1 The basic coalescent

1.1 Introduction

In this chapter, we first motivate the need for a mathematical model that can describe the process generating genetic data, with specific emphasis on human variation. We assume that genetic data are in the form of DNA sequence data. The sequences or genes are all homologous copies of the same genetic region in the genome of a species. Whether one or both copies of a gene in an individual is sampled does not matter here, what matters is their number and genetic type. Such data are collected from one or several present-day populations of a single species, and from this sample we want to infer details about the evolutionary processes that created the data. (A population is here best understood as a population of genes, rather than a population of individuals, because we focus at the level of genes.) The inferential analysis is retrospective; we seek to understand aspects of the sample's (and the population's) evolutionary past through analysis of the present day sample. Below we sketch an analysis of a data set from humans to make more clear the different types of questions that coalescent theory seeks to answer.

The human genome consists of twenty-two autosomal chromosomes, the two sex chromosomes X and Y, and the mitochondrion. One representative sequence of the human genome has recently been approximately determined, and is presently subject to refinement. Additionally, a major effort has now been directed towards determining the population variation in the human genome through determination of single nucleotide polymorphism (SNPs). These are positions that vary within or between human populations. Table 1.1 shows the sizes of each chromosome and the number of positions where variation have so far been detected. The human genome and the variation that is observed is the result of interaction among evolutionary forces, such as mutational and selectional processes, mixing of variation through recombination, and demographic factors, such as the size, history, and geographical structure of the populations. Effects of demography are illustrated by the major colonisations of the globe (Figure 1.1). The present human population migrated out of East Africa approximately 100,000 years ago

Table 1.1 The human genome[a]

Chromosome	Size in Mb	Length in cM	Genes	SNPs	SNP density
1	245	293	1,945	426	1.74
2	243	277	1,283	396	1.63
3	199	233	1,049	317	1.59
4	191	212	765	318	1.66
5	181	198	879	323	1.78
6	171	201	1,053	309	1.79
7	158	184	952	282	1.78
8	146	166	717	256	1.75
9	134	167	755	263	1.96
10	135	182	756	250	1.85
11	135	156	1,294	249	1.84
12	133	169	1,006	213	1.60
13	114	118	341	166	1.46
14	105	129	647	148	1.41
15	100	110	592	166	1.66
16	90	131	900	183	2.03
17	82	129	1,121	144	1.76
18	78	124	267	138	1.77
19	64	110	1,303	110	1.72
20	64	97	631	232	3.63
21	47	60	231	88	1.87
22	49	58	485	121	2.47
X	152	198	750	45	1.61
Y	50	1	94	30	1.60

[a] The second column shows the length of each chromosome in million bases (Mb), the third the length in centiMorgans (cM), the fourth the estimated number of genes for each chromosome, the fifth the number of currently identified SNPs (thousands), and the last column shows the current density of SNPs per kilo bases (kb). The detected number of SNPs and thus also the SNP density will go up within the next few years. The data is from www.ensembl.org, release 17.33.1 (July 2003), except for the genetic map lengths which are taken from the Genethon genetic map.

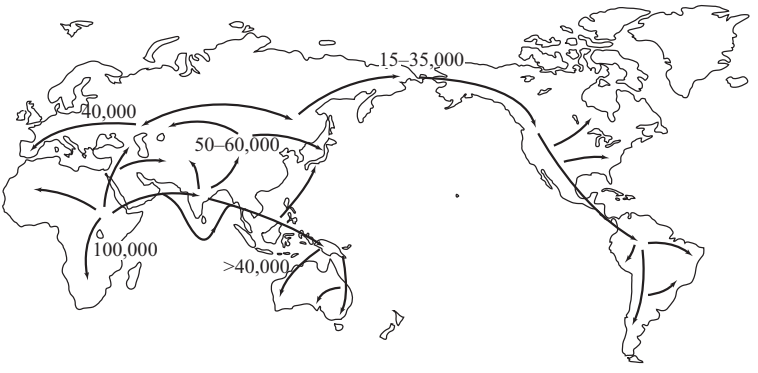

Figure 1.1 The world and historical migrations. The approximate dates of mass migrations (years ago) have mainly been determined by dating fossils found at different locations. Adapted from Cavalli-Sforza (2001).

Table 1.2 Human population growth during the last 12,000 years[a]

Time	10.000 BC	0	1750	1950	2000
Population (in millions)	6	252	771	2,521	6,055
Annual growth (%)	0.008	0.037	0.064	0.594	1.752

[a] The growth rates are point estimates at the different years.

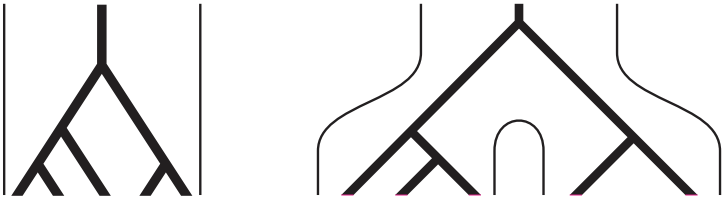

Figure 1.2 A population without geographical structure and a population consisting of two isolated subpopulations descending from a common ancestral population. The width between the thin black lines represents the size of the population at a given time in the past, such that the present time is at the bottom of the figure. The trees represent the ancestral relationship of the sampled genes. In each figure five genes are sampled, a line represents a gene's lineage as its history is back tracked. When two lines meet a common ancestor is encountered. After all genes have found a common ancestor there is just one lineage left, the lineage of the most recent common ancestor of the sample.

to colonise the world, arriving lastly at South America (for further discussion, see Chapter 8). The long term growth of the whole human population is shown in Table 1.2. It appears that the growth rate has been accelerating, that is, the human population is growing faster than exponential.

Geographical factors affect the possible patterns of genetic variation in populations. To illustrate the effect of two factors, population subdivision and population growth, we consider two simple and extreme scenarios. In Figure 1.2, the two geographical scenarios are shown—in the first there is one uniform population, where each gene has the same expected relationship to any other gene in the population (here 'gene' does not imply any knowledge about the type of the gene). The second scenario assumes that the population is made of two subpopulations of equal size that have been separated for, say, 100,000 years, but with low rate of migration between the two subpopulations. It is likely that we would find positions in the DNA-sequence that have, say, adenine (A) in most genes in subpopulation 1 and guanine (G) in most individuals in subpopulation 2. This could be the consequence of a mutation that occurred 90,000 years ago in a gene in subpopulation 2 and was randomly transmitted to most genes in this subpopulation over time. The mutant is only rarely expected to spread in subpopulation 1 unless the migration rate is high. This pattern could be repeated many times throughout the genome, reflecting the division of the two subpopulations; a pattern that would be extremely unlikely in a uniform population.

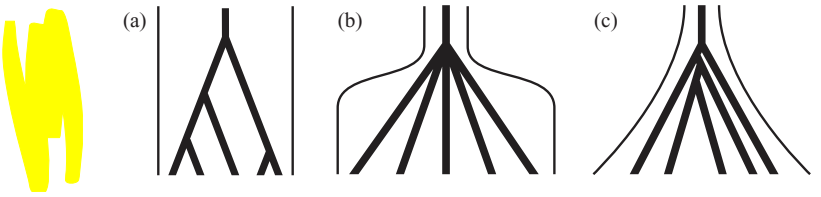

Figure 1.3 A population with (a) constant population size, (b) sudden large increase in population size (explosion), and (c) exponential growth. The width between the thin black lines represents the size of the population at a given time in the past. See the text for further explanation.

Population growth is illustrated by Figure 1.3. Three growth scenarios are shown. In the first, the population size is constant arbitrarily far back in time. In the second case, the population has quickly jumped from being very small to being very large some time ago. The conclusion we can draw about the ancestry of the sample of genes in this case depends on the relative sizes of the population before and after the explosion (time is here running in the usual direction, from the past towards the present time), and on when the explosion happened. If it happened too far back in time the genealogy of a sample will look like the genealogy in the first case, because all genes in the sample will find common ancestors after the explosion. At the other extreme it happened in recent times and all genes have distinct lineages at the time of the explosion. If the population size before the explosion is relatively small, most lineages would collapse instantly just before the explosion. Between these two extremes the genealogy will look like a distorted version of a genealogy from the population in the first case, depending on the sizes of the population before and after the explosion.

In the third case, the population is assumed to grow exponentially, that is, the population size decreases exponentially when we go back in time. This case has properties intermediate between the first two cases: There would be short internal branches just after the most recent common ancestor (MRCA) of the sample. However, we should not expect to see a sudden collapse of lineages as in the second case, because the population size is continuously decreasing. All these conclusions cannot be made conclusively without a complete specification of the model and we stress that they depend on the population sizes through time, how fast the sizes decreases as we go back in time, etc.

The genealogy influences the type of the genes observed in the sample. When a gene is passed on from parent to offspring there is a chance (however small) that the transmitted gene is a mutated copy of the parental gene. If the genealogy of a sample spans many generations there is a higher chance of seeing different types of genes than if the genealogy spans few generations. Also the shape of the genealogy is of importance: In Figure 1.3 mutations in (b) and (c) tend to produce singletons (genetic types that only occur in

one copy in the sample), in contrast to (a) where it is much more likely that a mutation is found in several members of the sample.

The lesson from these simple examples is that the population scenario has consequences for the probability of the observed data set. Mathematical models will give data exact probabilities as a function of underlying parameters describing population history.

1.2 A Y-chromosome data set

If we zoom in on a small fraction of the Y-chromosome (see Figure 1.4), we may look for variation between individual chromosomes in human populations. The Y-chromosome is easy to study since males carry only one copy and there is no recombination. Michael Hammer and coworkers have determined positions of variation along the chromosome and typed these in more than 2000 male individuals around the world. Figure 1.5 shows a subset of this data set. Each column is a segregating site, that is, a position that has two or more types in the sample. For simplicity, we encoded the data as 0s and 1s instead of the actual base pairs (A, T, C, and G), because at most two base pairs are present in each position. Sometimes it is possible to tell (or estimate) which of the two states is the oldest, the ancestral, and which is the youngest, the mutant. If this is the case we use zero as the ancestral state. For this data set, the ancestral state has been determined by comparison with a chimpanzee sequence. Figure 1.5 shows

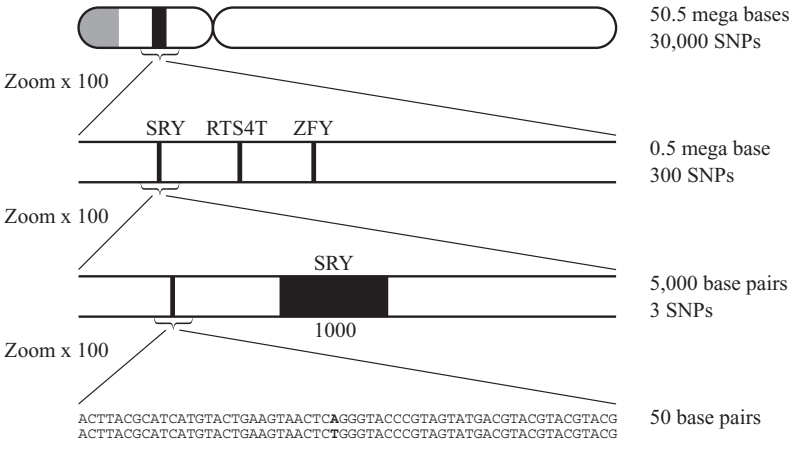

Figure 1.4 Zooming in on a small fraction of the Y-chromosome where a single nucleotide polymorphism is found (in the sex-determining gene, SRY). This polymorphism is one of the nineteen polymorphisms studied by Hammer et al. (2001) and shown in Table 1.3. The lower panel shows a place of polymorphism within a 50 bp interval.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	European	North Africa	South Asia	South Africa	Total
Haplotype	A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		10		47	57
	B	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				41	41
	C	0	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	1	0	0	1		4		5
	E	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	33	70		153	256
	F	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	6	1	19		26
	G	0	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	1	0	1	24	1	5		30
	H	0	1	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0			27		27
	I	0	1	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	47				47
	J	0	1	0	0	0	0	1	0	0	1	0	1	1	1	1	0	0	0	0	51	35	21	1	108
	K	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	6		1	11
	L	0	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	1	1	0	10				10
	N	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14		10		24
	P	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0			3		3
	Q	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	5		3		8
	R	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	160	8	41		209
Consensus		0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	355	131	133	243	862

Figure 1.5 Y-chromosome variation in four human population. There are nineteen segregating sites arranged in fifteen haplotypes named according to Hammer et al. (2001). The last four columns show the frequencies of the haplotypes in Europe, Northern Africa, sub-Saharan Africa, and South Asia. Missing entries indicate that a haplotype is not present. Zero denotes the estimated ancestral state of a segregating site as determined by comparison with a chimpanzee sequence. The bottom row is the consensus sequence, the sequence composed of the most frequent base in each column.

