Chapter 3

The Coalescent

To coalesce means to grow together, to join, or to fuse. When two copies of a gene are descended from a common ancestor which gave rise to them in some past generation, looking back we say that they coalesce in that generation. Seen forward in time, coalescent events are simply DNA replication events, and are only of special interest due to their place in the history of a particular sample. Kingman (1982a,b) showed that the joining up of lineages into common ancestors is described by a particular mathematical process, and he called this process the n-coalescent. Here we will see how Kingman's coalescent arises in the context of the two most commonly applied models of a population, the Wright-Fisher model and the Moran model, and discuss its applicability to a host of other models. From section 1.1 we have some familiarity with genealogies and their structure. In this chapter the coalescent genealogy of a sample is considered without reference to any observed variation in the sequences. This is possible, first because every sample of gene copies has a genealogy even if it displays no variation. Second, for the moment we assume that all variation is selectively neutral. By definition, this means that an individual's genotype has no effect on the number of descendents it leaves, and thus no effect on the genealogy of a sample. Much of the simplicity and elegance of the coalescent approach stems from the fact that, when variation is neutral, the genealogical process and the mutational process are independent and can be considered separately. Mutations and genetic data are the subject of Chapter 4.

3.1 Population Genetic Models

Theoretical studies of the genetics of populations rely on our ability to construct models which capture the essential biological features of populations but which are idealized enough to be mathematically tractable. Two such models have been the basis of most work in population genetics: the Wright-Fisher model and the Moran model. Neither of these was developed to fit the known biology of any particular organism. However, both are members of a broad class of models that describe many different breeding structures and which encompass a range of biologically reasonable assumptions about populations. Importantly, all of these models yield the coalescent under certain limiting conditions. The Wright-Fisher model represents a case of perfectly non-overlapping generations and the Moran model represents an idealized case of overlapping generations. Real populations might exist somewhere between these two extremes. The coalescent is an approximation to the ancestral process for a sample under the Moran model and the Wright-Fisher model when the population size is large, although some features of the coalescent are exact for the Moran model. We begin with a forward-time description of these two models, then later consider how the ancestral process is obtained.

3.1.1 The Wright-Fisher Model

The model introduced by Fisher (1930) and Wright (1931) assumes that all of the individuals in the population die each generation and are replaced by offspring. The population size N is assumed to be constant over time and finite. Because the population is finite in size and reproduction is a random process, some individuals may not contribute any offspring to the next generation. This random loss of genetic lineages forward in time is called genetic drift. Backward in time it is the source of the coalescent process. The Wright-Fisher model can be applied to haploid organisms, in which case the population will consist of N copies of the genome, or to diploid organisms, in which case there will be $2N$ copies. Assuming a diploid organism is probably the most common convention, but the coalescent best viewed at the start as a haploid model. In fact, many apparently diploid models can be reduced to haploid models, the exception being when diploidy has direct consequences on the dynamics of the population, such as when diploid migration occurs or when alleles exhibit dominance under natural selection. In most of what follows, we will assume a haploid organism. We will consider the applicability of the coalescent to diploid organisms in Chapter 7, but note here that it applies to diploids just as well as any other neutral population genetic model if we simply replace N below with $2N$.

The Wright-Fisher model assumes that the ancestors of the present generation are obtained by random sampling with replacement from the previous generation. Looking forward in time, consider the familiar starting point of classical population genetics: two alleles, A and a, segregating in the population. Let i be the number of copies of allele A, so that $N - i$ is the number of copies of allele a. Thus the current frequency of A in the population is $p = i/N$, and the current frequency of a is $1 - p$. We assume that there is no difference in fitness between the two alleles, that the population is not subdivided, and that mutations do not occur. This gives the familiar formula,

$$P_{ij} = \binom{N}{j} p^j (1-p)^{N-j} \qquad 0 \leq j \leq N, \tag{3.1}$$

for the probability that a gene with i copies in the present generation is found in j copies in the next generation. Let the current generation be generation zero and K_t represent the counts of allele A in future generations. Equation 3.1 states that K_1 is binomially distributed with parameters N and $p = i/N$, given $K_0 = i$. Therefore, from (2.39) and (2.40) we have

$$E[K_1] = Np = i, \tag{3.2}$$

$$\mathrm{Var}[K_1] = Np(1-p). \tag{3.3}$$

The number of copies of A is expected to remain the same on average, but in fact may take any value from zero to N. A particular variant may become extinct (go to zero copies) or fix (go to N copies) in the population even in a single generation. Over time, the frequency of A will drift randomly according to the Markov chain with transition probabilities given by equation 3.1, and eventually one or the other allele will be lost from the population. Ewens (2004) gives an excellent treatment of the forward-time dynamics of this model.

Perhaps the easiest way to see 3.1 is through a biologically motivated example. Imagine that before dying each individual in the population produces a very large number of gametes. However, the population size is tightly controlled so that only N of these can be admitted into the next generation. The frequency of allele A in the gamete pool will be i/N, and because there are no fitness differences, the next generation is obtained by randomly choosing N alleles. The connection to the binomial distribution, as discussed above in Section 2.1.2 is clear: we perform N trials, each with $p = i/N$ chance of success. Because the gamete pool is so large, it is not

depleted by this sampling, so the probability i/N is the same for each trial. The distribution of the number of A alleles in the next generation is binomial$(N, i/N)$ as equation 3.1 indicates.

Before we take up the backward, ancestral process for the Wright-Fisher model in the next chapter, we will use a classical derivation and result to see the difference in rates of genetic drift between this model and the Moran model described below. The heterozygosity of a population is defined to be the probability that two randomly sampled gene copies are different. For a randomly mating diploid population, this is equivalent to the chance that an individual is heterozygous at a locus. Let the current generation be generation zero, and let p_0 be the frequency of A now. The heterozygosity of the population now is equal to $H_0 = 2p_0(1-p_0)$, which is just the binomial chance that one allele A (and one a) is chosen in two random draws. Let the random variable P_t represent the frequencies of A in each future generation t. Then in the next generation the heterozygosity will be $H_1 = 2P_1(1-P_1)$. However, H_1 will vary depending on the random realization of the process of genetic drift described by equation 3.1. On average,

$$\begin{aligned} E[H_1] &= E[2P_1(1-P_1)] \\ &= 2\left(E[P_1] - E[P_1]^2 - \mathrm{Var}[P_1]\right) \\ &= 2p_0(1-p_0)\left(1-\frac{1}{N}\right) \\ &= H_0\left(1-\frac{1}{N}\right), \end{aligned}$$

and this shows that heterozygosity is lost through genetic drift. The derivation above uses $P_1 = K_1/N$ together with equations 3.2 and 3.3, and the simple rules of Section 2.1, such as equation 2.19. After t generations, we have

$$\begin{aligned} E[H_t] &= H_0\left(1-\frac{1}{N}\right)^t \\ &\approx H_0\, e^{-t/N} \end{aligned} \tag{3.4}$$

with the approximation being valid for large N (see equation 2.48). In the Wright-Fisher model, heterozygosity decays at rate $1/N$ per generation. The decrease of heterozygosity is a common measure of genetic drift, and we say that the drift occurs in the Wright-Fisher model at rate $1/N$ per generation.

3.1.2 The Moran Model

The Wright-Fisher model is the one most widely used in population genetics, but another model, due to Moran (1958,1962) is also very well studied. The Moran model has been important for two reasons. First, in contrast to the Wright-Fisher model, it applies to organisms in which generations are overlapping. Second, it has been important from the mathematical point of view, because many results can be derived exactly under the Moran model that are available only approximately under the Wright-Fisher model.

