(\theta_s\pi_1)} + \frac{\Gamma(\theta_s\pi_2 + n)}{\Gamma(\theta_s\pi_2)}\right], \qquad (19)$$

where $\theta_s = \theta/s$ is the per-site substitution rate. For the pyrimidine mtDNA data, the observed fraction of non-segregating sites is $180/201 = 0.896$, and the observed fractions of C (labeled 1) and T (labeled 2) bases are $\pi_1 = 0.604$ and $\pi_2 = 0.396$ respectively. Substituting these into (19) and solving for θ_s gives the moment estimator $\tilde{\theta}_s = 0.050$. This translates into an estimate of $\alpha = 2q_{12} = 0.050 \times 0.40 = 0.02$, and an estimate of $\beta = 2q_{21} = 0.03$. This estimate of θ_s should be contrasted with the per-base rate of 0.053 estimated from the Ewens sampling formula, and 0.022 from the infinitely-many-sites model. The variance of the moment estimator is hard to compute explicitly, although the top-down simulation method for the coalescent could be used to simulate the process, and therefore to construct empirical estimates of the variance.

A more detailed approach to rate estimation in the finite sites model is described by Lundstrom, Tavaré and Ward (1992a). The method is based once more on the exchangeability of the distribution of base frequencies between sites with the same mutation structure. Returning to the case in which there are K possible labellings at each site, define $V_{n,\boldsymbol{x}} \equiv V_{n,(x_1,x_2,\ldots,x_K)}$ to be the fraction of sites in which x_j individuals in the sample have nucleotide j at that site, for $1 \le j \le K$. The mean of $V_{n,\boldsymbol{x}}$ is given by

$$\mathbb{E}V_{n,\boldsymbol{x}} = \mathbb{P}(\boldsymbol{N} = \boldsymbol{x}) \equiv q(\boldsymbol{x}), \qquad (20)$$

the right-hand side being given by (14) for the independent mutation model, or by the solution of the recursion (17) in the general case. A least squares method obtains estimates by minimizing the squared error function

$$\sum_{\boldsymbol{x}} (V_{n,\boldsymbol{x}} - q(\boldsymbol{x}))^2.$$

This moment estimator makes fuller use of the data than the estimate based on the number of segregating sites. An alternative estimator, also described in Lundstrom et al. (loc. cit.) is based on the assumption that the sites are evolving independently. This approximation, which is reasonable for large substitution rates (where the between–sites correlations are effectively washed out), produces a likelihood function proportional to

$$\sum_{\boldsymbol{x}} V_{n,\boldsymbol{x}} \log q(\boldsymbol{x}),$$

that can then be maximized to obtain parameter estimates.

For the mtDNA pyrimidine data, the moment method and the (independent sites) maximum likelihood method gave estimates of the C to T rate as $\alpha = 2q_{12} = 0.02$, and the T to C rate as $\beta = 2q_{21} = 0.03$. These are in close agreement with the segregating sites estimator described above. To assess the variability in the estimates of α and β, we used the top-down simulation described in Section 2.2.1, arriving at empirical bootstrap confidence intervals of $(0.01, 0.04)$ for α, and $(0.02, 0.06)$ for β. These rates correspond to substitution probabilities of between 17×10^{-6} and 33×10^{-6} per site per generation for transitions from C to T, and between 25×10^{-6} and 50×10^{-6} per site per generation for transitions from T to C.

The adequacy of these estimates depends, of course, on how well the model fits the data. To assess this, we investigated how well key features of the data are reflected in simulations of the coalescent process with the given estimated rates. As might be expected, the overall base frequencies and the number of segregating sites observed in the data are accurately reflected in the simulations. One poor aspect of the fit concerned the number of *distinct sequences* observed in the simulations (9 to 17 per sample) compared to the 24 observed in the data. There are several reasons why such a poor fit might be observed, among them: (i) Site-specific variability in mutation rates; (ii) Admixture between genetically distinct tribes; and (iii) Fluctuations in population size that are not captured in the model. Further discussion of these

points may be found in Lundstrom, Tavaré and Ward (1992a), and in Section 4.

At this point, we have come to our mathematical vignette, where population genetics theory intersects with an interesting area in combinatorics. The mathematical level of the vignette is somewhat higher than our discussion of the coalescent; readers primarily interested in aspects of the coalescent might feel justified in skipping to Section 4.

3 Mathematical Vignette: Approximating combinatorial structures

Our mathematical vignette takes us from the world of population genetics to that of probabilistic combinatorics. We show how the Ewens sampling formula (ESF), whose origins in population genetics were described in Section 2.1, plays a central role in approximating the probabilistic structure of a class of combinatorial models. This brief account follows Arratia and Tavaré (1993), to which the interested reader is referred for further results. Our first task is to describe the combinatorial content of the ESF itself.