Table 1.3 Summary statistics for the Y-chromosome data set of Hammer et al. (2001)[a]

Population	Sample size	Segregating sites	Pairwise difference
European	355	16	2.48
North Africa	131	13	2.39
South Asia	133	14	2.56
South Africa	243	10	1.65

[a] 'Pairwise difference' is the average number of differences between two Y-chromosomes, for example, 2.48 is the average over $\binom{355}{2} = 62{,}835$ pairs of Y-chromosomes.

patterns of variation at nineteen variable positions in four human populations. The variation is arranged into fifteen distinct haplotypes. A haplotype is a sequence variant of the piece of the chromosomes being investigated. It can immediately be seen that different haplotypes predominate in different populations.

Table 1.3 shows summary statistics for the data set, including the number of segregating sites in each population, and the average number of differences between two chromosomes in each population. The results indicate that the South African sample has less variation than the other samples. The average number of differences between two chromosomes taken from

different populations is 3.18, which is larger than the within population differences. This suggests that there is genetic differentiation between the populations.

We used the computer program Genetree by Bahlo and Griffiths (2000) to estimate different quantities from the data set. This is possible when the data set satisfy the assumptions of the infinite sites model (Chapter 2) and no recombination (Chapter 5), which is the case for the present data set. The infinite sites assumption states that at most one mutation has happened in each position. The fact that only two types, 0 and 1, are present in each position, tell us that there are no obvious contradictions to this assumption. Combining the infinite sites assumption with the assumption of no recombination implies that the haplotypes can be explained or depicted by a gene tree, a graphical representation of the data that connects any two haplotypes in the sample with a series of mutation events, one for each variable site found in the two haplotypes. In this case the gene tree can be rooted because the state of the MRCA has been estimated from a chimpanzee sequence. Due to the convention that 0 represents the ancestral state, the root sequence is the haplotype with zeros only. The root sequence differs from the consensus sequence in three places.

Figure 1.6 shows the gene tree with mutations marked by their numbers (in Figure 1.5). The gene tree is uniquely determined by the data. When the root is given, we can read off events from the gene tree. For example, if we follow the lineage from the root down to haplotype F we encounter a split into three lineages, then a mutation event (mutation 2), a split into three lineages immediately after the mutation event, a mutation event (7), a split event, and finally a mutation event (8). In comparison of lineages the ordering of events along distinct lineages cannot be determined, only the relative order of events along a single lineage is known.

Using coalescent methods it is possible to estimate the time of events, depending on the assumptions of the underlying mathematical model. In this example we used a model with four distinct subpopulations, each exchanging migrants with a scaled rate 1.0 (see Chapter 4 for details). Further, each subpopulation has equal constant population size. The scaled rate is twice the number of migrants arriving in a subpopulation per generation. In the gene tree in Figure 1.6, we see that mutations found in more than one subpopulation or found in more than one gene are estimated to be older than those found in one subpopulation or in one copy only.

The probability of the observed data can be calculated as a function of parameters in the model, such as the scaled mutation rate, or scaled migration rates. We will see how this can be done in Chapter 2. Calculating the probability over the range of parameter values, one can draw the likelihood curve. The parameter value that corresponds to the maximum of the likelihood curve is the maximum likelihood estimate of the parameter, that

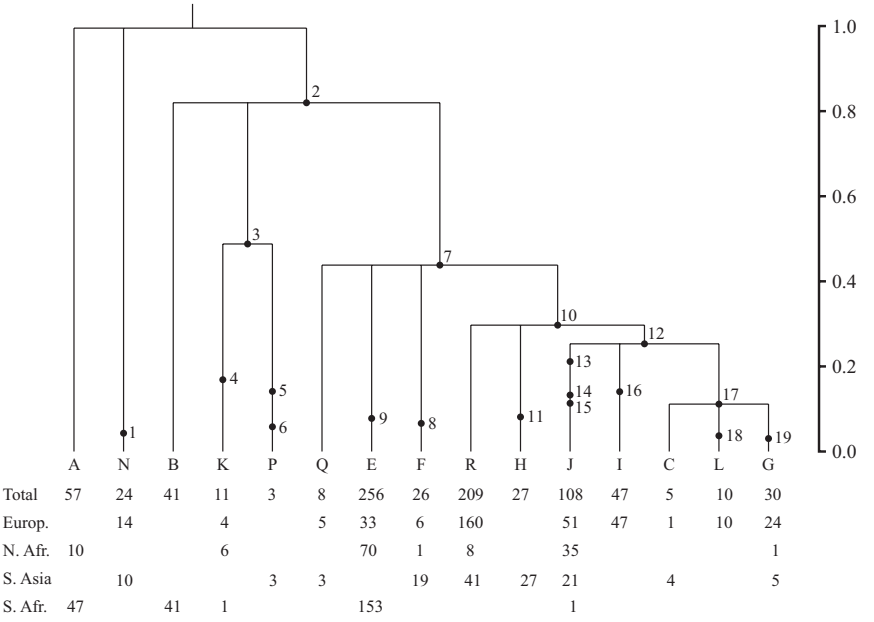

Figure 1.6 Gene tree estimated from the data set in Figure 1.5 using the Genetree program. The vertical axis shows time relative to the estimated time of the root.

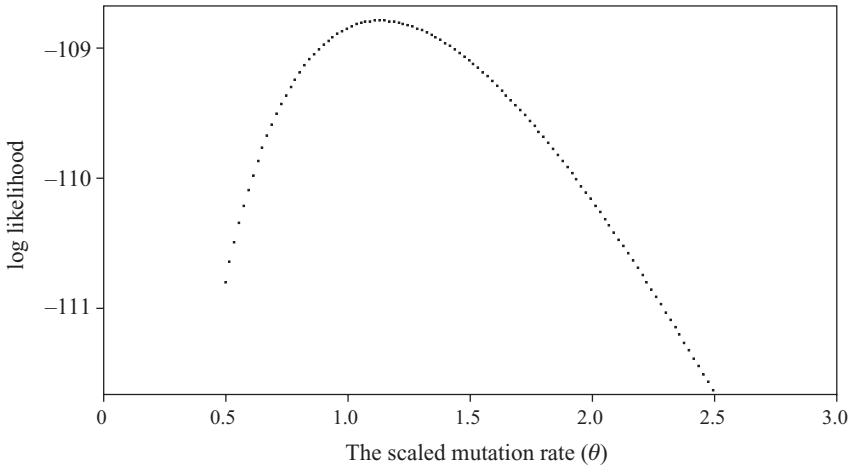

Figure 1.7 The likelihood curve for the scaled mutation rate $\theta = 4Nu$, where u is the mutation rate per generation and N the population size. The maximum likelihood estimate is approximately 1.2. The y-axis is on $\log_{10}$ scale.

is, the value of the parameter that makes the observed data under the chosen model most probable.

Figure 1.7 shows the likelihood surface of the scaled mutation rate θ, which is defined as $4Nu$, where N is the population size and u is the mutation rate per generation (see Chapter 2). This curve was calculated

under the same model as discussed above with fixed migration rates and fixed number of subpopulations. The maximum likelihood estimate of θ is approximately 1.2.

Another quantity of interest is the time of the MRCA of the sample (termed the TMRCA), that is, when was the first time the sample had one ancestor only. The probability density of this quantity given the data is termed the posterior density of the TMRCA. It is shown in Figure 1.8; again the same assumptions are used. Obviously, the haplotype of the

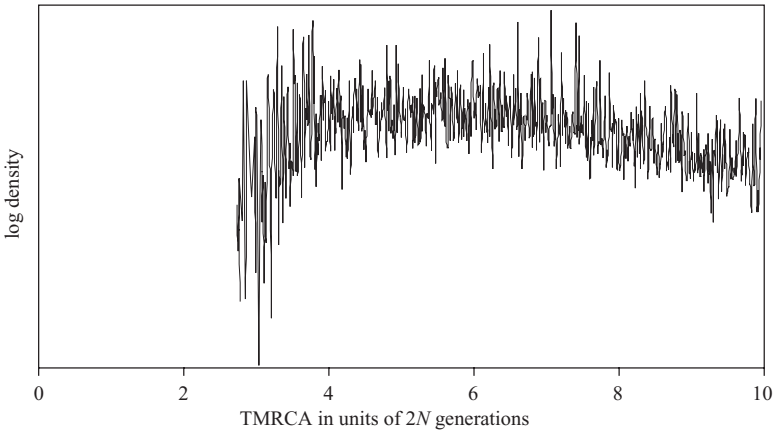

Figure 1.8 Posterior density for the TMRCA for the Y-chromosome data set. The maximum posterior estimate (the value of TMRCA that corresponds to the maximum of the curve) of the TMRCA is approximately six (after the curve has been smoothed). The posterior density is very flat which indicates that the maximum posterior estimate has a wide error margin.

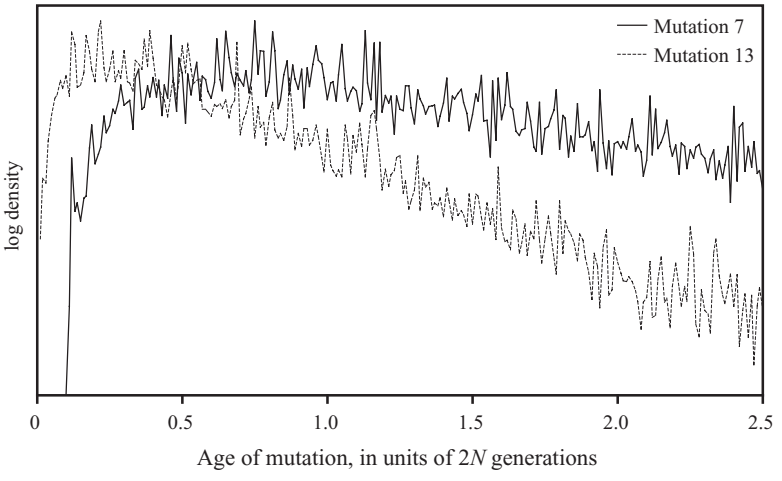

Figure 1.9 Posterior densities for the ages of mutation 7 and mutation 13 of the Y-chromosome data set.

MRCA consists of only zeros, so the TMRCA is at least as old as the most recent occurrence of the haplotype of the MRCA. Figure 1.8 shows that the TMRCA cannot be determined with great precision from this data set under the given model.

While Figure 1.6 showed the estimated relative ages for each mutation, Figure 1.9 shows the posterior density for the ages of two specific mutations. Clearly, the distributions are distinct with mutation 7 estimated to be the oldest in accordance with the point estimate embedded in Figure 1.6.

1.3 Data and theory

The main message from the Y-chromosome example is that explicit probability models of mutations, geography, and reproduction structures allow us to make statements about likely parameter values, ancestral events, dating of events, etc. Without such models, none of this is possible. The above examples are only illustrations of some simple questions one could ask. The relevant questions are both qualitative and quantitative in nature. Examples of qualitative questions are: Does the data show sign of population structure, recombination, population growth, or selection? Quantitative questions aim at estimating parameters, such as the scaled rate of recombination or migration, or the age of a mutation.

The theoretical field underlying the analysis of population data is *population genetics*. This field and many of its major theoretical results are quite old. It was pioneered by three major founding fathers: Sewall Wright (1889–1988), Ronald A. Fisher (1890–1962) and J. B. S. Haldane (1892–1964) during the 1920s and 1930s. They set the outlines for a prospective theory of the fate of genetic variation under migration, selection and random genetic drift. Later, the contributions of Motoo Kimura (1924–1994) made the theory more rigorous by advanced use of diffusion theory. He caused a major shift in the biological world view by introduction of the Neutral Theory. The Neutral Theory postulates that most of the genetic variation observed is selectively neutral. This does not imply that new mutations cannot be selected against, but that such variation is normally rapidly eliminated from the population by selection. The Neutral Theory does imply that only a very small fraction of new mutations are selectively advantageous. It provoked debate because it offered a much smaller role to natural selection than the most prevalent contemporary view; instead it emphasised the importance of stochastic factors such as variation in the frequency of an allele by random genetic drift.

The perspective in the field shifted further in the 1970s and 1980s when the emphasis changed from prospective (looking forward in time)

to retrospective (looking backward in time) methods of analysis. This development was a natural consequence of the availability of genetic data sampled at the present time but shaped by past processes. These processes were of interest to make inferences about.