The Moran model is formulated with haploid organisms explicitly in mind, and again we assume that the population size is N. In this model, at times $t = 0, 1, 2, \ldots$, two individuals are chosen at random with replacement from the population. These might be the same or they

might be different individuals. Each individual in the population has a $1/N$ chance of being chosen in each draw. The first individual chosen reproduces, *i.e.* copies itself, and the second one dies. Thus, if the same individual was chosen twice, it would reproduce itself then die and the state of the population would not change. Again let there be i copies of allele A and $N-i$ copies of allele a, and let j be the number of copies of allele A after one time unit. Now K_1 can assume only three possible values: $i+1$, i, and $i-1$. The probability that i increases is equal to the probability that an a allele is chosen to die times the probability that an A allele is chosen to reproduce. Again using $p = i/N$, and continuing this line of reasoning to the other two possible transitions gives

$$P_{ij} = \begin{cases} p(1-p) & \text{if } j = i+1, \\ p(1-p) & \text{if } j = i-1, \\ p^2 + (1-p)^2 & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, in contrast to a Wright-Fisher population, under the Moran model one of just three things must happen in one time unit: allele A increases in number by one, allele a increases in number by one, or the counts stay the same.

From this is not difficult to compute the expectation and variance of K_1 directly using equations 2.10 and 2.11 and with $i = Np$:

$$\begin{aligned} E[K_1] &= (Np+1)p(1-p) + (Np-1)p(1-p) + Np\left[p^2 + (1-p)^2\right] \\ &= Np\left[p(1-p) + p(1-p) + p^2 + (1-p)^2\right] \\ &= Np \qquad (3.5) \end{aligned}$$

$$\begin{aligned} \mathrm{Var}[K_1] &= (1)^2 p(1-p) + (-1)^2 p(1-p) + (0)^2\left[p^2 + (1-p)^2\right] \\ &= 2p(1-p). \qquad (3.6) \end{aligned}$$

As in the Wright-Fisher model, random genetic drift leads to variation in the number of copies of A, but since it is unbiased, the expected number in the next generation is equal to the number in the current generation.

Using these equations and considering the heterozygosity of the population, after one time unit,

$$\begin{aligned} E[H_1] &= E[2P_1(1-P_1)] \\ &= 2p_0(1-p_0)\left(1 - \frac{2}{N^2}\right) \\ &= H_0\left(1 - \frac{2}{N^2}\right) \end{aligned}$$

After t time units , we have

$$\begin{aligned} E[H_t] &= H_0\left(1-\frac{2}{N^2}\right)^t \\ &\approx H_0\, e^{-2t/N^2} \end{aligned} \tag{3.7}$$

Thus the rate of genetic drift per time unit in the Moran model is equal to $2/N^2$.

To make this comparable to drift in the Wright-Fisher model, we can define a generation under the Moran model to be equal to N steps, or birth-death events. Looked at from the point of view of an individual this makes sense as well. The probability that a particular individual dies in one time unit is $1/N$, so the lifetime of an individual is geometrically distributed with parameter $1/N$ (see equation 2.41). From 2.44 we can see that the lifetime of an individual has mean N steps, so it is natural to interpret this as one generation. If we rescale time accordingly by defining $\tau = t/N$, equation 3.7 becomes

$$E[H_\tau] \approx H_0\, e^{-2\tau/N} \tag{3.8}$$

Comparison to equation 3.4 shows that, with equivalent definitions of a generation, the rate of genetic drift is twice as fast in the Moran model as it is in the Wright-Fisher model. This is interesting from a biological standpoint because it means that differences in breeding structure can lead to differences in time scale of change in the population even though the way in which it changes (*e.g.* exponential decay as above) may be the same for different kinds of populations. This factor of two increase in the rate of drift in the Moran model is not a consequence of generations being overlapping. It is due, instead, to differences the distribution of offspring number among individuals in the population under Wright-Fisher-type versus Moran-type reproduction (Moran and Watterson, 1959; Feldman, 1966), as we will see in Section 3.2.3 below.

3.2 The Standard Coalescent Model

We begin with the simplest statement of the coalescent model. Kingman (1982a,b,c) proved this to be limiting ancestral process for a broad class of populations structures that includes the Wright-Fisher model and the Moran model. We trace the ancestral lineages, which are the series of genetic ancestors of the samples at a locus, back through time. The history of a sample of size n comprises $n-1$ coalescent events. Each coalescent event decreases the number of ancestral lineages by one. This takes the sample from the present day when there are n lineages through a series of steps in which the number of lineages decreases from n to $n-1$, then from $n-1$ to $n-2$, *etc.*, then finally from two to one. The single lineage remaining at the final coalescent event is the most recent common ancestor (MRCA) of the entire sample. At each coalescent event, two of the lineages fuse into one common-ancestral lineage. The result is a bifurcating tree like the one shown in figure 3.1. The times T_i on the right in figure 3.1 are the times during which there were exactly i lineages ancestral to the sample.

Thus, the coalescent is a stochastic process, like the ones considered in Chapter 2, only a little more complicated because it includes both a discrete tree structure and $n-1$ coalescence time intervals. The state space of genealogies is the set of all possible rooted birfurcating trees with labelled tips and nodes ordered in time, with coalescence times $0 < T_i < \infty$ for $2 \le i \le n$. Any particular genealogy, or realization of the coalescent process, will specify the branching pattern of relationships among the members of the sample and the coalescence times. Genealogies can

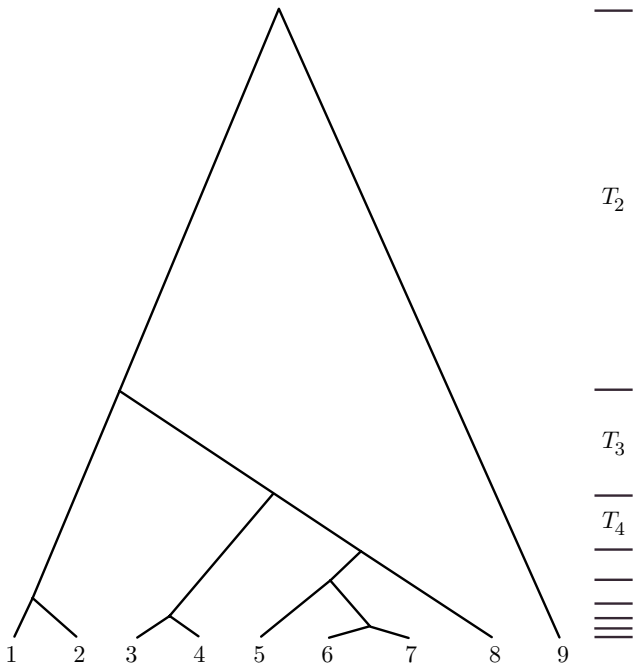


Figure 3.1: A coalescent genealogy of a sample of $n = 9$ items.

provide information about the population from which the sample was taken just as successive coin tosses provide information about the properties, *e.g.* fairness, of a coin. Thus, genealogies must be treated in a statistical setting. Unlike the result of a coin toss, however, genealogies cannot be observed directly. Information about genealogical history is inferred from patterns of polymorphism in a sample, which in turn result from another random process: mutation (see Chapter 4).

With a short but far-reaching list of assumptions about the population, it is possible to describe the probability distributions of both genealogical trees and coalescence times. These assumptions are:

1. Genetic differences have no consequences on fitness.
2. The population is not subdivided, geographically or otherwise.
3. The size of the population is constant over time.

The first and second assumptions above — that all genetic variation is selectively neutral and that the population is well-mixed, or panmictic — are two aspects of what is probably better veiwed as a single assumption. Namely, it is assumed that the number of offspring an individual has is independent of any labels that might be assigned to it, *e.g.* its allelic state or its geographic location. This can be seen clearly in sections 3.1.1 and 3.1.2, for the case of allelic states as labels. We say that the numbers of offspring among individuals in the population are exchangeable random variables. Exchangeability means identically distributed but not necessarily independent; for details see Kingman (1982c) and Aldous (1985). We will take a closer look at exchangeability in Section 3.2.3, but note for now that the non-independence of the numbers of offspring in the population is a consequence of the third assumption above, that the total number of offspring is fixed.