3.1 Approximations for the Ewens Sampling Formula

First, we recall Cauchy's formula for the number $N(\boldsymbol{a}) \equiv N(a_1, \ldots, a_n)$ of permutations of n objects that have a_1 cycles of length 1, a_2 cycles of length 2, $\ldots$, a_n cycles of length n:

$$N(\boldsymbol{a}) = \mathbf{1}(\sum_{l=1}^{n} l a_l = n)\, n! \prod_{j=1}^{n} \left(\frac{1}{j}\right)^{a_j} \frac{1}{a_j!}, \tag{21}$$

$\mathbf{1}(A)$ denoting the indicator of the event A. If each of the $n!$ permutations is assumed to be equally likely, then a random permutation has cycle index $\boldsymbol{a}$

with probability

$$\mathbb{P}(C_1 = a_1, \ldots, C_n = a_n) = \frac{N(\boldsymbol{a})}{n!} = \mathbf{1}(\sum_{l=1}^{n} l a_l = n) \prod_{j=1}^{n} \left(\frac{1}{j}\right)^{a_j} \frac{1}{a_j!}, \qquad (22)$$

where $C_j \equiv C_j(n)$ is the number of cycles of size j in the permutation. Comparison with (4) shows that $(C_1, \ldots, C_n)$ has the ESF with parameter $\theta = 1$. To give the permutation representation of the ESF for arbitrary θ, we need only suppose that for some $\theta > 0$,

$$\mathbb{P}(\pi) = c\theta^{|\pi|}, \ \pi \in S_n, \qquad (23)$$

where $|\pi|$ denotes the number of cycles in the permutation $\pi \in S_n$, where S_n is the set of permutations of n objects. The parameter c is a normalizing constant, which may be evaluated as follows. The number of permutations in S_n with k cycles is $|S_n^k|$, the absolute value of the Stirling number of the first kind. Hence

$$1 = \sum_{\pi \in S_n} \mathbb{P}(\pi) = \sum_{k=1}^{n} \sum_{\pi : |\pi| = k} \mathbb{P}(\pi) = c \sum_{k=1}^{n} |S_n^k| \theta^k = c\theta_{(n)},$$

so that $c^{-1} = \theta_{(n)}$. It follows that under this model,

$$\mathbb{P}(C_1 = a_1, \ldots, C_n = a_n) = \mathbf{1}(\sum_{l=1}^{n} l a_l = n) \, \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!}. \qquad (24)$$

We can see that θ-biasing a random permutation gives the ESF directly. The next ingredient in our story is the observation that the law in (24) may be represented as the joint law of *independent* Poisson random variables $Z_1, \ldots, Z_n$, having $\mathbb{E}Z_j = \theta/j$, conditional on $T \equiv \sum_{j=1}^{n} j Z_j = n$:

$$\mathcal{L}(C_1, C_2, \ldots, C_n) = \mathcal{L}(Z_1, Z_2, \ldots, Z_n | T = n). \qquad (25)$$

This follows because

$$\mathbb{P}(Z_1 = a_1, \ldots, Z_n = a_n | T = n) = \frac{\mathbf{1}(\sum_{l=1}^{n} l a_l = n)}{\mathbb{P}(T = n)} \prod_{j=1}^{n} \frac{e^{-\theta/j}(\theta/j)^{a_j}}{a_j!},$$

which agrees with (24) apart from a norming constant that does not vary with $a_1, \ldots, a_n$; since both formulas are probabilities, the norming constants must be equal.

Equation (25) suggests that we might usefully approximate the *dependent* random variables $C_1, \ldots, C_n$ by the independent random variables $Z_1, \ldots, Z_n$. This turns out to be too ambitious, but we can get away with just a little less. For any $b \in [n] \equiv \{1, 2, \ldots, n\}$, we can approximate the joint laws of $\boldsymbol{C}_b \equiv \boldsymbol{C}_b(n) \equiv (C_1, \ldots, C_b)$ by those of $\boldsymbol{Z}_b \equiv (Z_1, \ldots, Z_b)$, with an error that tends to 0 as $n \to \infty$ as long as $b = o(n)$, that is $b/n \to 0$.

As our measure of how well such an approximation might be expected to work, we use total variation distance as a metric on the space of (discrete) probability measures. Three equivalent definitions of the total variation distance $d_b(n)$ between (the law of) $\boldsymbol{C}_b$ and (the law of) $\boldsymbol{Z}_b$ are given below:

$$\begin{aligned} d_b(n) &\equiv d_{TV}(\mathcal{L}(\boldsymbol{C}_b(n)), \mathcal{L}(\boldsymbol{Z}_b)) \\ &= \sup_{A \subseteq \mathbb{N}^b} |\mathbb{P}(\boldsymbol{C}_b(n) \in A) - \mathbb{P}(\boldsymbol{Z}_b \in A)| \\ &= \frac{1}{2} \sum_{\boldsymbol{a} \in \mathbb{N}^b} |\mathbb{P}(\boldsymbol{C}_b(n) = \boldsymbol{a}) - \mathbb{P}(\boldsymbol{Z}_b = \boldsymbol{a})| \\ &= \inf_{couplings} \mathbb{P}(\boldsymbol{C}_b(n) \neq \boldsymbol{Z}_b). \end{aligned} \tag{26}$$