Theoretical population genetics has since the late 1990s been graced by increased attention from functional biology due to the appearance of completely sequenced genomes and associated data on population variation at the sequence level, turning population genetics into population genomics. This recent rise in importance stems both from the potential of mapping characters to genes (association mapping) and the possibility of a pharmacology tailored to the individual, when knowing the genotype of an individual is crucial for predicting drug metabolism and drug response. It is still an open question if these new fields can live up to the expectations, but almost irrespective of the final outcome, theoretical population genetics is bound to be central in functional biology to a degree that could not have been foreseen a decade ago.

The central approach of genealogical data analysis is a stochastic characterisation of the genealogies that relate the sequences. Evaluating the probability of a given data set then consists of two steps: First, model reproduction in the population which leads to a probabilistic description of the genealogical relationship of the sampled data. Second, each genealogy will generate the data with a specific probability when combined with a model of the mutation process.

1.4 The Wright–Fisher model

A simple model of populations describing the genealogical relationship among genes is that introduced by Wright (1931) and Fisher (1930). This basic model of reproduction provides a dynamic description of the evolution of an idealised population and the transmission of genes from one generation to the next. By genes we refer to a material entity transmitted from one generation to the next. If two copies of a gene can be distinguished we refer to them as different alleles. A sequence is a gene where the nucleotides have been determined. Thus, two sequences of the same gene are different alleles if they are not identical. The basic properties of the model in haploid and diploid versions are illustrated in Figures 1.10 and 1.11.

To facilitate comparison of haploid and diploid models we may assume a population size of $2N$ genes, corresponding to N diploid or $2N$ haploid individuals. Thus, haploid reproduction is modelled assuming $2N$ individuals. Note that other treatments of the Wright–Fisher model may assume N genes instead of $2N$ and that results therefore may differ by a factor of two reflecting this. In the haploid model, each of the genes of generation $t + 1$ are

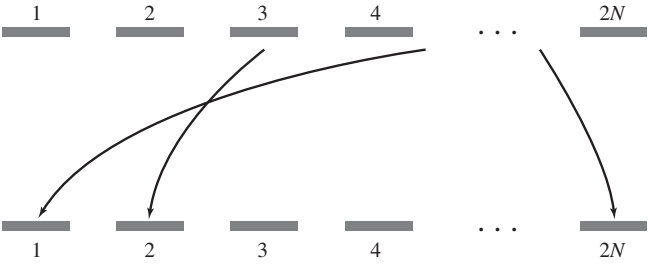

Figure 1.10 Haploid reproduction model. The genes making up the present generation (lower line) are drawn randomly with replacement from the parental generation.

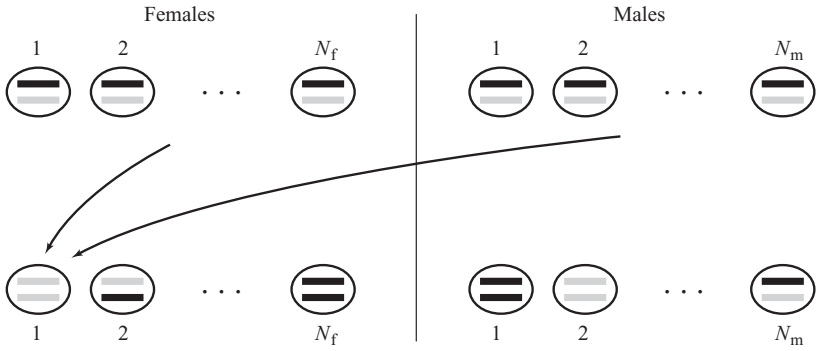

Figure 1.11 Diploid reproduction model. An individual in the present generation (lower line) draws randomly with replacement one of its genes from the female population and the other gene from the male population.

found by copying the gene of a random individual from generation t. This is repeated independently until $2N$ genes have been sampled (Figure 1.10). Each gene in generation $t + 1$ will thus have one parent gene in generation t, but it is a random one. A gene in generation t might not have any descendants in generation $t + 1$ and consequently its lineage has died out.

Diploid reproduction in the case of species with separate sexes assumes two subpopulations—females and males—of sizes N_f and N_m, respectively, with $N = N_f + N_m$, representing again $2N$ genes. Individuals in generation $t + 1$ are created one by one. Each individual chooses a male (father) and a female (mother) from the male and female populations, respectively, from generation t. Within the father and the mother one of the two genes is chosen with probability 0.5 each. This reproductive scheme is illustrated in Figure 1.11. Like the haploid model, each gene has one parent gene (in a male or a female), but each individual has two parents.

If the ancestry of two genes is traced back through time in the haploid model and the diploid model, respectively, there are certain restrictions on which parents a gene could choose in the diploid model relative to the

haploid model. In the haploid model all genes choose independently of each other, while in the diploid model the second gene must choose a different parent from the first gene. Genealogies in the diploid model and the haploid model are probabilistically similar for large choices of N, N_f, and N_m, if adjustments are made to how time is scaled. This will be taken up in Section 1.10. When nothing else is stated we assume the haploid model for convenience.

1.4.1 Assumptions of the Wright–Fisher model

A number of idealised and simplifying assumptions are explicitly and implicitly made in the Wright–Fisher model of reproduction. These are:

1. *Discrete and non-overlapping generations*. In the case of humans, this is equivalent to assuming that everybody has the same lifetime expectancy from conception to reproduction (about 25 years), and that reproduction and death is simultaneous for all individuals and synchronous among all individuals. Fortunately, it turns out that the assumption is of little practical consequence. Models assuming overlapping generations (not all genes give birth or die at the same time) give probabilistically similar genealogies.
2. *Haploid individuals or two subpopulations (males and females)*. In problems that do not relate to selection involving heterosis, it usually has little quantitative consequence to assume a haploid population of size $2N$ in place of a diploid population with N individuals, as noted previously.
3. *The population size is constant*. This is a genuine biological assumption and important quantities of the model will be different if the population is growing, shrinking, oscillating or has gone through a transient very small size, termed a population bottleneck (see Chapter 4).
4. *All individuals are equally fit*. This is a convenient assumption in the introduction of basic concepts, but presumably not realistic for all loci. Relaxing this assumption will be discussed in Chapter 4, most notably, because major questions in genealogical data analysis seek to investigate the presence and strength of natural selection.
5. *The population has no geographical or social structure*. Choosing parents randomly as in the Wright–Fisher model is not a realistic mechanism of reproduction in any real population. Population structure of any kind may greatly affect genealogies (Chapter 4) and this assumption is therefore important for analysis of many real data sets.
6. *The genes (or sequences) in the population are not recombining*. This is an important assumption that needs to be relaxed when analysing many real data sets. Recombination potentially occurs in most sequences, with Y-chromosome, and perhaps mitochondrial DNA, as the primary

exceptions. Unfortunately, relaxing the assumption makes analysis much more mathematically complex, mainly because the sequence sample is no longer related by a genealogical tree but rather a graph (the ancestral recombination graph, see Chapter 5) or a collection of trees.

1.4.2 The number of descendants of a gene in one generation

The number of descendants of a particular gene, i, in generation t is a stochastic variable. Its distribution is straightforward to calculate, since each time a new gene in generation $t+1$ is created it has probability $1/(2N)$ of picking the parent i in generation t and this sampling is performed repeatedly $2N$ times with replacement.

Let v_i be the number of descendants of gene i in generation t, $i = 1, 2, \ldots, 2N$, then

$$P(v_i = k) = \binom{2N}{k} \left(\frac{1}{2N}\right)^k \left(1 - \frac{1}{2N}\right)^{2N-k}. \tag{1.1}$$

This is an example of the binomial distribution, $\mathrm{Bi}(m, p)$, with parameters $m = 2N$ and $p = 1/(2N)$. Thus, the number of genes descending from a given gene is binomially distributed. The moments of the binomial distribution are well-known: For v_i the mean is

$$E(v_i) = mp = 2N\frac{1}{2N} = 1, \tag{1.2}$$

and the variance is

$$\mathrm{Var}(v_i) = mp(1-p) = 2N\frac{1}{2N}\left(1 - \frac{1}{2N}\right) = 1 - \frac{1}{2N}. \tag{1.3}$$

That the mean is one is a consequence of the population size being constant: If the mean number of descendants of a gene were larger/smaller than one the population would be increasing/decreasing in size. The covariance of the offspring numbers for two genes i and j is

$$\mathrm{Cov}(v_i, v_j) = E(v_i v_j) - E(v_i)E(v_j) = -\frac{1}{2N}, \tag{1.4}$$

and the correlation coefficient is

$$\mathrm{Cor}(v_i, v_j) = \frac{\mathrm{Cov}(v_i, v_j)}{\sqrt{\mathrm{Var}(v_i)\mathrm{Var}(v_j)}} = -\frac{1}{2N-1}. \tag{1.5}$$

Thus, v_i and v_j are almost independent of each other for large $2N$. Intuitively, a negative covariance (or correlation) is expected because if

gene i leaves many descendants in the next generation then j is more likely to leave few. This is because the total number of descendants of all genes in a generation is $2N$. Naturally, this effect is more pronounced in small populations than in large populations.

If $2N$ is large then v_i is almost Poisson distributed, Po(1),

$$P(v_i = k) \approx \frac{1}{k!}e^{-1} \tag{1.6}$$

with mean one and variance one. The probability that a gene does not leave descendants is $P(v_i = 0) = e^{-1} \approx 0.37$ and approximately a fraction of $1-e^{-1} \approx 0.63$ of all genes have descendants. Thus, in a large randomly mating population the present day population descends from a relatively small fraction of genes a few generations ago, namely approximately 0.63^t if t generations ago. For example, a population of size $2N = 10,000$ originates from about ten ancestral genes (0.1% of the total population) approximately fifteen generations ago ($10,000 \cdot 0.63^{15} \approx 10$). The lineages of the remaining genes (approximately $10,000 - 10 = 9990$) in the ancestral population fifteen generations ago did not survive until the present day generation.

As soon as the number of ancestral genes becomes small these calculations are no longer valid because they are based on large sample size properties. This is where the coalescent process comes in.

1.4.3 An example

Figure 1.12 shows the Wright–Fisher model of reproduction in a population of size 10 for fifteen reproduction cycles, corresponding to sixteen generations. Each gene is arbitrarily labelled with the position they have in the row (generation) they occupy. Each gene is linked to its ancestor gene in the previous generation. This diagram completely describes the genealogical relationships of all genes during these sixteen generations. Since the labels of the individuals are arbitrary, they can be sorted in each generation, such that all the children of the first parent comes first, then the children of the second parent, etc. This has been done in Figure 1.13. The embedded tree structure now becomes much more apparent, since it is easy to track the ancestry of any set of genes chosen anywhere in the diagram. It is a consequence of the model that any set of genes in a finite number of generations will have only one ancestor. It is quite possible that more than sixteen generations are needed to find the first common ancestor of all genes (the MRCA), even though this is not the case in this example. The concept of a MRCA of the whole population of genes occurring at some time is a 'back in time' statement that has a 'forward in time' consequence: At a certain generation all the lineages starting from the $2N$ genes will die out except for one. To see this, sample the whole population and trace their ancestry back

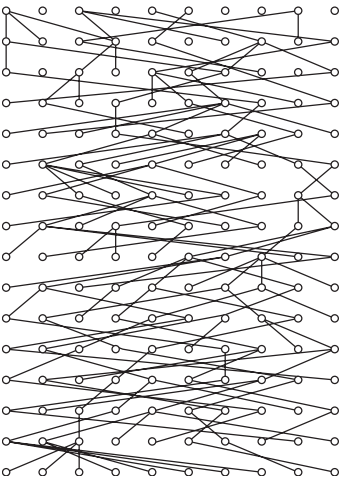

Figure 1.12 The haploid Wright–Fisher model with ten genes applied for sixteen generations, corresponding to applying the haploid model fifteen times starting with the population at the top row.

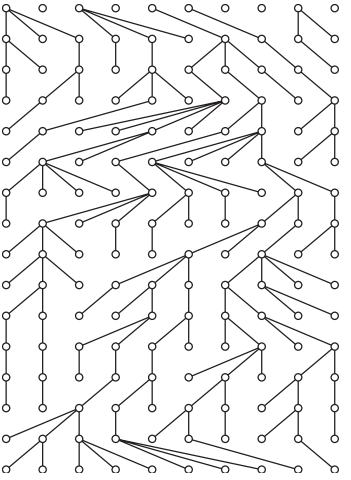

Figure 1.13 The same diagram as in Figure 1.12 but with lineages sorted such that the tree-like nature of the reproductive structure emerges.

until their MRCA has been found. All the other genes in that generation do not have any descendants in the most recent generation.