Kingman (1982a,b) showed that in the limit as N goes to infinity, the coalescence times T_i are independent and exponentially distributed as

$$f_{T_i}(t_i) = \binom{i}{2} e^{-\binom{i}{2} t_i} \qquad t_i \geq 0, \qquad i = 2, \ldots, n \tag{3.9}$$

when time is measured appropriately. In the next two sections, we will see what the appropriate units of time are under the Wright-Fisher and Moran models. Because they are exponentially distributed, the mean and the variance (see equation 2.51) of the times to coalescence are

$$E[T_i] = \frac{2}{i(i-1)}, \tag{3.10}$$

$$\mathrm{Var}[T_i] = \left(\frac{2}{i(i-1)}\right)^2. \tag{3.11}$$

From equation 3.10, it is clear that the most ancient coalescence time, the one in which the remaining two lineages coalesce into the MRCA of the entire sample, is expected to be the longest. The coalescence times in figure 3.1 are drawn in proportion to their expected values. Especially in a large sample, many coalescent events will occur over a very short period of time in the recent history of the sample. Because the coalescence times are mutually independent, we have

$$f_{T_n,\ldots,T_2}(t_n, \ldots, t_2) = \prod_{i=2}^{n} f_{T_i}(t_i). \tag{3.12}$$

In addition, at each coalescent event, every pair of lineages is equally likely to be the pair that coalesces. This means that every possible genealogical tree structure is equally likely. All of the remarkable results of the standard coalescent model follow directly from these two properties: the random-joining or random-bifurcating nature of coalescent trees, and the independent, exponential coalescence times.

The formal proof of the above statements for a general, exchangeable population model is a little too technical for us here; see Kingman's original papers and the recent work of Möhle (*e.g.*, 2001). With reference to the discussion of Poisson processes in Chapter 2, we can recognize that the exponential distribution in equation 3.9 is consistent with a Poisson process in which each of the $i(i-1)/2$ possible pairs coalesces independently with rate $\lambda = 1$. We can also suspect that the way the limiting, continuous-time coalescent is obtained within any particular model of a population must be like the way in which the binomial distribution became a Poisson and the geometric distribution became an exponential when the probability of success became very small (but here with $N \to \infty$). The next two sections illustrate these notions in heuristic derivations of the coalescent under the Wright-Fisher and Moran models, drawing heavily upon the excellent work of Watterson (1975), Hudson (1983a,1990), Tajima (1983), and Tavaré (1984). Interested readers should also consult the reviews of coalescent theory by Donnelly and Tavaré (1995) and Nordborg (2001).

3.2.1 Wright-Fisher Model Derivation

Kingman (1982a,b) proved that the coalescent process describes the ancestral genetic process for a sample of fixed size n in the limit as N approaches infinity in the Wright-Fisher model. The ancestral process starts from a present day sample of n gene copies, *e.g.* DNA sequences at some genetic locus, and traces the ancestral lineages of the sample back to the most recent common

ancestor. Again, a lineage at a particular generation in the past is represented by an individual whose genome contains material directly ancestral to one or more of the samples. The n gene copies, or sequences, which we can also think of as the lineages at time zero of the ancestral process, are assumed to have been sampled without replacement from the population. Sampling without replacement is what empiricists do in practice, unless there is something to prevent it, and this guarantees that all n members of the sample represent distinct genetic lineages. The requirement that N approaches infinity while n remains fixed is typically stated as $n \ll N$ (n is much less than N), because we use the coalescent as an approximation to the behavior of a relatively small sample from a large population rather than a truly infinite one.

Assume for the moment that N is not necessarily large. The Wright-Fisher model assumes that the j ancestors of i lineages are sampled randomly with replacement from the N individuals present in the previous generation. Each parent has chance $1/N$ of being chosen as the parent of each lineage, and we can think of this process as tossing i balls randomly into N boxes. If two or more balls wind up in the same box we say that those lineages have a common ancestor in the previous generation. Thus, when all i balls fall into distinct boxes, the number of ancestors, j, is equal to i. At the other extreme, if all i balls land in the same box, then j is equal to one and all the lineages share a common ancestor in the previous generation. This process leads to the following single-generation transition probability, the probability that i lineages are descended from j ancestors in the immediately previous generation:

$$G_{i,j} = \frac{S_i^{(j)} N_{[j]}}{N^i} \qquad 1 \le j \le i \tag{3.13}$$

(Watterson, 1975), in which $N_{[j]} = N(N-1)\cdots(N-j+1)$ is a descending factorial, and $S_i^{(j)}$ are Stirling numbers of the second kind. The distribution given by equation 3.13 is an example of an occupancy distribution; see Johnson, Kotz, and Kemp (1993) for a recent thorough account of these well-studied distributions.

The Stirling number of the second kind $S_i^{(j)}$ is the number of ways of a set of i elements can be partitioned into j subsets. For example, consider $G_{i,i-1}$, which is the probability that i lineages have $i-1$ ancestors in the previous generation. Recalling Table 2.1 for the case of coin tosses, we could enumerate all the possible ways of throwing i balls into N boxes, then group them according to the number of occupied boxes. Each arrangement in which $i-1$ boxes were occupied would represent the case where a single pair of lineages had a common ancestor and the other $i-2$ had distinct ancestors. Each of these arrangements would have the same probability $N_{[i-1]}/N^i$ and there would be

$$S_i^{(i-1)} = \binom{i}{2} = \frac{i(i-1)}{2}$$

of them because this is the number of possible pairs. Stirling numbers of the second kind can be generated recursively using $S_i^{(1)} = 1$ and

$$S_i^{(j)} = S_{i-1}^{(j-1)} + j S_{i-1}^{(j)} \tag{3.14}$$

for $j = 2, 3, \ldots, i-1$, and with $S_i^{(i)} = 1$. They also satisfy the equation

$$x^i = \sum_{j=1}^{i} S_i^{(j)} x_{[j]}, \tag{3.15}$$

which shows that the distribution given by equation 3.13 sums to one over $j = 1, 2, \ldots, i$. There are also Stirling numbers of the first kind, and we will see these in Section 4.2. Abramowitz and

		j									
		1	2	3	4	5	6	7	8	9	10
	10				0.017	0.129	0.345	0.356	0.136	0.016	
	20					0.008	0.062	0.224	0.372	0.268	0.065
	50						0.003	0.030	0.166	0.419	0.382
N	100							0.005	0.056	0.311	0.628
	200								0.016	0.187	0.796
	500								0.003	0.084	0.913
	1000								0.001	0.043	0.956

Table 3.1: The probability $G_{i,j}$, that $i = 10$ sequences have j ancestors in the immediately previous generation for different values of N. Values $< 10^{-3}$ are omitted for readability.

Stegun (1964) list many properties of Stirling numbers, tabulate their values, and give further references.

Returning to equation 3.13 we can see that Kingman's coalescent does not apply exactly to the Wright-Fisher model when the population size N is not large. In the Wright-Fisher model, i lineages might have anywhere from $j = 1$ to $j = i$ ancestors in the immediately previous generation. The coalescent, however, admits only $j = i$ and $j = i - 1$, that at most two out of the i share a common ancestor in any generation. Table 3.1 lists $G_{i,j}$ of equation 3.13 for a sample of size ten, or for ten lineages, as N increases. When N is equal to ten, it is most likely that there are six or seven ancestors of the ten lineages in the previous generation. Thus there will often be three or four coalescent events in one generation. Scanning down any column, we see that the chance that there are $j < i$ ancestors decreases rapidly as N increases, while the chance that there are $j = i$ approaches one. By the time N is as big as 1000, nearly all of the probability mass is found at $j = i - 1$ and $j = i$, and the probabilities for $j < i - 1$ become insignificant in comparison. This implies that the requirement of the coalescent, that at most one coalescent event occurs in given generation, is met, but it is difficult to extract much more than this from table 3.1.