In (26), the infimum is taken over all couplings of $\boldsymbol{C}_b$ and $\boldsymbol{Z}_b$ on a common probability space, and $\mathbb{N} \equiv \{0, 1, 2, \ldots\}$. Arratia, Barbour and Tavaré (1992) use a particular coupling to show that there is a universal constant $c = c(\theta)$ with $c(1) = 2$ such that

$$d_b(n) \leq c(\theta) \frac{b}{n}, \tag{27}$$

so that indeed $\boldsymbol{C}_b$ and $\boldsymbol{Z}_b$ may be coupled closely if (and, it turns out, only if) $b = o(n)$.

3.2 Combinatorial assemblies

The spirit of the approximations in Section 3.1 – replacing a dependent process by an independent one – carries over to other combinatorial structures. The first of these is the class of *assemblies*. These are labelled structures built as follows. The set $\{1, 2, \ldots, n\}$ is partitioned into a_k subsets of size k, for $k = 1, 2, \ldots, n$, and each subset of size k is marked as one of m_k indecomposable components of size k. For example, in the case of permutations $m_k = (k-1)!$, the components of size k being cycles on k elements. The number of structures $N(\boldsymbol{a})$ of weight n having a_i components of size i, $i = 1, 2, \ldots, n$ is therefore given by

$$N(\boldsymbol{a}) = \mathbf{1}(\sum_{l=1}^{n} l a_l = n)\, n! \prod_{j=1}^{n} \left(\frac{m_j}{j!}\right)^{a_j} \frac{1}{a_j!}, \tag{28}$$

and the total number $p(n)$ of structures of weight n is given by

$$p(n) = \sum_{\boldsymbol{a}} N(\boldsymbol{a}). \tag{29}$$

A random structure of weight n is obtained by choosing each of the $p(n)$ possibilities with equal probability. If $C_j \equiv C_j(n)$ denotes the number of components of size j, then

$$\mathbb{P}(C_1 = a_1, \ldots, C_n = a_n) = \frac{N(\boldsymbol{a})}{p(n)} \tag{30}$$

In the case of permutations, this reduces to (22), because then $m_j/j! = 1/j$. Note that for any $x > 0$, the probability above is proportional to

$$\mathbf{1}(\sum_{l=1}^{n} l a_l = n) \prod_{j=1}^{n} \left(\frac{m_j x^j}{j!}\right)^{a_j} \frac{1}{a_j!},$$

so that by comparison with (22) we see that $\mathcal{L}(C_1, \ldots, C_n) = \mathcal{L}(Z_1, \ldots, Z_n | T = n)$, where the Z_i are independent Poisson random variables with means

$$\mathbb{E}Z_i = \frac{m_i x^i}{i!}, i = 1, 2, \ldots.$$

In particular this implies that

$$\begin{aligned} d_b(n) &= d_{TV}(\mathcal{L}(\boldsymbol{C}_b), \mathcal{L}(\boldsymbol{Z}_b)) \\ &= d_{TV}(\mathcal{L}(R_b), \mathcal{L}(R_b|T=n)), \end{aligned}$$

where $R_b = \sum_{1 \le i \le b} iZ_i$. This observation reduces the calculation of a total variation distance between two processes to a total variation distance between two random variables. We focus our attention on the class of assemblies that satisfies the *logarithmic condition*

$$\frac{m_i}{i!} \sim \frac{\kappa y^i}{i}, \ i \to \infty \tag{31}$$

for some $\kappa, y > 0$. Among these are random permutations (for which (31) holds identically in i with $\kappa = y = 1$), and random mappings of $[n]$ to itself, for which $m_i = (i-1)! \sum_{j=0}^{i-1} i^j/j!, \kappa = 1/2, y = e$. The study of random mappings has a long and venerable history in the combinatorics literature, reviewed in Mutafciev (1984), Kolchin (1986), and Flajolet and Odlyzko (1991) for example.

For the logarithmic class we may choose $x = y^{-1}$, and then it is known (under a mild additional rate of convergence in (31)) that

$$d_b(n) = O\left(\frac{b}{n}\right), \tag{32}$$

just as for the ESF. Indeed, more detailed information is available. For example, Arratia, Stark, and Tavaré (1993) show that for fixed b,

$$d_b(n) \sim \frac{1}{2n}|\kappa - 1| \ \mathbb{E}|R_b - \mathbb{E}R_b|. \tag{33}$$

The term $|\kappa - 1|$ reflects the similarity of the structure to an ESF with parameter κ, whereas the term $\mathbb{E}|R_b - \mathbb{E}R_b|$ reflects the local behavior of the structure.