From the point of view of data analysis, only a sample of n (typically n is much smaller than $2N$, i.e. $n << 2N$) genes are taken from the present population and the genealogical ancestry of this sample is of interest. In Figure 1.14 three genes (1, 2, and 3, which are the first, third, and ninth of the whole population in Figure 1.12) have been sampled randomly in the present population. Edges back in time tracking the ancestors of these three sequences are highlighted. Two generations back in time, genes 1 and 2, find a common ancestor and this lineage might be labelled (1, 2) to reflect this fact. Seven generations further back in time back (1, 2) finds a common ancestor with 3 and the genealogical relationships of the three genes are now fully described. It is possible to follow the MRCA of all three genes further

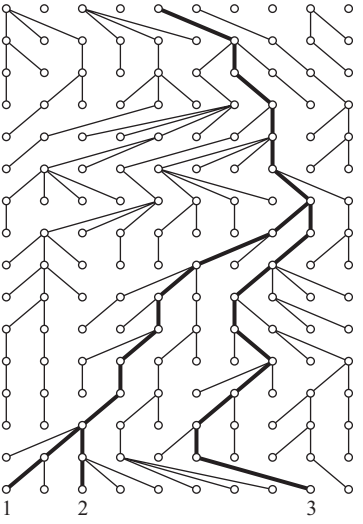

Figure 1.14 The genealogy of three randomly sampled sequences named 1, 2, 3 from left to right. The ancestry of the sequences are marked by bold lines sixteen generations back in time; nine generations back all three sampled sequences have found a common ancestor.

back in time, but there will never be any information on what happened to this MRCA further back in time from the sampled genes. Questions about the structure of genealogies of a population or a sample from a population are our main interest to address, for example:

1. What is the length of the epoch when there were only k lineages?
2. What it is the probability that the number of lineages shrinks by more than one in one generation?
3. What is the general structure of the genealogies generated by this process?

Getting answers to these question will facilitate the interpretation of actual sequence data.

1.5 The geometric distribution

Ubiquitous in population genetics and in coalescent theory in particular is the *Markov property*, which states that the probability of a specific next step in a discrete or continuous process only depends on the present state of the process, that is, the process is without memory of events prior to the present. In genetics it is natural to assume that the probability that something is going to happen (e.g. a mutation or finding a common ancestor) depends only on the situation at present (e.g. the present nucleotide, nucleotide sequence, or number of genes). In processes where time is discretely measured (e.g. in generations) the Markov property is closely associated with the *geometric* distribution. If time is continuous (e.g. as measured by a stopwatch) the

analogous distribution is the *exponential* distribution. The exponential distribution will be discussed in the next section.

In analogy with the binomial distribution, the geometric distribution can be obtained by independent and repeated experiments with two possible outcomes, for example, tossing a coin, or observing whether or not a dice shows six. The binomial distribution describes the outcome of repeating the experiment n times, where each experiment has a probability p of success, and then asking for the total number of successes. The geometric distribution also describes results from a repeated experiment. However, rather than describing the number of successes, it records the time (i.e. the number of trials) until the first success.

Thus, let $X_i, i = 1, 2, \ldots$, be a series of independent and identically distributed experiments with probability p of success and $1 - p$ of failure. In population genetics this could be going back in time and in each generation observing whether two genes had found a common ancestor. Denote success with one and failure with zero. Let T be the waiting time until the first success, that is, $T = \min\{i \mid X_i = 1, i = 1, 2, \ldots\}$. This leads to the geometric probability distribution,

$$P(T = j) = (1 - p)^{j-1} p \tag{1.7}$$

$j = 1, 2, \ldots$, since $T = j$ implies $j - 1$ failures followed by one success. We take $T \sim \text{Geo}(p)$ to mean that T is geometrically distributed with parameter p.

A series of simple properties of geometrically distributed variables can easily be derived. Assume that $t_2 > t_1$. Then

$$P(T > t_2 \mid T > t_1) = P(T > t_2 - t_1), \tag{1.8}$$

$$E(T) = \frac{1}{p}, \tag{1.9}$$

and

$$\text{Var}(T) = \frac{1 - p}{p^2}. \tag{1.10}$$

Let S be a second geometrically distributed variable, $S \sim \text{Geo}(p')$. Assume S is independent of T, then

$$\min(S, T) \sim \text{Geo}(p + p' - pp'). \tag{1.11}$$

Property (1.8) illustrates the lack of memory: Knowing that T has not occurred at time t_1 does not change the probability of when it will occur

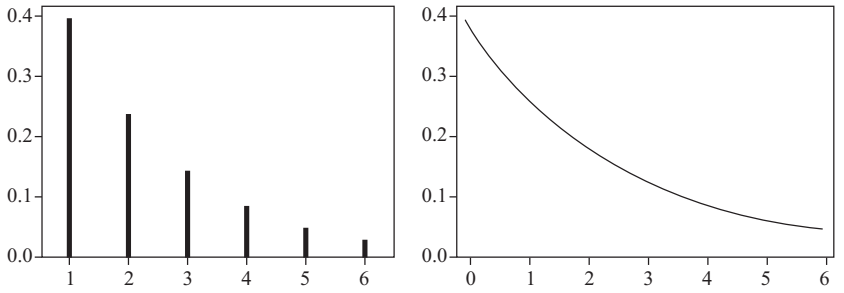

Figure 1.15 Comparison of the continuous exponential density and the discrete geometric density with mean 2.5.

in the future. An example of a geometric distribution with mean 2.5 is displayed in Figure 1.15. Note the long tail of the distribution, and the general fact that the probabilities are decreasing for increasing j.

1.6 The exponential distribution

The exponential distribution can be defined in a variety of ways. It arises naturally from the Markov property, but it can also be obtained as the limit distribution of a series of geometric distributions measured on a finer and finer ladder of time points. One then waits for an event (success) on a continuous line.

Let U have an exponential distribution, that is,

$$P(U \leq t) = 1 - e^{-at}. \tag{1.12}$$

The density function of the distribution is found by differentiation of $P(U \leq t)$ with respect to t,

$$f(t) = \frac{dP(U \leq t)}{dt} = ae^{-at}. \tag{1.13}$$

The exponential distribution is characterised by one parameter only, the intensity a which can be interpreted as the expected number of events in an interval of length one, if the waiting times between events were all independent and had exponential distributions with intensity a. Alternatively, the parameter also characterises the probability that an event occurs soon, because $P(U < t) \approx at$ for small t. The exponential distribution with the same mean as the geometric distribution is also shown in Figure 1.15.

The distribution has properties analogous to those of the geometric distribution. Let V be a second exponentially distributed variable with intensity b.

Assume that V is independent of U, and that $t_2 > t_1$, then

$$P(U > t_2 \mid U > t_1) = P(U > t_2 - t_1), \tag{1.14}$$

$$E(U) = \frac{1}{a}, \tag{1.15}$$

$$\mathrm{Var}(U) = \frac{1}{a^2}, \tag{1.16}$$

$$P(U < V) = \frac{a}{a+b}, \tag{1.17}$$

and

$$\min(U, V) \sim \mathrm{Exp}(a+b). \tag{1.18}$$

Property (1.14) restates the Markov property, that all that matters is the present state (and not how long the process has been in that state or which states it resided in previously). The joint properties of U and V are important because waiting for competing, but independent events, such as coalescent events, mutation events, migration events, etc., is at the core of coalescent theory. Waiting for the first of two possible events to occur is again an exponential variable and the probability that one of two types of events is the first to occur only depends on the relative values of the intensities (property (1.17)).

The geometric distribution can be approximated with an exponential distribution in various ways. Here we focus on one particular approximation that is useful in deriving the continuous time coalescent. Let T have a geometric distribution with parameter p, that is,

$$P(T \geq j) = (1-p)^j. \tag{1.19}$$

If p is small T is typically large and one might measure T on a smaller time scale to counterbalance that p is small. Assume M is some large number such that $a = pM$ and $t = j/M$ are both small compared to M. In the coalescent context M will be of the order of $2N$, p of the order of $1/(2N)$, and j of the order of $2N$. Rewriting $(1-p)^j$ as

$$\left(1 - \frac{pM}{M}\right)^{M \cdot j/M} = \left(1 - \frac{a}{M}\right)^{tM}, \tag{1.20}$$

we obtain

$$P(T \geq j) = P(T/M \geq t) \approx e^{-at} \tag{1.21}$$

from standard mathematical analysis. In consequence, $U = T/M$ is approximately exponential with intensity a.

This approximation involves a change in how time is measured: Instead of measuring in discrete units, time is measured in continuous time such that one unit of continuous time corresponds roughly to M discrete units. In coalescent theory this will be $2N$ discrete generations. This is accomplished by dividing T by M. To jump between the two time scales we either divide by M or multiply by M. As just mentioned, to go from discrete time units we divide j by M and, conversely, we multiply t by M to obtain discrete time from continuous time: for example, if $t = 2.353$ and $M = 100$ then $j = tM \approx 235$.

1.7 The discrete-time coalescent

With the concept of the Wright–Fisher model of reproduction and the properties of the associated important probability distributions, the binomial distribution, the geometric distribution, and the exponential distribution, at hand, we are in a position to derive the basic coalescent process. The presentation does not follow the original formulation by Kingman (1982*a*, *b*) because his formulation is mathematically advanced. What will be derived is what Kingman termed the *n-coalescent* or just the coalescent for a sample of n genes. The coalescent for an infinite (or very large) sample will briefly be discussed at the end of this chapter.

1.7.1 Coalescence of a sample of two genes

What is the distribution of the waiting time until the MRCA of two genes sampled in a haploid model with $2N$ genes? The probability that these two genes find an ancestor in the first generation back in time is $1/(2N)$—the first can choose its parent freely, but the second gene must choose the same parent as the first gene, which is one out of $2N$ possibilities. The probability that the two genes have different ancestors is therefore $1 - 1/(2N)$.

Since sampling in different generations is independent of each other, the probability that two genes find a common ancestor j generations back in time is

$$\left(1 - \frac{1}{2N}\right)^{j-1} \frac{1}{2N}. \tag{1.22}$$

In the first $j - 1$ generations they chose different ancestors, and then in generation j they chose the same ancestor. Thus the coalescence time T_2

for two genes to find a MRCA is distributed as

$$P(T_2 = j) = \left(1 - \frac{1}{2N}\right)^{j-1} \frac{1}{2N}, \tag{1.23}$$

$j = 1, 2, \ldots$, which implies that T_2 is geometrically distributed with parameter $1/(2N)$. The mean of T_2 is therefore $E(T_2) = 1/(1/(2N)) = 2N$ generations. Thus, the expected time until a MRCA is the same as the number of genes in the population. In the example shown in Figure 1.14, where $2N$ is ten, the probability that two randomly picked genes in one generation has the same parent is $1/10$ and the expectation of the waiting time for them to find a common ancestor is ten generations.

1.7.2 Coalescence of a sample of *n* genes

The waiting time for k ($\leq n$) genes to have less than k ancestral lineages can also be calculated. The probability that k genes have k different ancestors in the previous generation is

$$\begin{aligned}\frac{(2N-1)}{2N}\frac{(2N-2)}{2N}\cdots\frac{(2N-k+1)}{2N} &= \prod_{i=1}^{k-1}\left(1-\frac{i}{2N}\right)\\ &= 1 - \sum_{i=1}^{k-1}\frac{j}{2N} + O\left(\frac{1}{N^2}\right) = 1 - \binom{k}{2}\frac{1}{2N} + O\left(\frac{1}{N^2}\right),\end{aligned} \tag{1.24}$$

where $O\left(1/N^2\right)$ is all terms which are divided by N^2 or any higher power of N. Since we assume that n is much smaller than N, $O\left(1/N^2\right)$ is negligible and can be ignored (the effect of this is exemplified below). This approximation is equivalent to ignoring the possibility that more than one pair of genes find a common ancestor in the same generation.

The reasoning behind equation (1.24) is analogous to the two genes case. The first gene can choose freely among the $2N$ genes, the second gene must choose a different parent and can only choose between $2N - 1$ parents. The third gene can only choose among $2N - 2$ possible genes and so forth. Thus, when n is much smaller than N, the probability that no coalescence event occurs is

$$1 - \binom{k}{2}\frac{1}{2N}, \tag{1.25}$$

and the probability of a coalescence event in a given generation is

$$\binom{k}{2}\frac{1}{2N}. \tag{1.26}$$

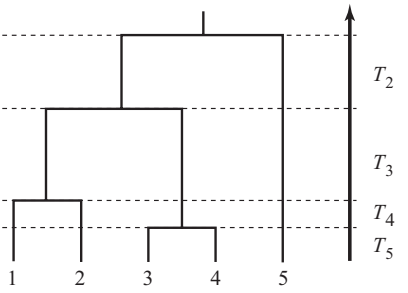

Figure 1.16 The different time epochs in a coalescent tree. T_j, $j = 2, 3, 4, 5$ is the time while there are j ancestors to the sampled five genes.