Using equation 3.13 and the image of balls and boxes, we can show that the Wright-Fisher model yields the coalescent when N is very large. Consider $G_{i,i}$, the probability that i lineages have i distinct ancestors in the immediately previous generation. The first of ball is thrown randomly, and it lands in one of the N boxes. This is the ancestor of the first sequence. Now there are $N - 1$ empty boxes, so the chance that the next ball thrown lands in an unoccupied box is equal to $(N - 1)/N$. This is the probability that the first two sequences have different ancestors, that they do not coalesce. The chance that the third ball thrown also lands in an empty box is then $(N - 2)/N$, and so on. Continuing, and simplifying, we obtain

$$\begin{aligned} G_{i,i} &= \left(\frac{N-1}{N}\right)\left(\frac{N-2}{N}\right)\cdots\left(\frac{N-(i-1)}{N}\right) \\ &= \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{i-1}{N}\right) \\ &= 1 - \frac{\sum_{j=1}^{i-1} j}{N} + o\left(\frac{1}{N}\right), \end{aligned}$$

where, as in equation 2.49, the notation $o(1/N)$ represents terms that decrease to zero faster than $1/N$ as N tends to infinity. The sum in the numerator of the second term on the right above is equal to the binomial coefficient $i(i-1)/2$, which can be seen from equation 2.38 and Table 2.2). Similarly, from equation 3.13 we obtain

$$\begin{aligned} G_{i,i-1} &= \frac{S_i^{(i-1)} N_{[i-1]}}{N^i} \\ &= \frac{\binom{i}{2}}{N}\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{i-2}{N}\right) \\ &= \frac{\binom{i}{2}}{N} + o\left(\frac{1}{N}\right) \end{aligned} \tag{3.16}$$

since $S_i^{(i-1)} = i(i-1)/2$ as noted above. All other $G_{i,j}$, with $j < i-1$, are $o(1/N)$.

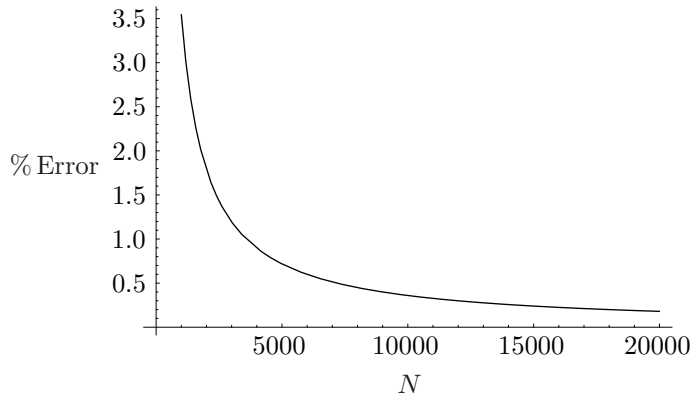


Figure 3.2: The Wright-Fisher model's convergence to the coalescent.

Thus, as N becomes larger and larger, the ancestral process for i lineages becomes like a series of Bernoulli trials with a constant probability $G_{i,i-1} = i(i-1)/(2N)$ each generation of success. Success in this case means that a single pair of lineages coalesces. Figure 3.2 shows the percent error of this approximation for $i = 10$, as a function of N. Specifically, the curve plots the difference between the full expression for $G_{i,i-1}$ from equation 3.16 and the approximation $G_{i,i-1} = i(i-1)/(2N)$ as a percentage of the full $G_{i,i-1}$. Using equation 3.16 we can show that this will be very close to $(i-1)(i-2)/(2N)$, or $36/N$ when $i = 10$, as long as N is not too small. As an aside, note that this exposes a shortcoming of the use of the relatively weak $o(1/N)$ conditions above, for example in equation 3.16. In fact, we know that these terms are of order, or proportional to, $1/N^2$ and so will decrease to zero much more quickly than $o(1/N)$ requires ($1/N^p$ where $p > 1$; see equation 2.49). The curve in figure 3.2 begins at $N = 1000$, which is the largest value of N in table 3.1 with an error of only about 3.6%, and it drops quickly to less than 1% when N is greater than 3600. This illustrates that the coalescent can be a reasonable approximate model for a large finite population.

Formally, in the limit as N tends to infinity the ancestral process under the Wright-Fisher model converges to the continuous-time coalescent process described by Kingman. Time is

measured in units of N generations, and we can express this limiting results in terms of (one minus) the distribution function, or

$$P\{T_i^{(N)} > t\} \quad = \quad (1 - G_{i,i})^{[Nt]} \quad \longrightarrow \quad e^{-\binom{i}{2}t} \qquad \text{as} \quad N \to \infty,$$

which is identical to that of the exponential distribution (see equation 2.61) with parameter $\binom{i}{2}$. The notation $[Nt]$ above means the integer part of Nt. It simply recognizes the fact that, while t can assume any value greater than zero, the geometric probability $(1 - G_{i,i})^{[Nt]}$ makes sense only for whole generations; this discrepancy become negligible as N approaches infinity.

3.2.2 Moran Model Derivation

In the previous section, we saw that the coalescent holds in the Wright-Fisher model only in the limit of very large population size. For finite N it was necessary to consider the possibility of multiple coalescent events in a single generation. However, the derivation above was relatively simple because the Wright-Fisher model is formulated in a way that makes it well-suited for a retrospective approach: the parents of the current generation are obtained by random sampling with replacement from the previous generation. The Moran model provides an important counterpoint to this. First, there is no possibility of multiple coalescent events in a single time step, so the structure of the finite-N process is less complicated than in the Wright-Fisher model. Second, the Moran model does not include a simple, ready-made description of an ancestral process. Instead, the ancestral process must be obtained by considering both the sampling of lineages and the process of reproduction forward in time in the population. This is required in the analysis of most models, for example those in the next section, and the Moran model provides an instructive setting for becoming familiar with this approach.

As before, the ancestral process begins with a sample of size n taken randomly without replacement from the population, and the same considerations apply to the ancestry of i lineages that existed at some time in the history of the sample. Now we must account for the various possible states of the population when the sample was taken. Fortunately, under the reproductive scheme of the Moran model, in a single time step only two things can happen in the population. With probability $1/N$, the same individual is chosen to reproduce and to die. We note in passing that here a mutation might occur, although we continue to ignore mutation until Chapter 4. What is important here is that, in this case, a single offspring replaces its parent, so a common ancestor event between two lineages is impossible, both in the whole population and among the lineages ancestral to a sample. On the other hand, with probability $1 - 1/N$, the individual chosen to reproduce is different than the individual chosen to die. In this case, the individual who reproduces survives and its offspring replaces the individual who dies. This represents the birfucation of one lineage, so looking backwards in time a common ancestor event occurs in the total population. There is no possibility of multiple coalescent events in a single time step.

However, a common ancestor event somewhere in the population is not guaranteed to occur among some smaller number, i, of ancestral lineages. This requires, in addition, that the i lineages contain both the individual who reproduced and its offspring. We label the offspring 1 and its parent 2, and these now coexist in the population. Then the probability that i lineages randomly sampled without replacement include both of these individuals can be computed as

$$P\{1 \text{ in sample } \cap \text{ 2 in sample}\} \quad = \quad 1 \; - \; P\{1 \text{ not in sample } \cup \text{ 2 not in sample}\},$$

or one minus the probability that 1 or 2 (or both) are not in the sample. The term on the right

is readily calculated using as

$$\begin{aligned} P\{\text{1 not in sample } \cup \text{ 2 not in sample}\} &= P\{\text{1 not in sample}\} + P\{\text{2 not in sample}\} \\ &\quad - P\{\text{1 not in sample } \cap \text{ 2 not in sample}\} \end{aligned}$$

which is a straightforward application of equation 2.8.