The θ-biased structures, those with probability proportional to the number of components, may also be studied in this way. In particular (30) holds, the Poisson-distributed Z_i now having means

$$\mathbb{E}Z_i = \frac{\theta x^i m_i}{i!}.$$

The accuracy of the approximation of $\boldsymbol{C}_b$ by $\boldsymbol{Z}_b$ for the logarithmic class is still measured by (32) and (33), where κ is replaced by $\theta\kappa$. The ESF is the case in which $x = \kappa = 1$, since (31) holds identically in i.

A rather weak consequence of the bounds typified by (32) and (33) is the fact that for each fixed b, $(C_1(n), \ldots, C_b(n)) \Rightarrow (Z_1, \ldots, Z_b)$, so that the component counting process $\boldsymbol{C}$ converges in distribution (in $\mathbb{R}^\infty$) to the independent process $\boldsymbol{Z}$. For each n, we are comparing the combinatorial process to a single limiting process. This recovers the classical result of Goncharov (1944) showing that the cycle counts of a random permutation are asymptotically independent Poisson random variables with means $1/i$. The analog for random mappings is due to Kolchin (1976).

There are many uses to which such total variation estimates can be put. In essence, functionals of the dependent process that depend mainly on the small component counts (that is, on components of size $o(n)$) are well approximated by the corresponding functional of the independent process, which is often much easier to analyse. A typical example shows that the total number of components in such a structure has asymptotically a Normal distribution, with mean and variance $\kappa\theta \log n$. A corresponding functional central limit theorem follows by precisely the same methods. In addition, these estimates lead to bounds on the distance between the laws of such functionals. Some examples that illustrate the power of this approach may be found in Arratia and Tavaré (1992) and Arratia, Barbour and Tavaré (1993a).

3.3 Other combinatorial structures

The strategy employed for assemblies also works for other combinatorial structures, including *multisets* and *selections*. We focus just on the multiset case. To build such structures, which are now unlabelled, imagine a supply of m_j irreducible components of weight j, and build an object of total weight n by choosing components with replacement. The number $N(\boldsymbol{a})$ of structures of weight n having a_j components of size j, $j = 1, 2, \ldots, n$ is

$$N(\boldsymbol{a}) = \prod_{j=1}^{n} \binom{a_j + m_j - 1}{a_j} \mathbf{1}(\sum_{l=1}^{n} l a_l = n), \tag{34}$$

and the total number of structures of weight n is $p(n) = \sum_{\boldsymbol{a}} N(\boldsymbol{a})$. A random multiset of size n has a_j components of size j with probability

$$\frac{1}{p(n)} \prod_{j=1}^{n} \binom{a_j + m_j - 1}{a_j} \mathbf{1}(\sum_{l=1}^{n} l a_l = n). \tag{35}$$

The ingredient common to assemblies and multisets is the fact that

$$\mathcal{L}(C_1, \ldots, C_n) = \mathcal{L}(Z_1, \ldots, Z_n | T = n),$$

but the approximating independent random variables $\{Z_j\}$ are no longer Poisson, but rather negative binomial with parameters m_i and x^i:

$$\mathbb{P}(Z_i = k) = \binom{m_i + k - 1}{k} (1 - x^i) x^{ik}, \quad k = 0, 1, \ldots, \tag{36}$$

valid for $0 < x < 1$. In the θ-biased case, the Z_i are negative binomial with parameters m_i and θx^i, for any $\theta < x^{-1}$.

The most studied example in this setting concerns the factorization of a random monic polynomial over the finite field $GF(q)$. The components of size i are precisely the irreducible factors of degree i, there being

$$m_i = \frac{1}{i} \sum_{j|i} \mu(i/j) q^j$$

of them. The function $\mu(\cdot)$ is the Möbius function: $\mu(k) = -1$ or 1 according as k is the product of an odd or even number of distinct prime factors, and $\mu(k) = 0$ if k is divisible by the square of a prime. The logarithmic condition

$$m_i \sim \frac{\kappa y^i}{i}, \ i \to \infty, \tag{37}$$

is satisfied by random polynomials with $\kappa = 1$ and $y = q$. For this logarithmic class the total variation estimates (32) and (33) apply once more (with appropriate modification for the θ-biased case), and the same techniques described at the end of the previous section may then be used to study the behavior of many interesting functionals. In particular, examples describing the functional central limit theorem, with error estimates, for the random polynomial case, may be found in Arratia, Barbour and Tavaré (1993a).

3.4 The large components

Thus far we have described how we might approximate a complicated dependent process (the counts of small components) by a simpler, independent process, with an estimate of the error involved. It is natural to ask what can be said about the large component counts. To describe this, we return once more to the ESF.