In consequence the probability that two genes out of the k genes finds a common ancestor $T_k = j$, $j = 1, 2, \ldots$, generations ago is

$$P(T_k = j) \approx \left\{1 - \binom{k}{2}\frac{1}{2N}\right\}^{j-1}\binom{k}{2}\frac{1}{2N}, \tag{1.27}$$

and T_k has approximately a geometric distribution with parameter $\binom{k}{2}/(2N)$. The terminology is shown in Figure 1.16. Because all pairs of genes are equally likely to find a common ancestor, the pair that finds a common ancestor is chosen with equal probability among the $\binom{k}{2}$ possible pairs. The times $T_2, \ldots, T_n$ are independent.

1.7.3 Example: Effect of approximations

Relating the quantities of the previous sections to the example of Figure 1.14, the probability that all three genes have one ancestor in the previous generation is $\frac{1}{10}\frac{1}{10} = \frac{1}{100}$, since the first gene can choose a parent freely, while the next two genes must choose the same parent as the first gene. The probability that three genes have three different ancestors is $\frac{10 \cdot 9 \cdot 8}{10 \cdot 10 \cdot 10}$. The remaining possibility that the three genes have two parents in the previous generation (i.e. one pair of the three possible pairs has only one ancestor) is then $1 - \frac{10 \cdot 9 \cdot 8}{1000} - \frac{1}{100} = \frac{27}{100}$. Using the approximate probability it comes out as

$$\binom{3}{2}\frac{1}{10} = \frac{3}{10}. \tag{1.28}$$

There is a slight difference between the true value and the approximate value, since $\frac{27}{100} - \frac{3}{10} = -\frac{3}{100}$. If the approximate value of $\frac{3}{10}$ is used here, the expected number of generations before any of the three genes have common ancestors is $10/3 \approx 3.33$. $2N = 10$ is very small. For $2N > 100$, the agreement is excellent.

The accuracy of the approximation for large N leads to a formulation of the coalescent with two convenient properties: A model that uses continuous

time and that additionally is independent of $2N$. To accomplish this we need the exponential distribution that was introduced previously.

1.8 The continuous time coalescent

In the Wright–Fisher model time is measured in discrete units, generations. It is conceptually and computationally advantageous to consider continuous time approximations. A natural choice for the coalescent has been to scale in continuous time, so that one unit of time corresponds to the average time for two genes to find a common ancestor, which was just shown to be $2N$ generations. Again, note that other treatments of the coalescent process prefer scaling time by N (or occasionally $4N$) rather than $2N$, leading to results differing by a factor of two. Using any of these transformations of time, the coalescent becomes independent of the population size. It will only be used if we want to translate time back into generations. This emphasises that the structure of the coalescent process is the same for any population as long as the sample size n is small compared to the population size $2N$; only the time scale differs between populations when $2N$ differs.

To derive the continuous coalescent process we let $t = j/(2N)$, where j is time measured in generations. It follows that $j = 2Nt$ translates continuous time t back into generations j. (If $j = 2Nt$ is not an integer j is truncated to the nearest lower integer. For example, $2Nt = 2 \cdot 10^4 \cdot 1.23241$ is truncated to 24648.) The waiting time, T_k^c, in the continuous representation for k genes to have $k-1$ ancestors is exponentially distributed, $T_k^c \sim \text{Exp}(\binom{k}{2})$, that is,

$$P(T_k^c \leq t) \;=\; 1 - e^{-\binom{k}{2}t}. \tag{1.29}$$

This is derived from equations (1.21) and (1.27) letting $t = j/(2N)$, $M = 2N$, and $p = \binom{k}{2}/(2N)$. A continuous time realisation of the coalescent is shown in Figure 1.17, with time scaled in generations on the left and in units of $2N$ generations on the right.

It is now possible to describe a stochastic algorithm that samples genealogies for n genes. Only the version for continuous time is given; the discrete version is analogous to the continuous. We refer to the continuous time coalescent as the basic coalescent or the basic coalescent process.

Algorithm 1

1. Start with $k = n$ genes.
2. Simulate the waiting time T_k^c to the next event, $T_k^c \sim \text{Exp}(\binom{k}{2})$.

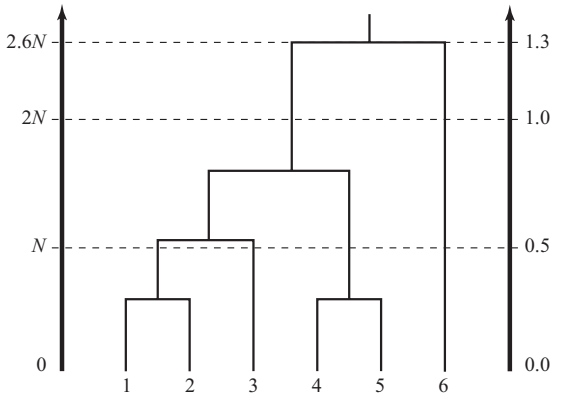

Figure 1.17 A continuous time genealogy with time measured in units of generations (left) and in units of $2N$ generations (right).

3. Choose a random pair (i, j) with $1 \leq i < j \leq k$ uniformly among the $\binom{k}{2}$ possible pairs.
4. Merge i and j into one gene and decrease the sample size by one, $k \to k - 1$.
5. If $k > 1$ go to 2, otherwise stop.

A computational advantage of going back in time when simulating the history of a present day sample relative to an alternative forward approach is that in the former one only needs to keep track of the ancestry of the sample of interest, while in a forward approach the whole population needs to be traced. Referring back to Figure 1.14, a forward approach would have to calculate all edges in the whole illustration and additionally make sure that sufficiently many generations had been simulated, such that the MRCA to all extant sequences had been found. A backward approach would only have to find the highlighted edges.

From now on we will only refer to the continuous model unless stated otherwise and the superscript c is dropped, thus T_k^c is denoted by T_k. Whenever we think of a particular population with a certain size $2N$, time can be translated back into generations by multiplication by $2N$.

1.9 Calculating simple quantities on a coalescent tree

In the continuous coalescent, the properties of the exponential distribution make it easy to calculate a number of important quantities on the genealogy.

1.9.1 The height of a tree

Consider the coalescent tree depicted in Figure 1.16. The height, H_n, of the tree of a sample of size n is the sum of time epochs, T_j, while there

are $j = n, n-1, \ldots, 2$ ancestors. The distribution of H_n is obtained as a convolution of the exponential variables,

$$P(H_n \leq t) = \sum_{k=1}^{n} e^{-\binom{k}{2}t} \frac{(-1)^{k-1}(2k-1)n_{[k]}}{n_{(k)}}, \tag{1.30}$$

where $n_{[k]} = n(n-1)\cdots(n-k+1)$, and $n_{(k)} = n(n+1)\cdots(n+k-1)$. However, the mean of H_n is easily obtained

$$E(H_n) = \sum_{j=2}^{n} E(T_j) = 2\sum_{j=2}^{n} \frac{1}{j(j-1)} = 2\left(1 - \frac{1}{n}\right). \tag{1.31}$$

Similarly, because the T_js are independent, the variance of H_n is

$$\mathrm{Var}(H_n) = \sum_{j=2}^{n} \mathrm{Var}(T_j) = 4\sum_{j=2}^{n} \frac{1}{j^2(j-1)^2}. \tag{1.32}$$

The sum in (1.31) grows towards 2 (scaled in $2N$ generations) as the number of genes increases. The expected waiting time for n genes to find their common ancestor is less than twice that of the expected waiting time for two genes to find their common ancestor ($E(H_2) = E(T_2) = 1$). This is a counterintuitive fact for most people when initially introduced to the coalescent. The example also shows that in the continuous time coalescent we can at least formally consider a sample of infinite size because an infinite sample finds a common ancestor in expected time $2(1 - 1/\infty) = 2$, thus in finite time. It is mathematically convenient that the coalescent model applies for even large sample sizes; sample sizes that might be larger than the size ($2N$) of the Wright–Fisher population it seeks to approximate. These considerations connect back to the discussion at the end of section 1.4.2: In a large population it takes only a few generations before the whole population has only a small number of ancestors; the number of generations is much smaller than the size of the population. The variance of H_n goes to a finite value for increasing n: $\frac{4}{3}(\pi^2 - 9) \approx 1.159$. The sum converges very quickly. One property of H_n is a disproportionately large contribution from T_2, the time from two ancestors to one ancestor. Table 1.4 shows the contribution of different epochs to the variance of H_n.

The property that T_2 contributes significantly to the time of the MRCA can also be seen in Figure 1.18. This figure shows that the variation in the height of the coalescent tree for a sample of twenty is large, and that most of it is caused by the variance in T_2.

Table 1.4 Variance of H_n and L_n and the ratio of the variance of H_n, respectively L_n, to that of T_2

n	2	3	4	5	6	10	15	20
Var(H_n)	1.000	1.111	1.139	1.149	1.153	1.158	1.159	1.159
T_2 contribution	1.000	0.900	0.877	0.870	0.867	0.864	0.863	0.863
Var(L_n)	4.000	5.000	5.444	5.694	5.854	6.159	6.304	6.375
T_2 contribution	1.000	0.800	0.734	0.702	0.683	0.649	0.635	0.627

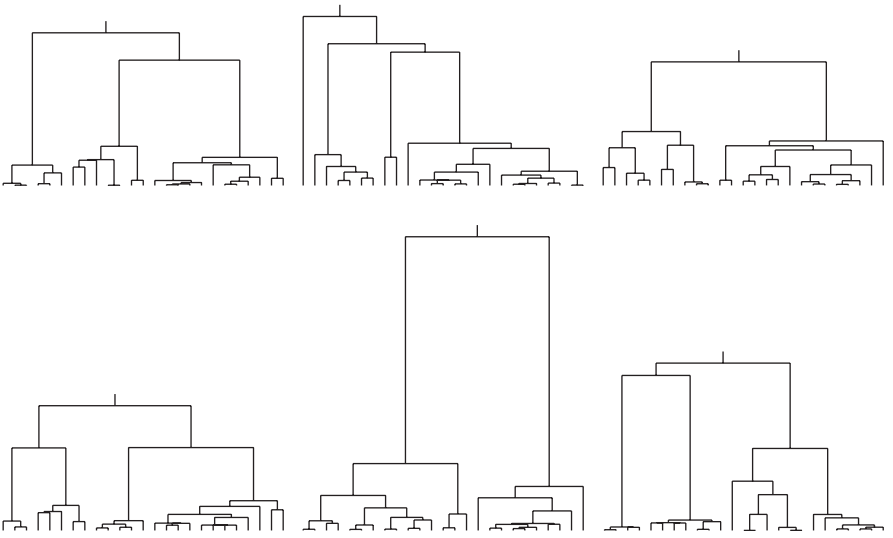

Figure 1.18 A sample of six realisations from the coalescent relating twenty-five genes.

1.9.2 The total branch length of a tree

In contrast to H_n, the distribution of the total branch length L_n has a nice form

$$P(L_n \leq t) = (1 - e^{-t/2})^{n-1}. \tag{1.33}$$

The mean of L_n is most easily obtained by weighting the coalescent times by the number of lineages that exist in that epoch,

$$E(L_n) = \sum_{j=2}^{n} j E(T_j) = 2 \sum_{j=1}^{n-1} \frac{1}{j}. \tag{1.34}$$

This sum does not converge for large n, but grows slowly with n. In fact, it is proportional to the natural logarithm of n,

$$\sum_{j=1}^{n-1} \frac{1}{j} \approx \log(n). \tag{1.35}$$

We can use the mean of L_n to get a sense of how much history genes in a sample share. The genes would share the least history if they all sprung from a common ancestor (assuming they had a common ancestor) some time ago and then evolved along distinct lineages. If the time until the common ancestor is the same as in the basic coalescent, namely $E(H_n) = 2(1 - 1/n)$, then the total branch length would be n times that quantity, or $2n(1 - 1/n) = 2(n - 1)$. Comparing this to the mean of L_n in the basic coalescent, we find the ratio

$$\frac{E(L_n)}{2(n-1)} = \frac{\sum_{j=1}^{n-1} \frac{1}{j}}{n-1} \approx \frac{\log(n)}{n-1}. \tag{1.36}$$

This number is small even for small n; $n = 5$: 52%, $n = 10$: 31%, and $n = 100$: 5.2%. For example if $n = 10$ then on average $100\% - 31\% = 69\%$ of the total time in the genealogy of a sample is shared between two or more genes. Thus coevolution of genes is the rule rather than the exception.