Random sampling without replacement can be envisioned as tossing balls into boxes, but with the provision that occupied boxes are prohibited from receiving any more balls. By computing the probabilities that box 1 remains empty after each ball is tossed and mutliplying these together, we have

$$P\{\text{1 not in sample}\} = \left(\frac{N-1}{N}\right)\left(\frac{N-2}{N-1}\right)\cdots\left(\frac{N-1-(i-1)}{N-(i-1)}\right) = \frac{N-i}{N}.$$

The same considerations for box 2 show that $P\{\text{2 not in sample}\}$ is identical to this. Using the same approach, we have

$$\begin{aligned} P\{\text{1 not in sample } \cap \text{ 2 not in sample}\} &= \left(\frac{N-2}{N}\right)\left(\frac{N-3}{N-1}\right)\cdots\left(\frac{N-1-(i-1)}{N-(i-1)}\right) \\ &= \frac{(N-i)(N-1-i)}{N(N-1)}. \end{aligned}$$

Putting all of this together gives

$$\begin{aligned} P\{\text{1 in sample } \cap \text{ 2 in sample}\} &= 1 - 2\frac{N-i}{N} + \frac{(N-i)(N-1-i)}{N(N-1)} \\ &= \frac{i(i-1)}{N(N-1)}, \end{aligned}$$

which again is the probability that the i lineages contain both the parent and its offspring, and thus that two of sample lineages have a common ancestor in the previous generation, given that such an event can occur.

In all, the chance that a common ancestor event occurs among the i lineages is equal to the probability that reproduction in the population makes it possible, *i.e.* that the offspring individual does not replace its parent, multiplied by the probability that both the offspring and its parent are among the i sample lineages:

$$\begin{aligned} G_{i,i-1} &= \left(1-\frac{1}{N}\right)\frac{i(i-1)}{N(N-1)} \\ &= \binom{i}{2}\frac{2}{N^2}. \end{aligned} \tag{3.17}$$

Because we know that only one other event is possible. *i.e.* no common ancestor event, we have

$G_{i,i} = 1 - G_{i,i-1}$. For completeness, we can calculate $G_{i,i}$ easily using the above logic:

$$\begin{aligned} G_{i,i} &= \frac{1}{N} + \left(1 - \frac{1}{N}\right)\left(1 - \frac{i(i-1)}{N(N-1)}\right) \\ &= 1 - \binom{i}{2}\frac{2}{N^2}. \end{aligned} \tag{3.18}$$

Thus, as noted above, one aspect of the coalescent is an exact result for the Moran model: only two lineages can coalesce at a time. However, to obtain the continuous-time ancestral process given by equation 3.9, it is still necessary to take the limit as N goes to infinity, and to measure time in units of $N^2/2$ Moran model time steps.

3.2.3 Breeding Structure and Exchangeability

The previous two sections show that the ways in which time must be rescaled in order to obtain Kingman's coalescent process in the Wright-Fisher model and in the Moran model are the same as the rates of genetic drift, specifically the loss of heterozygosity, in these two models calculated in Section 3.1. This is not too surprising because, in some fundamental sense, the coalescent process is genetic drift viewed backwards in time. More than three decades ago, Felsenstein (1971) showed that the rate of loss of alleles in a population that contains i alleles now is related to $G_{i,i}$, and a number of other intimate connections between forward and backward processes in population genetic models have been established. Ewens (1990) reviews many of these, and Möhle (*e.g.* 2001) has made important recent extensions. In this section, we will see how the time scales of the ancestral processes in the Wright-Fisher model and in the Moran model are related to Kingman's (1982b) definition of the *effective size* of the population: $N_e = N/\sigma^2$ where σ^2 is the variance in the numbers of offspring of individuals in a large population (see below). More importantly, we will return to the concept of exchangeability introduced in Section 3.2, and investigate its biological meaning in a simple example.

Cannings (1974) described the following class of exchangeable-type population models. Let the random variable Y_i count the number of offspring of individual i in the population, and let y_i be a particular instance of Y_i. Each individual in the population is assumed to have the same distribution of offspring number, but of course these are correlated because the total population number N is assumed to be constant. That is, every realization $(y_1, y_2, \ldots, y_N)$ of the process of reproduction in the population must satisfy the constraint $\sum_{i=0}^{N} y_i = N$. Thus, the Y_i are exchangeable random variables, which means that anything we wish to compute will not depend on the labels of the individuals (Aldous, 1985). We can take exchangeability to mean identically distributed but not independent. Because they are identically distributed and must sum to N, the expected number of offspring is $E[Y_i] = 1$ for all such models. It is further assumed that the offspring-number distribution does not change over time. Finally, we note an important property of the population, which is that the numbers of offspring of an individual in different generations are independent, and we can see this as a consequence of fact that the individuals can be relabelled each generation without any effect.

In the Wright-Fisher model, the joint distribution of the numbers of offspring each generation of the N individuals in the population is multinomial with parameters N and $p_1 = p_2 = \cdots = p_N = 1/N$. The multinomial distribution is just a generalization of the binomial distribution, in which several different outcomes are possible in each trial. Here, the different possible outcomes are that individual i ($1 \le i \le N$) is the parent of some member of the next generation. We have

$$P(Y_1 = y_1, \ldots, Y_N = y_N) = \frac{N!}{y_1! \cdots y_N!} p_1^{y_1} \cdots p_N^{y_N} \tag{3.19}$$

and with $p_1 = p_2 = \cdots = p_N = 1/N$, we obain $E[Y_i] = Np_i = 1$ and

$$\begin{aligned} \mathrm{Var}[Y_i] &= Np_i(1-p_i) = 1 - \frac{1}{N}, \\ \mathrm{Cov}[Y_i, Y_j] &= -Np_ip_j = -\frac{1}{N}, \end{aligned}$$

for the Wright-Fisher model. See Chapter 35 of Johnson, Kotz and Balakrishnan (1997) for a description of the multinomial distribution and its properties. The binomial distribution is a special case of the multinomial distribution, so the equations above can be compared to equations 2.39 and 2.40. Note that y_i can be any number from zero to N, but because the total number of offspring must be equal to N, the these numbers are strongly correlated when N is small. For example, if $N = 2$ and one individual has two offspring, the other must have no offspring. As N increases, these correlations become weak. In the limit as N goes to infinity, the distribution of the number of offspring of an individual becomes Poisson with expectation (and variance) equal to one, which is how Fisher (1922) conceived of this model.

While every generation in the Wright-Fisher model begins with N newly-produced offspring, under the Moran model individuals can persist. Therefore, we take "offspring" in the Moran model to include both the individual itself, if it persists, and its offspring in the usual sense. The joint distribution of $Y_1, \ldots, Y_N$ in the Moran model is not one of the well-known statistical distributions. It is obtained by considering the choice of one individual to reproduce and one individual to die, where in both cases the chance that a particular individual is chosen is equal to $1/N$. Thus, every one of the N^2 possible pairs of individuals is equally likely. There are N pairs in which the same individual is chosen to die and to reproduce. In this case the offspring replaces its parent and every member of the population contributes one individual to the next generation ($Y_1 = \cdots = Y_N = 1$). There are $N(N-1)$ pairs in which different individuals are chosen to reproduce and to die, and again each of these has probability $1/N^2$. In this case, the individual i who reproduces has $Y_i = 2$ and the individual $j \neq i$ has $Y_j = 0$. Therefore, we have

$$P(Y_1 = y_1, \ldots, Y_N = y_N) = \begin{cases} \dfrac{1}{N} & \text{if } y_1 = \cdots = y_N = 1, \\ \dfrac{1}{N^2} & \text{if } (y_i, y_j) = (2,0)\, i \neq j, y_r = 1 \text{ for all } r \neq i, j, \\ 0 & \text{otherwise.} \end{cases} \tag{3.20}$$

The top term on the right includes all the possibilities for choosing the same individual to reproduce and to die.

The expectation and variance of the number of offspring Y_i of an individual in the Moran model can be obtained using equation 3.20, or directly from equations 3.5 and 3.6 in Section 3.1.2 by considering an allele in single copy, that is with frequency $p = 1/N$. The covariance of Y_i and Y_j can be obtained from equation 3.20 by noting that the product $(Y_i - E[Y_i])(Y_j - E[Y_j]) = (Y_i - 1)(Y_j - 1)$, is only non-zero when one individual leaves two descendents and the other leaves zero. Again $E[Y_i] = 1$, and we have

$$\begin{aligned} \mathrm{Var}[Y_i] &= \frac{2}{N}\left(1 - \frac{1}{N}\right), \\ \mathrm{Cov}[Y_i, Y_j] &= -\frac{2}{N^2}. \end{aligned}$$

Again we can see that the covariance approaches zero as N grows. In constrast to the Wright-Fisher model, the variance also has this property, although the approach to zero is N times slower than for the covariance.