Let $L_1 \equiv L_1(n) \geq L_2 \geq \cdots \geq L_K$ denote the sizes of the largest cycle, the second largest cycle, ..., the smallest of the K cycles in a θ-biased random permutation. We will define $L_j \equiv L_j(n) = 0$, $j > K$. It is known from the work of Kingman (1974, 1977) that the random vector $n^{-1}(L_1, L_2, \ldots, L_K, 0, 0 \ldots)$ converges in distribution to a random vector $(X_1, X_2, \ldots)$ satisfying $\sum X_j = 1$ almost surely. The vector $\boldsymbol{X} = (X_1, X_2, \ldots)$ has the *Poisson-Dirichlet distribution* with parameter θ, which we denote by $\mathcal{PD}(\theta)$. There are a number of characterizations of $\mathcal{PD}(\theta)$, among them Kingman's original definition: Let $\sigma_{(1)} \geq \sigma_{(2)} \geq \cdots > 0$ denote the points of a Poisson process on $(0, \infty)$ having mean measure with density $\theta e^{-x}/x, x > 0$,

and set $\sigma = \sum_{i \geq 1} \sigma_{(i)}$. Then

$$(X_1, X_2, \ldots) \stackrel{\mathrm{d}}{=} \left(\frac{\sigma_{(1)}}{\sigma}, \frac{\sigma_{(2)}}{\sigma}, \ldots \right).$$

We now know that the large component sizes, those that are $O(n)$, of a θ-biased random permutation are described asymptotically by the $\mathcal{PD}(\theta)$ law. What can be said about the large components of the other combinatorial structures we have seen? We will focus once more on the logarithmic structures that satisfy either condition (31) or (37), where population genetics has a crucial role to play once more.

In approximating the behavior of counts of large components $(C_{r+1}, \ldots, C_n)$ we should not expect to be able to compare to an *independent* process because, for example, there can be at most $\lfloor n/j \rfloor$ components of size j or greater, this condition forcing very strong correlations on the counts of large components. However, we should be able to compare the component counting process $\boldsymbol{C}^r \equiv (C_{r+1}, \ldots, C_n)$ of the combinatorial structure to the ESF process $\hat{\boldsymbol{C}}^r \equiv (\hat{C}_{r+1}, \ldots, \hat{C}_n)$, say. The approximating process is still discrete and, although not independent, it has a simpler structure than the original process. For random polynomials, it is shown in Arratia, Barbour and Tavaré (1993a) that

$$d_{TV}(\mathcal{L}(\boldsymbol{C}^r), \mathcal{L}(\hat{\boldsymbol{C}}^r)) = O\left(\frac{1}{r}\right), \tag{38}$$

so that the counts of factors of large degree can indeed be compared successfully to the corresponding counts for the ESF. The estimate in (38) has as a consequence the fact that the (renormalized) factors of largest degree have asymptotically the $\mathcal{PD}(1)$ law, a result that also follows from work of Hansen (1991). In addition, a rate of convergence is also available. In fact, (38) essentially holds for any of the logarithmic class; cf. Arratia, Barbour and Tavaré (1993b).

In conclusion, we have seen that a variety of interesting functionals of the component structure of certain combinatorial processes may be approximated

in total variation norm by either that functional of an independent process, or that functional of the ESF itself. The important aspect of this is the focus on *discrete* approximating processes, rather than those found by renormalizing to obtain a continuous limit. In a very real sense, our knowledge of 'the biology of random permutations', as described by the ESF, has provided a crucial ingredient in one area of probabilistic combinatorics.

4 Where to next?

In the preceding sections, we have illustrated how coalescent techniques may be used to model the evolution of samples of selectively neutral DNA sequence data. Some simple techniques for estimating substitution rates, some based on likelihood methods and some on more ad hoc moment methods, were reviewed. We also illustrated how the probabilistic structure of the coalescent might be used to simulate observations in order to assess the variability of such estimators.

4.1 Likelihood methods

Notwithstanding the lack of recombination and selection, inference about substitution rates in such regions poses some difficult statistical and computational problems. Most of these are due to the apparently heterogeneous nature of the substitution process in different regions of the sequence. One of the outstanding open problems in this area is the development of practical likelihood methods for sequence data. Inference techniques for sequence data from a fixed (but typically unknown) tree are reviewed in Felsenstein (1988). The added ingredient in the population genetics setting is the random nature of the coalescent itself – in principle, we have to average likelihoods on trees over the underlying coalescent sample paths. The computational problems involved in this are enormous. The likelihood can be thought of as a sum

(over tree topologies) of terms, in each of which the probability of the configuration of alleles given the branching order and coalescence times $T_n, \ldots, T_2$ is averaged over the law of $T_n, \ldots, T_2$. Monte Carlo techniques might be employed in its evaluation. One approach, using a bootstrap technique, is described by Felsenstein (1992).