The variance of L_n is

$$\mathrm{Var}(L_n) = \sum_{j=2}^{n} j^2 \mathrm{Var}(T_j) = 4 \sum_{j=1}^{n-1} \frac{1}{j^2}, \tag{1.37}$$

which converges to $2\pi^2/3 \approx 6.579$ as n increases. This implies that for large n, L_n is narrowly centered around $E(L_n)$. Table 1.4 shows the variance contribution of T_2 to the total variance of L_n.

1.9.3 The effect of sampling more sequences

Sampling more sequences is less effective than most would intuitively think. Figure 1.19 shows an example where a sample of ten sequences (in bold) is embedded in a larger sample of fifty sequences. Most of the deep branches (those nearer the root) are already in the tree with ten sequences. This is not surprising since the expected height of a tree for fifty genes is 1.96, while the expected height for a tree for ten genes is 1.80. Thus, a fivefold increase in genes on the average leads to less than 10% increase in height. The increase in branch length is also slow. In the present case the relative increase is $[E(L_{50}) - E(L_{10})]/E(L_{10}) = 58\%$, which again should be compared to

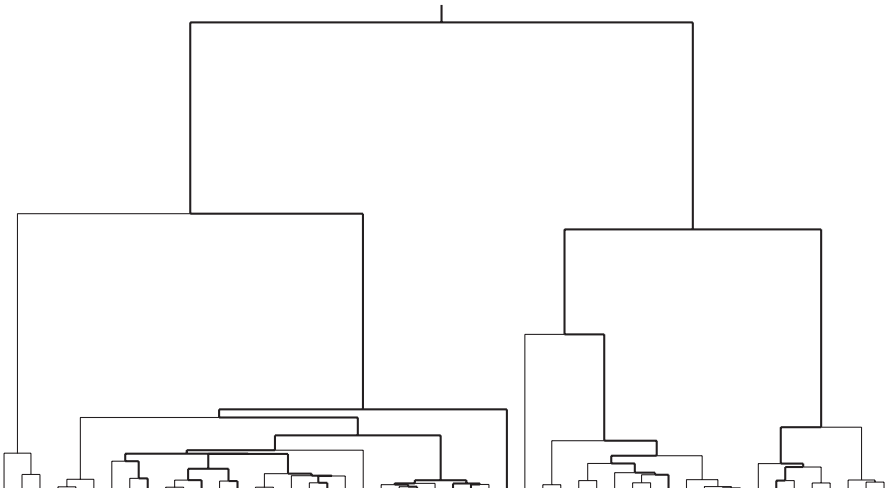

Figure 1.19 The effect of adding more sequences to a sample. A sample of ten sequences (in bold) are extended with forty more sequences. This mainly changes the lower part of the coalescent tree.

a fivefold increase in sample size. The effect of sampling more sequences will be taken up again in more detail in Section 3.4. For example, the probability that the smaller sample shares MRCA with the larger sample is derived.

1.10 The effective population size

As mentioned previously a real physical population is not likely to behave reproductively as the Wright–Fisher model. Most real populations show some form of reproductive structure, either due to geographical proximity of individuals (or genes) or due to social constraints (e.g. all females may not be allowed to marry any male, and vice versa). Also it is likely that the number of descendants of a gene in one generation does not follow the Poisson distribution with intensity one. If the real population has been of constant size over time then the mean number of descendants is bound to be one, but the variance of the distribution might differ from the variance in the Poisson (here one).

When the Wright–Fisher model, or the basic coalescent, is used to model a real physical population, the size ($2N$) of the population in the (haploid) Wright–Fisher model cannot be taken to be the size of the real population. For example, many human genes had a MRCA less than 200,000 years ago. If we count one generation as 20 years then N should be less than $200,000/(4 \cdot 20) = 2.500$ (the expected TMRCA of the whole population is $4N$), which is unrealistically small for the real population. However, it

suggests that the real population might be approximated by a Wright–Fisher model with $N = 2.500$.

For a real population or some model population (e.g. a diploid Wright–Fisher model) the population size of the haploid Wright–Fisher that in some sense best approximates the real population (or the model) is called the effective population size, N_e. Thus, in the example above N_e might be 2.500. There are several ways of defining N_e (e.g. see Ewens 2004). Often different definitions agree or agree approximatively for large values of N_e. Here we shall only be concerned with one definition of N_e and a generalisation of that definition that turns out to be particularly useful for models with variable population sizes (Chapter 4). The measure is called the *inbreeding effective population size*, which Ewens (2004) denoted by $N_e^{(i)}$ to distinguish it from other related quantities. Here it is just denoted by N_e. It is defined by

$$N_e = \frac{1}{2P(T_2 = 1)}, \tag{1.38}$$

where T_2 is in generations. The generalisation is defined by

$$N_e^{(t)} = \frac{E(T_2)}{2} \tag{1.39}$$

where t stands for time and T_2 again is in generations. The main and important difference between the two measures is that N_e is related to the immediate past (the previous generation), whereas $N_e^{(t)}$ is related to the number of generations until a MRCA is found. For the haploid Wright–Fisher model the two definitions agree. Here, $P(T_2 = 1) = 1/(2N)$ and $E(T_2) = 2N$. Thus, $N_e = N$ and $N_e^{(t)} = N$. In the diploid model with N_f females and N_m males ($N_\mathrm{f} + N_\mathrm{m} = N$),

$$P(T_2 = 1) = \left(1 - \frac{1}{2N}\right) \frac{N}{8N_\mathrm{f}N_\mathrm{m}} \tag{1.40}$$

and, thus,

$$N_e = N_e^{(t)} \approx \frac{4N_\mathrm{f}N_\mathrm{m}}{N} = 4c(1-c)N \tag{1.41}$$

for large N, $N_\mathrm{f} = cN$, and $N_\mathrm{m} = (1-c)N$. Again the two definitions agree. In particular, if $N_\mathrm{f} = N_\mathrm{m}$ ($c = \frac{1}{2}$) then $N_e = N_e^{(t)} = N$. If population size varies with time then the two definitions disagree because N_e depends on the particular generation we focus on.

To give a few other, more extreme, examples, assume that in each generation three genes are chosen and each creates a third of the next generation. The probability that two randomly chosen genes have the same

parent would then be a third, independently of population size. Then $N_e = N_e^{(t)} = 1.5$. Or assume that $\sqrt{2N}$ genes are chosen each to have $\sqrt{2N}$ offspring. The probability of having identical parents in the previous generation is then $1/\sqrt{2N}$ and $N_e = N_e^{(t)} = \sqrt{2N}/2$. These are two extreme examples.

Approximating the evolution of a real (or model) population by a haploid Wright–Fisher model with population size $2N = 2N_e$ implies that one particular aspect of the real population is mimicked in the haploid Wright–Fisher model, namely the expected TMRCA of two genes. Other quantities might differ, for example, the variance of the TMRCA, or the TMRCA of a sample of size n, and so on.

In Chapter 4 models with population structure and variable population sizes are introduced and the effective population size provides a standard by which these models can be compared.

1.11 The Moran model

Moran (1958*a*) proposed an alternative model to the Wright–Fisher model. Moran's model has overlapping generations, in contrast to the Wright–Fisher model that has non-overlapping generations. The population consists of $2N$ haploid individuals or genes. (A diploid version can be formulated as well. Here we focus on the haploid version for convenience.) A new generation is formed from the previous generation by sampling randomly one gene to give birth to a new gene, and one gene to die, see Figure 1.20. The gene that dies is not allowed to be the gene chosen to give birth. (Other formulations of the Moran model allow for the two genes to be one and the same.) All other genes survive to the next generation. The way the Moran model is constructed automatically rules out the possibility of multiple coalescent events in the same generation or that more than two genes share the same common ancestor in the previous generation.

The probability that two genes share a common ancestor in the previous generation is $1/(N(2N-1))$ because one out of the possible $\binom{2N}{2}$ pairs has the desired property. Thus the waiting time until a MRCA of two genes is a geometric distribution with parameter $1/(N(2N-1))$ and the natural time scale is in units of $N(2N-1)$ generations, rather than in units of

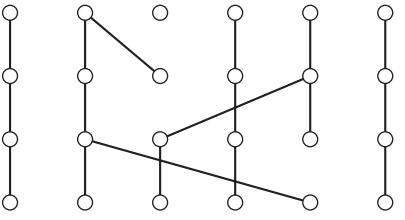

Figure 1.20 The Moran model. Shown is a few overlapping generations in the Moran model. In each generation one gene is chosen to give birth to a new gene and one gene is chosen to die. Because only one gene reproduces in each generation the appropriate time scale is different from that in the Wright–Fisher model.

$2N$ generations. When adjusted for the differences in time scale, the basic coalescent serves as an approximation for both models.

The Moran model is often applied by theoreticians because many calculations turn out to be particularly simple in this model, whereas a similar calculation in the Wright–Fisher model might be highly intractable. However, the Moran model appears to have less appeal to biologists and the Wright–Fisher model is thus more frequently used as a model of population reproduction.

1.12 Robustness of the coalescent

It has already been mentioned that the haploid and the diploid Wright–Fisher models, as well as the Moran model, tend to give probabilistically similar genealogies for large N. This is an example of what is called robustness of the coalescent: The continuous time coalescent can be used as an approximation of many different discrete time population reproduction models. Kingman showed that this is the rule rather than the exception. The effective population size defines the appropriate time scale and provides the transversion factor between discrete time units and continuous time.

Among the 1982 contributions from Kingman, this must be the most practical, in the sense that it legitimises the the use of the coalescent beyond the simple Wright–Fisher based models and the Moran model.

Recommended reading

Ewens, W. J. (2004), *Mathematical Population Genetics*, 2nd edn, Springer Verlag.

Hudson, R. R. (1991), 'Gene genealogies and the coalescent process', *Oxford Surveys in Evolutionary Biology* **7**, 1–49.

Kingman, J. F. C. (1982*b*), 'On the genealogy of large populations', *J. Appl. Prob.* **19A**, 27–43.

2 From genealogies to sequences

In the previous chapter we discussed how genes or sequences are related in a population through their common ancestry. However, to model real data a model of how mutations cause changes in the DNA is necessary. To meet this several mutation models have been developed. Historically, the infinite alleles model appeared first (Kimura and Crow 1964), followed by the infinite sites model (Kimura 1969), and the finite sites model (Jukes and Cantor 1969). This development illustrates an increased focus towards analysis of nucleotide sequence data sets, but also a development towards increasing complexity of analysis.

In this chapter we first introduce the three different types of mutation models, then a mathematical framework for working with probabilities of a sample configuration is discussed. We will mainly focus on the infinite alleles and infinite sites models and relate these models to the underlying genealogy of a sample. Mutations are assumed to be selectively neutral. This has the desirable effect that the mutation process can be separated from the genealogical process, because, in the absence of selection, the mutational process and the transmission of genes from one generation to the next are independent processes (genes have the same probability of transmission whatever their type). Thus, a sample configuration for n genes can be simulated using a two step procedure: (1) simulate the genealogy of n genes; (2) add mutations to the genealogy according to the chosen mutation model.

We will make use of forwards and backwards perspectives to calculate and explain quantities of interest. The backwards perspective implies that we are looking backwards in time: what could have happened prior to the present time? The forwards perspective implies that we are looking forwards in time: what could be the next event? We will frequently jump between the two perspectives. In the first two sections, 2.1 and 2.2, the time perspective is with the natural passage of time—from the past towards the present.

2.1 Mathematical models of alleles

2.1.1 The infinite alleles model

This model assumes that the information we have about alleles only allows us to say if they are identical or different. This was conceptually inspired

by isozymes, which are differently charged forms of an enzyme. There is no spatial or quantitative information about observed differences. A mutation in an existing allele will always create a new allele not observed before. If we observe alleles, the only information that would be available, would be which alleles are identical. For instance $(1, 4)(2)(3, 5, 7)(6)$ could be such an observation where the notation signifies that four types have been observed, 1 is identical to 4, and 3, 5 and 7 form an identical set. The unique types, 2 and 6 were only observed once. Having observed two different alleles, say 4 and 6, it is impossible to know whether 4 mutated into 6 through a single mutation, or 6 into 4, or whether several mutations separate one of the types from the other.

The actual labels (such as 4 and 6) are without allele specific information and have no biological interpretation. What matters are the sizes of the groups. In the present example that is: two groups of size 1, one of size 2, and one of size 3. This is often abbreviated as $1^2 2^1 3^1$, which is not a product, but a shorthand for the sizes of the sets observed. The superscript denotes the number of groups of a given size.