In addition to the Wright-Fisher model and the Moran model, Kingman (1982b) showed that the coalescent holds for a subset of the exchangeable-type population models of Cannings (1974) in the limit as N tends to infinity and with time rescaled appropriately. In particular, Kingman assumed that variance of offspring number in this limit,

$$\lim_{N\to\infty} \mathrm{Var}[Y_i] \quad = \quad \sigma^2,$$

was finite and non-zero ($0 < \sigma^2 < \infty$). The Wright-Fisher model satisfies this criterion, and has $\sigma^2 = 1$, but the Moran model does not, with $\sigma^2 = 0$. The Moran model must be treated separately, and yet as Section 3.1.2 shows, it still has the coalescent as its limiting ancestral process. In the general case, the coalescent is obtained when time is rescaled by the factor $N_e = N/\sigma^2$, and we can see that this is the correct time scale for the Wright-Fisher model ($N_e = N$), and for the Moran model ($N_e = N^2/2$) despite the fact that the Moran model does not satisfy the condition for $\mathrm{Var}[Y_i]$.

The coalescent, with its effective population size $N_e = N/\sigma^2$, can be obtained in these general models by considering the possible realizations of the process of reproduction, then sampling i individuals randomly without replacement, and computing $G_{i,j}$ following Gladstein (1978). Convergence to the coalescent results from the fact that

$$G_{i,j} \quad = \quad \begin{cases} 1 - \binom{i}{2}\sigma^2/N \ + \ o(1/N) & \text{if } j = i, \\ \binom{i}{2}\sigma^2/N \ + \ o(1/N) & \text{if } j = i-1, \\ o(1/N) & \text{otherwise.} \end{cases} \tag{3.21}$$

If necessary, for example to examine errors or rates of convergence as in figure 3.2, we could make the stronger statement than the largest parts of the $o(1/N)$ terms above are on the order of $1/N^2$. Consider the probability that two lineages have a common parent in the previous generation. This requires that both lineages are among the offspring of a single individual. For a particular outcome of reproduction in the population, we can use the logic of Section 3.2.2 to obtain

$$P\{\text{two have same parent}|Y_1 = y_1, \ldots, Y_N = y_N\} \quad = \quad \sum_{i=1}^{N} \frac{y_i(y_i-1)}{N(N-1)}.$$

The average of this over the distribution of $(Y_1, Y_2, \ldots, Y_N)$ gives

$$G_{2,1} \quad = \quad E\left[\sum_{i=1}^{N} \frac{y_i(y_i-1)}{N(N-1)}\right] \quad = \quad \frac{E[y_1(y_1-1)]}{N-1} \quad = \quad \frac{\mathrm{Var}[y_1]}{N-1} \quad = \quad \frac{\sigma^2}{N} \ + \ o(1/N).$$

in which we have used the fact that $E[y_i(y_i-1)]$ is the same for every i, and that $E[y_i] = 1$, so that $E[y_i(y_i-1)] = \mathrm{Var}[y_i]$ (Kingman, 1982b).

Before moving on, we note that there have been many different definitions of effective population size, depending on what measure of genetic drift is used, and that these do not always agree (Ewens, 1982). The above, $N_e = N/\sigma^2$, might be termed the *coalescent effective size*. Sjödin *et al.* (2005) have recently argued for the use of this term in a slightly broader setting, which we will turn to in Chapter 7 when we consider the robustness of the coalescent.

From the biological standpoint, the important feature of exchangeability is that the reproductive capacities of every individual in every generation is the same. There can be no transmission of reproductive potential from parents to offspring, as would be the case if heritable variation in survivorship or fecundity existed in the population, nor can there be any correlations in reproductive potential due to other factors, such as geographic location. To be exchangeable, it must be possible to randomly reassign these labels (fitnesses of alleles, geographic locations, etc.) without effect. So far, we have seen this property as a consequence of the biological assumptions of panmixia and neutrality in the Wright-Fisher model and the Moran model. However, it is possible to construct models with non-trivial biological structure, but within which the offspring numbers are still exchangeable, and this illustrates the meaning of exchangeability.

Let us assume that the habitat is structured in such a way as to determine the distribution of offspring numbers. Note that we have already made one assumption of this sort: that the population size is constant over time, implicitly fixed by external factors. We can call this new model the "nest-site" model. At the start of every generation, each individual has an equal chance of securing any given nest site, but nest sites differ in quality. There are many different ways to proceed at this point, and for the sake of illustration we choose one. Assume that there are K different kinds of nest sites. Nests of type i comprise a fraction β_i of the total number of nest sites. The quality of nest sites is fixed so that the individuals who occupy sites of type i account for a fraction α_i of offspring. Let us further assume that the $N\alpha_i$ offspring are produced by their $N\beta_i$ parents via Wright-Fisher sampling.

Consider the ancestry of a sample of size two under this model. The probability that the two individuals come from the same parent in the immediately previous generation is given by

$$P\{\text{coal}\} \quad = \quad \sum_{i=1}^{K} \alpha_i \left(\frac{N\alpha_i - 1}{N} \right) \left(\frac{1}{N\beta_i} \right).$$

This is the probability that both samples, taken without replacement, came from the part of the population that was produced by individuals in type i nest sites times the chance that they had the same parent given this. As N increases, this probability of coalescence becomes

$$P\{\text{coal}\} \quad \approx \quad \frac{1}{N} \sum_{i=1}^{K} \frac{\alpha_i^2}{\beta_i}. \tag{3.22}$$

Now consider the number Y_1 of offspring of a single newborn individual when the population size is large. With probability β_i the individual will have a Poisson number of offspring with mean and variance equal to α_i/β_i. Then the expected number of its offspring is equal to one, which is true of course of any constant-size population model. By conditioning on the type of nest site the individual ends up occupying, we have

$$\begin{aligned} \sigma^2 \quad &= \quad \sum_{i=1}^{K} \beta_i \left[\frac{\alpha_i}{\beta_i} \; + \; \left(\frac{\alpha_i}{\beta_i} \right)^2 \right] \; - \; 1 \\ &= \quad \sum_{i=1}^{K} \frac{\alpha_i^2}{\beta_i}. \end{aligned} \tag{3.23}$$

The term in brackets above is equal to the expected value Y_1^2, given that it occupies a nest site of type i. Comparing equation 3.23 to equation 3.22 we see that $N_e = N/\sigma^2$ under this nest-site model, and since this is a Cannings model, Kingman's coalescent is the ancestral process in the limit as N goes to infinity and time is measured in units of N_e generations, provided that $0 < \sigma^2 < \infty$.

When $\alpha_i = \beta_i = 1/K$, equation 3.23 gives $\sigma^2 = 1$ and $N_e = N$ as in the Wright-Fisher model. In all other cases, $\sigma^2 > 1$, and $N_e < N$ in the nest-site model. For example, if there are just two types of nests in the frequencies $\beta_1 = 1/4$ and $\beta_2 = 3/4$, and type-1 nests are the only ones that permit reproduction ($\alpha_1 = 1$), then $\sigma^2 = 4$ and $N_e = N/4$. Equation 3.23 says that whenever some indivuals produce a disproportion number of offspring, the coalescent effective size will be smaller than the actual size of the population. Despite the obvious biological structure of the population, convergence to the coalescent means that only effect of the structure is on N_e: the shape of the ancestral process for a sample, and thus sampled data,is exactly the same as if there were no structure at all. Again, the key feature of the nest-site model, which makes it an exchangeable-type model, is that nest sites are not inherited, but assigned randomly every generation. Cases in which structure alters the ancestral process more dramatically will be seen in Chapters 5 and 7.