An alternative approach is to compute likelihoods numerically using the recursion in equation (17). The probabilistic structure of the coalescent takes care of the integration, and the problem is, in principle at least, simpler. For small sample sizes and simple mutation schemes this is possible (see Lundstrom (1990) for example), but it is computationally prohibitive even for samples of the size discussed earlier. An alternative is the Markov chain Monte Carlo approach in Griffiths and Tavaré (1993), in which equation (17) is used to construct an absorbing Markov process in such a way that the probability $q(\boldsymbol{n})$ in (17) is the expected value of a functional of the process up to the absorption time. That is, represent $q(\boldsymbol{n})$ as

$$q(\boldsymbol{n}) = \mathbb{E}_{\boldsymbol{n}} \prod_{j=0}^{\tau} f(\mathbf{N}(j)), \tag{39}$$

where $\{\mathbf{N}(j), j = 0, 1, \ldots\}$ is a stochastic process determined by (17), and τ is the time it takes this process to reach a particular set of states. Classical simulation methodology may now be used to simulate independent observations with mean $q(\boldsymbol{n})$. The scheme in (39) may be modified to estimate the entire likelihood surface from a single run, providing a computationally feasible method for approximating likelihood surfaces.

As an illustration, we return to the mitochondrial data described in Section 2.3. We saw that of the 21 segregating sites in the sample, 14 were consistent with an infinitely-many-sites model. The remaining 7 sites are described in Table 2. These data comprise a sample of 63 individuals from a $K = 2^7 = 128$ allele model. The allele frequencies are given in Table 2.

The observed fraction of T nucleotides is $\pi_T = 207/441 = 0.469$, and so

Table 2: Incompatible sites and frequencies

Sequence	Site 0	1 T	9 T	10 C	13 C	17 T	18 T	19 C	frequency
1		0	0	0	1	0	0	0	8
2		0	0	0	0	0	0	0	12
3		1	0	0	0	0	0	0	3
4		0	1	0	0	0	0	0	12
5		0	0	0	1	1	0	0	2
6		0	0	1	0	0	0	1	1
7		0	0	0	0	1	1	0	1
8		0	0	0	0	0	1	0	9
9		1	0	0	0	0	1	0	3
10		0	0	1	0	0	1	0	1
11		0	0	0	0	0	1	1	7
12		0	0	1	0	0	1	1	3
13		0	1	1	0	0	1	1	1

Data from Table 1.
The row labelled 0 gives the nucleotide corresponding to 0 at that site.
The last column gives the frequencies of the alleles in the sample.

$\pi_C = 0.531$. We use these to determine the per-site mutation rate matrix Q in (15):

$$Q = \frac{1}{2}\begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} \equiv \frac{\theta}{2s}\left(\begin{pmatrix} \pi_C & \pi_T \\ \pi_C & \pi_T \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right),$$

where $s = 7$. Assuming that π_C and π_T are given by their observed frequencies, there is just the single parameter θ to be estimated. Preliminary simulation results give the maximum likelihood estimate of θ at about $\hat{\theta} = 17$. This corresponds to a *per site* $C \to T$ rate of $\alpha = 1.14$, and a *per site* $T \to C$ rate of $\beta = 1.28$. These rates are about fifty times higher than those based on the analysis in Section 2.4 using all 201 sites. Of course, this set of sites was chosen essentially because of the high mutation rates in the region, and so should represent an extreme estimate of the rates in the whole molecule. Nonetheless, the results do point to the lack of homogeneity in substitution rates in this molecule. For other approaches to the modeling of hypervariable sites, see Lundstrom, Tavaré and Ward (1992b).

4.2 Discussion

The emphasis in this chapter has been on the development of inference techniques for the coalescent, a natural model for the analysis of samples taken from large populations.

An interesting development in the mathematical theory has been the study of measure-valued diffusions initiated by Fleming and Viot (1979). This is a generalization of the 'usual' diffusions so prevalent in the classical theory of population genetics, described for example in Ewens (1979, 1990) and Tavaré (1984). A comprehensive discussion of the Fleming-Viot process appears in Ethier and Kurtz (1993), where the probabilistic structure of a broad range of examples such as multiple loci with recombination, infinitely-many-alleles with selection, multigene families and migration models are discussed in some detail.

Perhaps the most important aspect of the theory that has seen rather little theoretical treatment thus far is the area that might loosely be called *variable population size processes*, and their inference. These issues are becoming more important in the analysis and interpretation of human mitochondrial sequence data. Two recent articles in this area are Slatkin and Hudson (1991) and Rogers and Harpending (1992). Lundstrom, Tavaré and Ward (1992b) note that the effects of variable population size on gene frequency distributions may readily be confounded with the effects of hypervariable regions in the sequences. A careful assessment of the interaction of these two effects seems important, as does a detailed treatment of the effects of spatial structure and population subdivision on the analysis of sequence diversity. The Monte Carlo likelihood methods developed for sequence data in Griffiths and Tavaré (1993) adapt readily to situations like this. They offer a practical approach to inference from very complicated stochastic processes. These techniques are based on genealogical arguments that provide the cornerstone of a firm quantitative basis for the analysis of DNA sequence data, and our

understanding of genomic diversity.

5 References

5.1 General Purpose References

ARRATIA, R. and TAVARÉ, S. (1993a) Independent process approximations for random combinatorial structures. *Adv. Math.*, in press.