How this can be coupled to the evolution of a set of genes is described in Figure 2.1. The figure provides an informal and intuitive way of understanding the infinite alleles model. To be able to derive mathematical results for this model (and also the other mutation models to be introduced) we shall return to the Wright–Fisher model in Section 2.2 and explain how the Wright–Fisher model can be extended with a mutation process.

Let us assume that we can determine the types of these genes at any time point in the order they appear. In the beginning there would only be

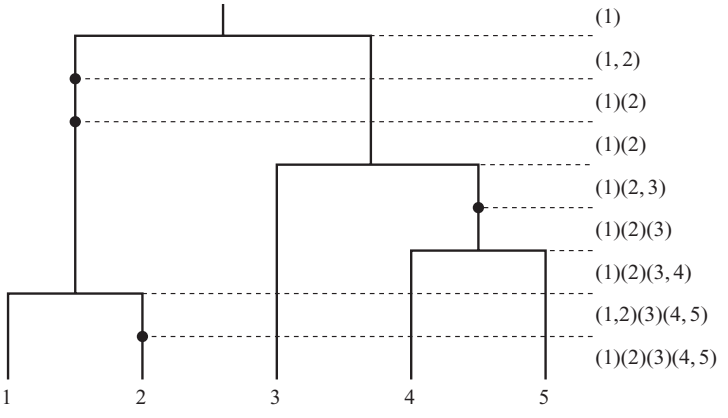

Figure 2.1 Infinite alleles model. Only the set size notation is shown here. The sample's history is followed from the root of the tree towards the present day. The configurations shown to the right of the tree represent the sample configuration as it is between dotted lines. The bullet '•' denotes a mutation event. The labels attached to the genes are those used in the configuration lowest in the figure.

one type (1) or 1^1. The direction of time is here from the past towards the present, from the root towards the present day sample. A split event (a coalescent event when going backwards in time) would then create two identical types, 2^1. A mutation would then modify one of these, creating a different type, leaving two different alleles in one copy each, 1^2. A split event will in the set notation take an allele and duplicate it. In the set size notation, it will remove a set of size k and add a set of size $k+1$. A mutation event will remove an allele and create a completely new type. In the set size notation, a set of size k will be removed and two sets of sizes 1 and $k-1$ will be included. (If $k=1$ the configuration is left unchanged.) In Figure 2.1, $(1)(2,3)$, or $1^1 2^1$, becomes $(1)(2)(3)$, or 1^3, such that two sets of size one are created and the set of size two is removed. Also $(1)(2)$, or 1^2, becomes $(1)(2,3)$, or $1^1 2^1$. It is important to understand that the numbers are arbitrary labels and that they do not carry any information about the alleles themselves. For example, $(1)(2)$ remains $(1)(2)$ after the second mutation event counted from the MRCA, and $(1)(2)(3,4)$ becomes $(1,2)(3)(4,5)$ after the third split event.

2.1.2 The infinite sites model

The infinite sites model can be interpreted as describing the evolution of very long DNA strings with low mutation rate at each position. (Therefore, genes are frequently denoted as sequences in this model.) As a consequence a mutation will always happen in a new position. Biologically, the model is justified using the following argument: The number of variable sites in a sample of real sequences is typically small compared to the number of sites which are identical in all sequences. Further, often only one or two nucleotides are found in a given position, indicating that mutations happen rarely. The latter point is of course questionable because mutations in a given position could preferably be between two specific nucleotides or from one specific nucleotide into another.

In the infinite sites model there will always be one or two states in a position of a set of sequences, never more, because each position mutates at most once. All that matters in comparison of two sequences at a given position is whether they are different or not. The two possible alleles are labelled 0 and 1 with no biological meaning attached to the labels. If labels are swapped the interpretation will remain the same. The model also implies that all mutations that have happened in the history of a sample are recoverable, in contrast to the infinite alleles model, where two consecutive mutations in one lineage will only be registered as one overall change. The position of a mutation is random along the sequence.

It is convenient to represent a sequence as a series of zeros and ones, leaving out all positions that are not segregating in the sample. Attach to each segregating site its position in the sequence. An example is given in

0.073	0.294	0.550	0.894
1	1	0	0
1	1	0	1
0	0	0	0
0	0	1	0
0	0	1	0

Figure 2.2 Representation of infinite sites data. Shown is a sample of size five with four segregating sites. The top row shows the positions, relative to the length of the sequence, as decimal numbers. All positions that are invariable are left out. A history of this sample is shown in Figure 2.3.

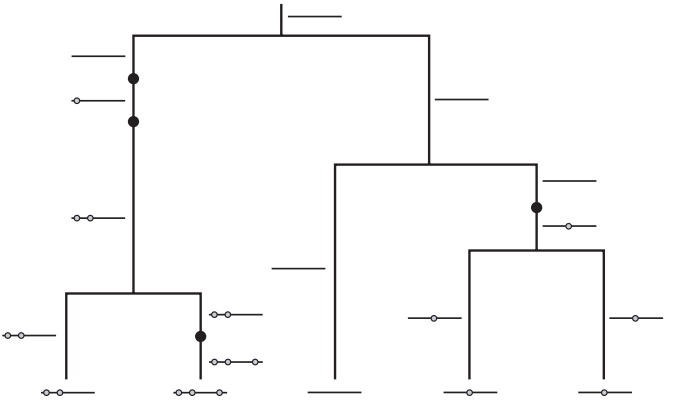

Figure 2.3 Infinite sites model. The sequence is represented by a continuous string with mutations attached. In the example the circle denotes the derived state. However this information is only available if we know the ancestral states of all positions, that is if we know the root sequence. If another combination of states was used as root sequence the tree might not be compatible with the sequence data. All mutations are visible in the present day sequences.

Figure 2.2. Figure 2.3 illustrates an evolutionary history of the sample in Figure 2.2 using a different, but commonly applied, representation of a sequence. The infinite sites model is also discussed in the context of the Wright–Fisher model later in this chapter and some consequences of the infinite sites model will be derived formally.

A series of points should be noted at this stage. A mutation always partitions the set of sequences into two groups. These two groups correspond to an edge on the true tree relating the sequences, in the sense that if this edge was removed the tree would be cut into two trees where all leaves on the first tree shared one character state, while leaves on the other tree shared another character state. In other words, a mutation induces a bipartition of the sample. If the mutation rate was sufficiently high, the true unrooted tree topology of the sequences can be inferred because mutations will be dense on every branch and all sequences in the sample will thus have their own distinctive patterns of mutations. The number of mutations in the two histories in Figures 2.1 and 2.3 are the same, but in the infinite sites example, it is apparent that four mutations have occurred, while in the infinite alleles example, the two mutations on the same branch will only be visible as one

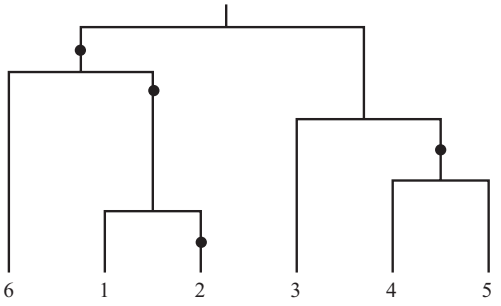

Figure 2.4 Sampling and the infinite alleles model. After sampling of allele 6 both mutations on the branch connecting the ancestor of 2 and 3 with the MRCA become visible in the sample.

mutation, unless sequences were sampled that contained the first but not the second of the two mutations (as in Figure 2.4).

2.1.3 Finite sites model

Modelling has moved towards increased realism. A completely realistic model of DNA evolution would imply a complete specification of the DNA sequence. In population studies such a sequence will have a fixed length and will be assumed to evolve only by mutations. In principle a model should contain a process describing insertions and deletions, but these occur so rarely in population data that sequences normally can be unambiguously aligned, even if insertions/deletions have occurred. One might then choose to discard insertions and deletions.

In the example of Figure 2.5, an original sequence of length eight encountered mutations over time as it evolved. In this series of events, two mutations happened in the same position (position 7), which would be prohibited under the infinite sites model. So four mutations only created three segregating sites in this case. Mutations happening in the same position (but not necessarily in the same lineage) are called recurrent mutations. They create two kinds of partitions that are not possible under the infinite sites model: (1) partitions based on three or four types, as induced by the seventh position in Figure 2.5; (2) bipartitions that do not correspond to a partition defined by a branch in the tree, because of two (or more) mutations in the same direction (e.g. T $\rightarrow$ A twice) or in opposite directions (e.g. T $\rightarrow$ A and A $\rightarrow$ T in different branches). In addition there is the possibility of 'back-mutation', a mutation that erases the effect of the first mutation, for example, T $\rightarrow$ A and subsequently A $\rightarrow$ T in the same branch.

The simplest finite sites model is the Jukes–Cantor model, named JC69, introduced by Jukes and Cantor in 1969: All positions are equally likely to mutate and the mutant is chosen with equal probability among the three

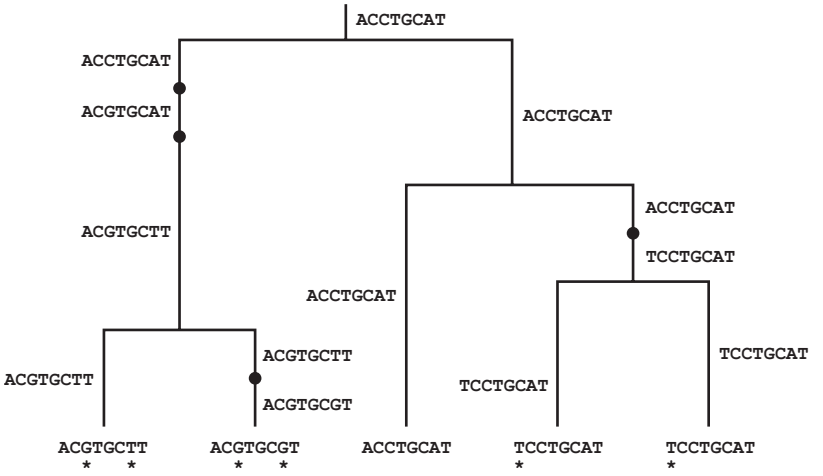

Figure 2.5 Finite sites model. Each nucleotide position might experience more than one mutation. Two sequences can be identical even if several mutations have occurred in their history: Forth and back mutations can erase the stamp of mutations, for example, if A → T at some point back in time and T → A in the same position but closer to the present time.

possible nucleotides. This model was later modified by Kimura (1980) to accommodate the fact that transition events (A ↔ T and C ↔ G) occur at a faster rate than transversion events (all other events). It is called K80. Both JC69 and K80 are unrealistic in the sense that all nucleotides are expected to occur with the same frequency (viz. 0.25) in a random sequence, which is not likely to be the case for any sequence. Felsenstein (1981) modified JC69 to take into account unequal base frequencies (F81), and Hasagawa et al. (1985) combined F81 and K80 to allow for unequal base frequencies and transition/transversion bias at the same time. Still other more sophisticated models have been introduced to account for subtle differences in substitution rates.

An important assumption of the finite sites models discussed above is that positions evolve independently of each other. (This is also an assumption of the infinite sites model because a mutation in one position does not influence the chance of a mutation in another position.) It is possible to model coupled or dependent evolution between neighbouring positions. Such models become increasingly difficult to analyse though they might be more appropriate for the analysis of many real sequences. One special kind of such models are codon models. The biological fact that DNA sequences are translated into proteins creates an additional level of complexity. Each triplet of nucleotides (a *codon*, 64 in total) codes for a specific amino acid that differs from other amino acids in respect of chemical and physical properties. When coding DNA evolves, a substitution might or might not change the encoded amino acid. An amino acid change might have

functional or phenotypical consequences and thereby be of severe importance for the cell's or organism's ability to function and survive. Ideally, a model of DNA sequence evolution should incorporate information about the encoded amino acids. However, a population data set rarely shows more than a few differing amino acids and it becomes practically impossible to perform inferences in such models because of the huge number of model parameters compared to the sparse data.

In the infinite alleles model and the infinite sites model the mutation rate can be taken as an overall rate for all types of events that change the type of a sequence: Thus the rate can be taken to comprise both the rate of nucleotide mutations and the rate of insertions and deletions. All that matters is that a mutation, insertion, or deletion introduces a change in the type of the gene and that this change is unique to the sample. In contrast, only nucleotide substitutions are modelled in the finite sites model.