3.3 Some Properties of Coalescent Genealogies

Twenty years after the birth of coalescent theory, the field abounds with results concerning the sizes and shapes of genealogies. Some of the properties that have been studied are of natural interest considering the mathematical structure of the coalescent. Most have been of interest because they are related to the measurement of biological diversity. Given the important association between coalescent theory and the collection and analysis of genetic data, there is a good deal of overlap between the two. For example, the time T_{MRCA} back to the most recent common ancestor of the sample is equal to the stopping-time of the coalescent, but it can also be a quantity of great interest to biologists studying the history of populations. Section 3.3.1 below considers T_{MRCA} and another measure, T_{total}, or the total length the genealogy, which is of inherent interest to biologists since it is equal to the time over which mutations might have occurred in the history of the sample. Section 3.3.2 then considers the branching structure of genealogies. These structures and their associated probabilities are also of interest both mathematically and biologically. In addition, an understanding of them is essential before predictions about measures of sequence polymorphism that depend on tree structure — such as the distribution of the site frequencies introduced in Chapter 1 — can be made in Chapter 4.

3.3.1 Two Measures of the Size of a Genealogy

The mathematical simplicity of the coalescent derives from the fact that the coalescence times T_i are (i) independent of one another and (ii) independent of the branching structure of the genealogy. Both of these properties follow directly from Poisson process of coalescence with rate equal to one for every pair of lineages. As a result, it is straightforward for make predictions about many quantities, including two of enduring interest to population geneticists: the time to the most recent common ancestor of the entire sample, T_{MRCA}, and the total length of all the branches in the genealogy, T_{total}. Because T_i is the time in the history of the sample during which there were exactly i ancestral lineages,

$$T_{\mathrm{MRCA}} = \sum_{i=2}^{n} T_i \tag{3.24}$$

and

$$T_{\mathrm{total}} = \sum_{i=2}^{n} i T_i \tag{3.25}$$

Equation 3.24 is just the sum of all $n-1$ coalescence times, and equation 3.25 is the sum of the lengths of all the branches in the genealogy, broken up into the coalescence time intervals,

T_i. Remembering section 1.1 above, we might naively have defined some τ_i to be the length of the i-th branch in the genealogy, where $1 \leq i \leq 2n-2$, and then T_{total} would be the sum of these: $\sum_{i=1}^{2n-2} \tau_i$. If we then wanted to calculate the expectation and variance of T_{total}, or its probability function, we would have faced serious problems because the τ_i and their distributions would be different for different genealogies. Thankfully, this is unnecessary. We know that all genealogies have i lineages during time T_i regardless of their structure, and this makes it easy to "integrate" over all possible genealogies to obtain the properties of T_{total} (and T_{MRCA}).

Because T_{MRCA} and T_{total} are simple functions of independent exponential random variables, we can use equations 2.16 and 2.27, together with equation 2.51, to compute the expectations of T_{MRCA} and T_{total}. Thus,

$$E[T_{\text{total}}] = \sum_{i=2}^{n} iE[T_i] = \sum_{i=2}^{n} i\frac{2}{i(i-1)} = 2\sum_{i=1}^{n-1}\frac{1}{i} \tag{3.26}$$

and

$$\begin{aligned} E[T_{\text{MRCA}}] &= \sum_{i=2}^{n}\frac{2}{i(i-1)} = 2\sum_{i=2}^{n}\left(\frac{1}{i-1}-\frac{1}{i}\right) \\ &= 2\left(1-\frac{1}{2}+\frac{1}{2}-\frac{1}{3}+\frac{1}{3}-\ldots-\frac{1}{n-1}+\frac{1}{n-1}-\frac{1}{n}\right) \\ &= 2\left(1-\frac{1}{n}\right) \end{aligned} \tag{3.27}$$

The variances of T_{MRCA} and T_{total} are also computed easily using equations 2.19 and 2.29, together with equation 2.51. These turn out to be

$$\text{Var}[T_{\text{total}}] = 4\sum_{i=1}^{n-1}\frac{1}{i^2} \tag{3.28}$$

and

$$\text{Var}[T_{\text{MRCA}}] = 8\sum_{i=2}^{n}\frac{1}{i^2} - 4\left(1-\frac{1}{n}\right)^2 \tag{3.29}$$

Equations 3.26 and 3.28 are due to Watterson (1975), while Hudson (1990) and Donnelly and Tavaré (1995) derive and review equations 3.27 and 3.29.

Tajima (1993) and Tavaré *et al.* (1997) point out that $E[T_{\text{MRCA}}]$, $\text{Var}[T_{\text{MRCA}}]$, and $\text{Var}[T_{\text{total}}]$ converge to constant values 2, $4\pi^2/3 - 12 \approx 1.16$, and $2\pi^2/3 \approx 6.58$, respectively, as the sample size n goes to infinity. In contrast, $E[T_{\text{total}}] \approx 2(\log(n)+\gamma)$ and so increases without bound as n grows — the constant Euler's $\gamma \approx 0.577216$ is defined to be $\lim_{n\to\infty}\sum_{i=1}^{n} 1/i - \log(n)$. Figure 3.3 shows how $E[T_{\text{MRCA}}]$ and $E[T_{\text{total}}]$ depend on n. Although $E[T_{\text{total}}]$ does increase without bound, it does so more slowly for larger n. As equation 3.26 shows, sampling an $(n+1)$st sequence adds only $2/n$ to what may already be a sizable number. This has consequences for the measurement of DNA sequence polymorphism, which we will explore in Chapter 4. Similarly, from figure 3.3 or equation 3.27, we can see that $E[T_{\text{MRCA}}]$ is close to its asymptotic value of 2 even for moderate n. Figure 3.1, in which the lengths of the coalescence times are drawn in proportion to their expected values, shows the consequences this has on the shapes of genealogies under the standard coalescent model. For all but the smallest samples, there will likely be a large number of coalescent events in the very recent history of the sample. Seen from another perspective, the most ancient coalescence times comprise a large fraction of any genealogy.

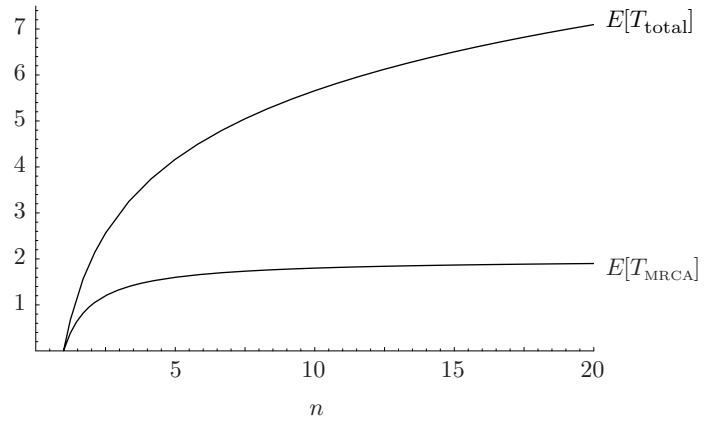


Figure 3.3: The relationship between sample size and the expected values of T_{MRCA} and T_{total}.

Although the moments of T_{MRCA} and T_{total} are quite easy to obtain, the fact that the coalescence times T_i are mutually independent makes the derivations of the full probability distributions of T_{MRCA} and T_{total} almost as straightforward. The distribution of T_{MRCA} is simply the sum of $n-1$ independent exponential random variables, T_i, with parameters $i(i-1)/2$ for $2 \leq i \leq n$. We can use equation 2.66 to immediately obtain

$$f_{T_{MRCA}}(t) = \sum_{i=2}^{n} \binom{i}{2} e^{-\binom{i}{2}t} \prod_{\substack{j=2 \\ j \neq i}}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}}. \tag{3.30}$$

Equation 3.30 is due to Takahata and Nei (1985) who used a Laplace transform of $T_{\text{MRCA}} = \sum_{i=1}^{n} T_i$ rather than directly performing $n-2$ convolutions. However, the distribution of T_{MRCA} had been obtained previously by Tavaré (1984) working directly from Kingman's ancestral Markov chain and using matrix methods. The resulting equation,

$$f_{T_{\text{MRCA}}}(t) = \sum_{i=2}^{n} \frac{(2i-1)(-1)^i n_{[i]}}{n_{(i)}} \binom{i}{2} e^{-\binom{i}{2}t}, \tag{3.31}$$

in which

$$n_{[i]} = n(n-1)\dots(n-i+1) \tag{3.32}$$

$$n_{(i)} = n(n+1)\dots(n+i-1), \tag{3.33}$$

may look quite different, but is identical to equation 3.30 above.