AVISE, J.C. (1986) Mitochondrial DNA and the evolutionary genetics of higher animals. *Phil. Trans. R. Soc. Lond. B* 312, 325-342.

ETHIER, S.N., and KURTZ, T.G. (1993) Fleming-Viot processes in population genetics. *SIAM J. Control and Optimization*, in press.

EWENS, W.J. (1979) *Mathematical Population Genetics.* Springer-Verlag, New York.

EWENS, W.J. (1990) Population genetics theory – the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory.* Edited by S. Lessard, pp. 177 – 227. Kluwer Dordrecht, Holland.

FELSENSTEIN, J. (1982) Numerical methods for inferring evolutionary trees. *Quarterly Review of Biology* 57, 379-404.

FELSENSTEIN, J. (1988) Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* 22, 521-565.

HUDSON, R.R. (1991) Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, Volume 7. Edited by D. Futuyma and J. Antonovics, pp 1 – 44.

TAVARÉ, S. (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoret. Popn. Biol.* 26, 119-164.

5.2 Detailed References

ANDERSON, S., BANKIER, A., BARRELL, B., deBRUIJN, M., COULSON, A., DROUIN, J., EPERON, I., NIERLICH, D., ROE, B., SANGER, F., SCHREIER, P., SMITH, A., STADEN, R. & YOUNG, I. (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290, 457-465.

ARRATIA, R. and TAVARÉ, S. (1992) Limit theorems for combinatorial structures via discrete process approximations. *Rand. Struct. Alg.* 3, 321-345.

ARRATIA, R., BARBOUR, A.D. and TAVARÉ, S. (1992) Poisson process approximations for the Ewens sampling formula. *Ann. Appl. Probab.* 2, 519-535.

ARRATIA, R., BARBOUR, A.D. and TAVARÉ, S. (1993a) On random polynomials over finite fields. *Math. Proc. Camb. Phil. Soc.*, in press.

ARRATIA, R., BARBOUR, A.D. and TAVARÉ, S. (1993b) Logarithmic combinatorial structures. *Ann. Probab.*, in preparation.

ARRATIA, R., STARK, D. and TAVARÉ, S. (1993) Total variation asymptotics for Poisson approximations of logarithmic combinatorial assemblies. *Ann. Probob.*, submitted.

CANN, R., STONEKING, M., and WILSON, A.C. (1987) Mitochondrial DNA and human evolution, *Nature* 325, 31-36.

Di RIENZO, A., and WILSON, A.C. (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA, *Proc. Natl. Acad. Sci. USA* 88, 1597-1601.

ESTABROOK, G.F., JOHNSON, C.S. Jr., and McMORRIS, F.R. (1976) An algebraic analysis of cladistic characters. *Discrete Math.* 16, 141-147.

ETHIER, S.N. and GRIFFITHS, R.C. (1987) The infinitely–many–sites model as a measure valued diffusion. *Ann. Probab.*, 15, 515-545.

EWENS, W.J. (1972) The sampling theory of selectively neutral alleles. *Theoret. Popn. Biol.* 3, 87-112.

FELSENSTEIN, J. (1992) Estimating effective population size from samples of sequences: a bootstrap Monte Carlo approach. *Genet. Res. Camb.*, in press.

FLAJOLET, P. and ODLYZKO, A.M. (1990) Random mapping statistics. In *Proc. Eurocrypt '89*, J.-J. Quisquater, editor, pp. 329-354. Lecture Notes in C.S. 434, Springer-Verlag.

FLEMING, W.H. and VIOT, M. (1979) Some measure-valued Markov processes in population genetics theory. *Indiana Univ. Math. J.* 28, 817-843.

GONCHAROV, V.L. (1944) Some facts from combinatorics. *Izvestia Akad. Nauk. SSSR, Ser. Mat.* 8, 3-48. See also: On the field of combinatory analysis. *Translations Amer. Math. Soc.* 19, 1-46.

GRIFFITHS, R.C. (1987) An algorithm for constructing genealogical trees. Statistics Research Report #163, Department of Mathematics, Monash University.

GRIFFITHS, R.C. (1989) Genealogical-tree probabilities in the infinitely-many site model. *J. Math. Biol.* 27, 667-680.

GRIFFITHS, R.C. and TAVARÉ, S. (1993) Simulating probability distributions in the coalescent process. *Theoret. Popn. Biol.*, in press.

HANSEN, J.C. (1991) Order statistics for decomposable combinatorial structures. *Rand. Struct. Alg.*, submitted.

HORAI, S. & HAYASAKA, K. (1990) Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA, *Amer. J. Hum. Gen.* 46, 828-842.

KIMURA, M. (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady influx of mutations. *Genetics* 61, 893-903.

KIMURA, M. and CROW, J.F. (1964) The number of alleles that can be maintained in a finite population. *Genetics* 49, 725-738.

KINGMAN, J.F.C. (1974) Random discrete distributions. *J. Royal Statist. Soc.* 37, 1-22.