2.2 The Wright–Fisher model with mutation

The Wright–Fisher model was introduced in Chapter 1 as a purely reproductive model without any information about genetic type. We can impose a process of mutation on top of the process of reproduction in the following simple way: Each gene chosen to be passed on is subject to a mutation event with probability u. That is to say, with probability $1 - u$ the gene is copied without modification (or error) to the offspring and with probability u a mutation occurs. Under the infinite alleles model a new type not previously seen in the population is introduced, under the infinite sites model a position along the gene is randomly chosen and the type of that position is changed (from zero to one, or vice versa), and under the finite sites model a site is chosen randomly and a mutation occurs in that position according to a given model. Thus, the overall structure of the mutation process is the same irrespective of whether an infinite alleles model, an infinite sites model or a finite sites model is assumed.

In Figure 2.6 three generations are shown. All mutations are selectively neutral in the sense that the type of the parent does not influence the

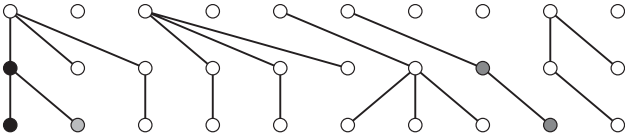

Figure 2.6 The Wright–Fisher model with mutation. Shown are three generations of the Wright–Fisher model with mutation. In the second generation, counted from the top row, two descendants of the previous generation are mutated copies of their parent gene. One descendant of the leftmost mutant in generation two is again a mutated copy of its parent.

probability that a mutation occurs or that the offspring survives. If we follow a lineage from the present time into the past (e.g. the lineage defined by the leftmost gene in Figure 2.6) then there is a chance, u, that the type of the parental gene in generation t differs from the type of the offspring gene in generation $t+1$. Consequently, the probability that a lineage experiences the first mutation j generations into the past, counting from the present time, is

$$P(T_M = j) = u(1-u)^{j-1}. \tag{2.1}$$

Here T_M denotes the number of generations until the first mutation event. It follows that T_M is a geometric variable with parameter u. If time is measured in units of $2N$ generations (as in the basic coalescent process) then

$$P(T_M \leq j) = 1-(1-u)^j \approx 1-e^{-\theta t/2} = P(T_M^c \leq t), \tag{2.2}$$

for large $2N$, where $t = j/(2N)$, $\theta = 4Nu$, and T_M^c denotes time in units of $2N$ generations. The parameter θ is called the population mutation rate or the scaled mutation rate. There are no fundamental reasons why θ is defined as $4Nu$ rather than $2Nu$, but some formulas turn out to simplify when scaled in $4N$, rather than in $2N$. The parameter θ can be interpreted as the expected number of mutations separating a sample of two sequences, since the expected coalescence time for two sequences is $2N$, and thus $2Nu$ mutations are expected on each branch. It transpires that T_M^c is an exponential variable with parameter $\theta/2$. Using T_M^c as a continuous time approximation to T_M is accurate as long as N is large. Henceforth we focus exclusively on the continuous time approximation. (Superscript c in T_M^c is dropped when ambiguities are not possible.)

If we instead consider n disjoint lineages then the time until the first mutation event in any of the n lineages is exponential with parameter $n\theta/2$, that is, n times that of a single lineage because lineages evolve independently of each other. Whether there is a mutation in one gene in one generation is independent of whether there is another mutation in another gene in the same generation. The form of the parameter $n\theta/2$ is now a consequence of property (1.18).

If we wait for mutation events and coalescence events then the parameter of the exponentially distributed waiting time is the sum of the two parameters. Again this is a consequence of property (1.18), because the two types of events are independent of each other. It follows that the parameter is

$$\binom{n}{2} + \frac{n\theta}{2} = \frac{n(n-1+\theta)}{2}. \tag{2.3}$$

Whether the first event is a coalescence event or a mutation event is determined by tossing a biased coin. With probability

$$\frac{\binom{n}{2}}{\binom{n}{2}+\frac{n\theta}{2}} = \frac{n-1}{n-1+\theta}, \tag{2.4}$$

the event is a coalescence event and with probability

$$1 - \frac{n-1}{n-1+\theta} = \frac{\theta}{n-1+\theta}, \tag{2.5}$$

the event is a mutation event. The pair that merges is chosen randomly among all pairs and the lineage undergoing mutation is chosen randomly among the n lineages.

2.3 Algorithms for simulating sequence evolution

This section deals with simulation of sample histories and simulation of sample configurations, two closely related subjects. Figures 2.4 and 2.5 provide examples of sample histories. A sample history is a sequence of dated events that created the sampled genes from a gene at the root. The sample configuration is obtained by discarding the history, keeping the configuration at the time of sampling. Naturally, if we can simulate sample histories, we can also simulate sample configurations.

Simulation of sample configurations and sample histories are of importance for many reasons. First of all, it provides means to study variation in random samples by repeated generation of sample configurations. Many quantities of interest and distributions of random variables cannot be found explicitly and we must resort to simulation. Second, it gives intuition into the dynamics of the coalescent and the mutation processes. And finally, efficient sampling of sample histories becomes a major issue for inferential procedures (Chapter 6).

Equations (2.3)–(2.5) form the basis of the first algorithm for simulating a set of genes (or sequences) with mutations.

Algorithm 2

1. Put $k = n$, where n is the sample size.
2. Choose an exponential variable with parameter $k(k-1+\theta)/2$.
3. With probability $(k-1)/(k-1+\theta)$ the event is a coalescence event and with probability $\theta/(k-1+\theta)$ it is a mutation event.

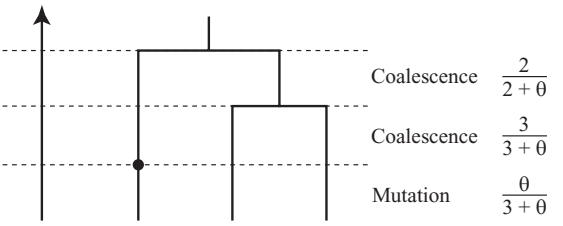

Figure 2.7 Algorithm 2. Simulation of a sample history for $n = 4$ genes. The first event is a coalescence event with probability $(n-1)/(n-1+\theta) = 3/(3+\theta)$ and a mutation event with probability $\theta/(3+\theta)$. In this example the event was a mutation event. The algorithm continues until there is only one lineage in the sample. To the right is shown the type of the event and the probability that it occurred.

4. If a coalescence event occurs choose a pair randomly to coalesce. Update k, $k \to k-1$.
5. If a mutation event occurs choose a lineage to mutate. Leave k unchanged.
6. Continue until k is one.

The algorithm is an extension of Algorithm 1 given in Chapter 1. It is illustrated in Figure 2.7. To determine the types of the present day genes (or sequences) start at the MRCA and move forward in time. Each time a mutation event is encountered the type of the gene is modified according to the chosen mutation model, as illustrated in Figures 2.4 and 2.5. Each time a split event is encountered copy the parent gene onto both descending lineages.

The second algorithm derives from the fact that the waiting time until a mutation occurs along a lineage is exponential with parameter $\theta/2$. This is mathematically equivalent to the following. Consider a branch of length t. The number, M_t, of mutations on the branch is Poisson distributed with intensity $t\theta/2$, that is,

$$P(M_t = j) = \frac{(t\theta)^j}{j!2^j} e^{-t\theta/2}. \tag{2.6}$$

In particular the mean of equation (2.6) is $t\theta/2$. Given the number of mutations on a branch the time of each mutation is random. The numbers and times of mutations on different branches are independent of each other such that branches can be treated one at a time. The process placing mutations on the branches according to equation (2.6) is called a Poisson process.

Algorithm 3

1. Simulate the genealogy of n sequences according to the coalescent process with rate $\binom{k}{2}$ while there are k lineages, that is, following Algorithm 1.

2. For each branch draw a number, M_t, from a Poisson distribution with intensity $t\theta/2$, where t is the length of the branch.
3. For each branch the times of the M_t mutation events are chosen randomly on the branch.

Algorithm 3 uses explicitly that mutations can be tossed onto the genealogy after the genealogy has been simulated. This is in fact an extremely useful property because it makes it easy to generalise Algorithm 3 to other scenarios by changing the way the genealogy is constructed (see Chapter 4). As long as the mutation process is neutral, mutations can be simulated according to the Poisson process and added to the genealogy in a second step. Figure 2.8 illustrates the algorithm.

If we are not interested in the sample histories as such, but just the configuration of the sample, Algorithm 3 can be simplified. Instead of simulating all mutation events on a particular branch (of length t) we can simulate the allelic state of the gene at the end of the branch given the allelic state of the gene at the beginning of the branch. Under the infinite alleles model this amounts to calculating the probability of at least one mutation over time t which is one minus the probability of no mutations, that is, $1 - e^{-t\theta/2}$, and simulate a binary random variable: With probability $1 - e^{-t\theta/2}$ the gene is mutated, and with probability $e^{-t\theta/2}$ the type of the gene is left unchanged.

Under the infinite sites model we simulate a Poisson number of mutations, M_t, and add them to the sequence at the end of the branch. The position of a mutation is chosen randomly along the sequence. (Note that for a given branch of length t with M_t mutations the time of an event is chosen randomly on the branch and the position of a mutation is chosen randomly on the length of the sequence.)

Finally under the finite sites model we focus first on a single nucleotide, instead of the whole sequence. The rate of mutations for a single nucleotide

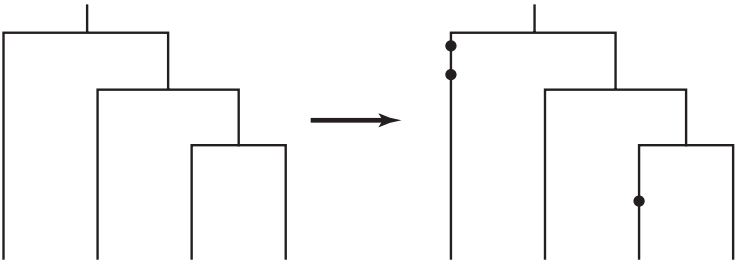

Figure 2.8 Algorithm 3. Two step simulation of sample histories. First, the genealogy of the sample is simulated. Second, mutations are put onto each branch according to a Poisson process with intensity $\theta t/2$, where t is the length of the branch. Mutations are randomly spaced on the branch. The sample configuration is determined by starting at the root moving forward in time and modifying the type of the genes as mutations are encountered.

is $\theta_0 = \theta/L$ if there are L nucleotides. The probability that a nucleotide is in state j given it is in state i, t time units in the past, is

$$P_{ij}(t) = 0.25\left\{1 - e^{-2t\theta_0/3}\right\}, \tag{2.7}$$

if $i \neq j$, where $i, j \in \{\mathrm{A}, \mathrm{G}, \mathrm{T}, \mathrm{C}\}$, and

$$P_{ii}(t) = 0.25\left\{1 + 3e^{-2t\theta_0/3}\right\}, \tag{2.8}$$

(see also Figure 2.9). Here we have assumed a Jukes–Cantor model such that each time a mutation happens the new nucleotide is equally likely to be any of the other three nucleotides.

If a sequence is L nucleotides long, the probability that the sequence evolves into some other sequence after time t is obtained as the product of L terms, one for each nucleotide. In particular, the probability that two sequences differ in d positions has probability

$$P(D_t = d) = \binom{L}{d}\left(0.75\left\{1 - e^{-2t\theta_0/3}\right\}\right)^d\left(0.25\left\{1 + 3e^{-2t\theta_0/3}\right\}\right)^{L-d}, \tag{2.9}$$

where D_t is the number of differences after time t and the binomial coefficient is the number of ways d nucleotides can be chosen out of L possible. The positions of the d differences are chosen randomly among the L possible positions.

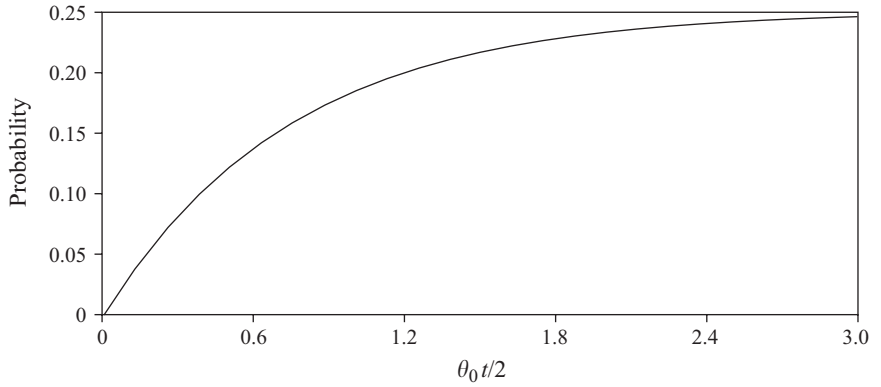

Figure 2.9 Finite sites model. The probability that a nucleotide has changed to a specific different nucleotide as a function of $\theta_0 t/2$. This probability starts as being 0, but then converges from below to 0.25 as $\theta_0 t/2$ becomes large.