KINGMAN, J.F.C. (1977). The population structure associated with the Ewens sampling formula. *Theor. Pop. Biol* 11, 274-283.

KINGMAN, J.F.C (1982a) On the genealogy of large populations. *J. Appl. Prob.* 19A, 27-43.

KINGMAN, J.F.C. (1982b) The coalescent. *Stochastic Processes Appl.* 13, 235-248.

KOLCHIN, V.F. (1976) A problem of the allocation of particles in cells and random mappings. *Theory Probab. Applns.* 21,

48-63.

KOLCHIN, V.F. (1986) *Random Mappings*, Optimization Software, Inc., New York.

LUNDSTROM, R. (1990) *Stochastic models and statistical methods for DNA sequence data.* Ph.D. thesis. Department of Mathematics, University of Utah.

LUNDSTROM, R., TAVARÉ, S. and WARD, R.H. (1992a) Estimating mutation rates from molecular data using the coalescent. *Proc. Natl. Acad. Sci. USA* 89, 5961-5965.

LUNDSTROM, R., TAVARÉ, S. and WARD, R.H. (1992b) Modelling the evolution of the human mitochondrial genome. *Math. Biosci.* 122, 319-336.

MADDISON, D.R. (1991) African origin of human mitochondrial DNA reexamined. *Systematic Zoology* 40, 355-363.

McMORRIS, F.R. (1977) On the compatibility of binary qualitative taxonomic charaters. *Bull. Math. Biol.* 39, 133-138.

MUTAFCIEV, L. (1984) On some stochastic problems of discrete mathematics. In *Mathematics and Education in Mathematics* (Sunny Beach), pp. 57-80. Bulgarian Academy of Sciences, Sophia, Bulgaria.

NEI, M. (1992) Age of the common ancestor of human mitochondrial DNA. *Mol. Biol. Evol.* 9, 1176-1178.

ROGERS, A., and HARPENDING, H. (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* 9, 552-569.

SCHURR, T., BALLINGER, S., GAN, Y., HODGE, J., MERRIWETHER, D.A., LAWRENCE, D., KNOWLER, W., WEISS,

K., & WALLACE, D. (1990) Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages, *Am. J. Hum. Genet.* 47, 613-623.

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

TEMPLETON, A.R. (1992) Human origins and analysis of mitochondrial DNA sequences. *Science* 255, 737.

VIGILANT, L., PENNINGTON, R., HARPENDING, H., KOCHER, T. & WILSON, A.C. (1989) Mitochondrial DNA sequences in single hairs from a South African population, *Proc. Natl. Acad. Sci. USA* 86, 9350-9354.

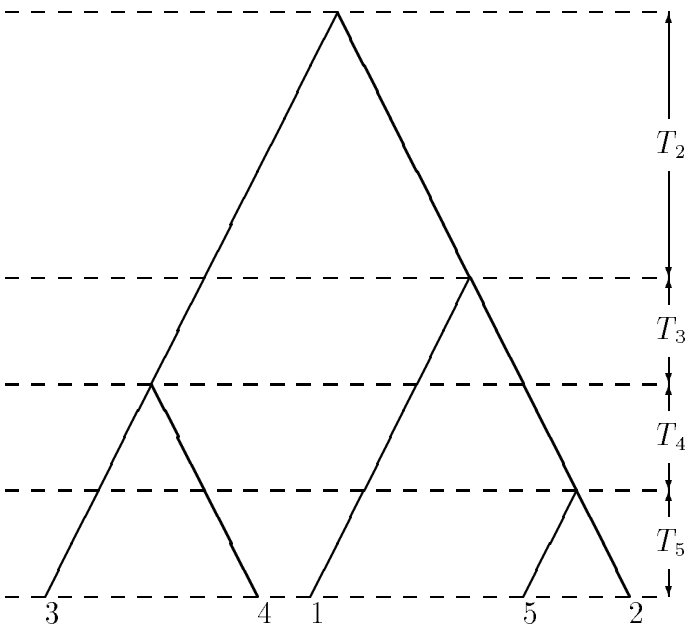
VIGILANT, L., STONEKING, M., HARPENDING, H., HAWKES, K., and WILSON, A.C. (1991) African populations and the evolution of human mitochondrial DNA. *Science* 253, 1503-1507.

WARD, R.H., FRAZIER, B.L. DEW, K. & PÄÄBO, S. (1991) Extensive mitochondrial diversity within a single Amerindian tribe. *Proc. Natl. Acad. Sci. USA* 88, 8720-8724.

WATTERSON, G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theoret. Popn. Biol.* 7, 256-276.

WRIGHT, S. (1968) *Evolution and the Genetics of Populations.* Volume 2, University of Chicago Press, Chicago.

Figure 1: Sample path of the coalescent for n=5



T_j denotes the time during which the sample has j distinct ancestors.

T_j has an exponential distribution with mean $2/j(j-1)$.