The distribution of the total length of the genealogy, T_{total}, can also be obtained. Note that, if we define $T_i^* = iT_i$, then from equation 2.52, T_i^* also follows an exponential distribution,

$$f_{T_i^*}(t) = \frac{i-1}{2} e^{-\frac{i-1}{2}t}. \tag{3.34}$$

Thus, similarly to T_{MRCA}, the total tree length $T_{\text{total}} = \sum_{i=2}^{n} T_i^*$ is the sum of $n-1$ independent but not identically distributed exponential random variables, and its distribution can also be

obtained directly from equation 2.66. Using this method, it is given by

$$f_{T_{\text{total}}}(t) = \sum_{i=2}^{n} \frac{i-1}{2} e^{-\frac{i-1}{2}t} \prod_{\substack{j=2 \\ j\neq i}}^{n} \frac{j-1}{j-i}. \tag{3.35}$$

As with equation 3.31 above, the alternative form

$$f_{T_{\text{total}}}(t) = \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \frac{i-1}{2} e^{-\frac{i-1}{2}t} \tag{3.36}$$

is available using Tavaré's (1984) matrix method. Equation 3.36 can be simplified further using the binomial theorem, equation 2.36, to give

$$f_{T_{\text{total}}}(t) = \frac{n-1}{2} e^{-\frac{t}{2}} \left(1 - e^{-\frac{t}{2}}\right)^{n-2} \tag{3.37}$$

and this may sometimes be preferred over equations 3.35 and 3.36.

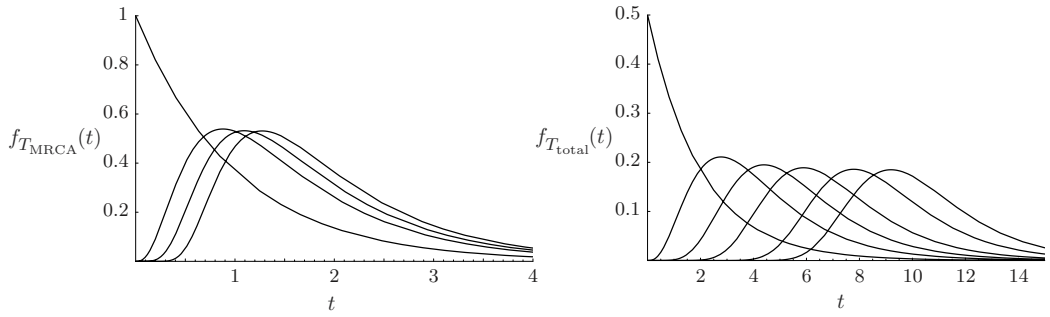


Figure 3.4: The distributions of T_{MRCA} and T_{total} for $n = 2, 5, 10, 20, 50, 100$ (from left to right). The curves for $n = 20$ and $n = 50$ are omitted for $f_{T_{\text{MRCA}}}(t)$ because they are very close to the curve for $n = 100$.

Figure 3.4 plots of probability functions of T_{MRCA} and T_{total}, given by equation 3.30 (or 3.31) and equation 3.35 (or 3.36 or 3.37) for a series of sample sizes. When n is equal to two, $f_{T_{\text{MRCA}}}(t)$ has mean equal to one and a mode at zero. As n increases, the curves for $f_{T_{\text{MRCA}}}(t)$ converge on a distribution with mean equal to two (see equation 3.27). This mean of two corresponds to a period of $4N$ generations under the Wright-Fisher model. The distribution $f_{T_{\text{MRCA}}}(t)$ has a mode at about 1.093 when $n = 10$ and 1.274 when $n = 100$. The asymmetry of $f_{T_{\text{MRCA}}}(t)$ (*e.g.* that the mode is less than the mean) reflects the strong influence of the most ancient coalescence time, T_2, which makes up a significant fraction of T_{MRCA} even when n is large. In contrast to the distribution of T_{MRCA}, as n increases the curves for $f_{T_{\text{total}}}(t)$ continue to move to the right, indicating larger and larger genealogies, even when the sample size is large. This is due to the fact, illustrated by equation 3.26, that the next sample taken adds a non-negligible increment to T_{total} even when n is already large. Although T_{total} is less strongly influenced than T_{MRCA} is by the earliest coalescence times, in particular T_2, $f_{T_{\text{total}}}(t)$ remains asymmetric as n grows. It does not, for instance, approach a Normal distribution in the limit of large n.

It is possible to obtain the limiting form of $f_{T_{\text{total}}}(t)$ as n increases. So that the limiting distribution will be centered around zero, define $T^*_{\text{total}} = T_{\text{total}} - 2(\log(n) - \gamma)$, where again

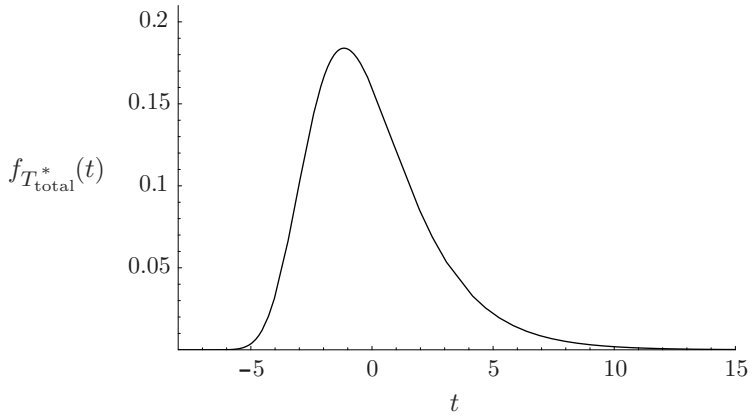


Figure 3.5: The limiting distribution $T^*_{\text{total}} = T_{\text{total}} - 2(\log(n) + \gamma)$ as $n \to \infty$.

$\gamma = \lim_{n\to\infty} \sum_{i=1}^{n} 1/i - \log(n) \approx 0.577216$. In other words T^*_{total} is to the deviation of T_{total} from its expected value for large n. With this change of variable, and using equation 2.47, as n approaches infinity equation 3.37 gives

$$f_{T^*_{\text{total}}}(t) \quad = \quad \frac{1}{2} e^{-t/2-\gamma-e^{-t/2-\gamma}}. \tag{3.38}$$

This is depicted in figure 3.5, and illustrates that the series of distributions of the total length of the genealogy shown in figure 3.4 assumes a stable shape in the limit of large sample size. The distribution 3.38 is an example of an extreme value distribution, which are reviewed in Chapter 22 of Johnson, Kotz, and Balakrishnan (1995). In particular, equation 3.38 is identical to the distribution 22.25 in Johnson, Kotz, and Balakrishnan (1995) with their $\xi = -2\gamma$ and their $\theta = 2$. An extreme value distribution makes sense because we can think of the distribution in equation 3.34 as the waiting time to the first event among $i - 1$ independent exponential processes, each with rate equal to $1/2$, so that $f_{T_{\text{total}}}(t)$ is identical to the distribution of the maximum of $n - 1$ exponential waiting times. Finally, we have $E[T^*_{\text{total}}] = 0$, and since T^*_{total} differs from T_{total} by a constant, $\text{Var}[T^*_{\text{total}}] = \text{Var}[T_{\text{total}}] = 2\pi^2/3$.

3.3.2 The Branching Structure of Genealogies

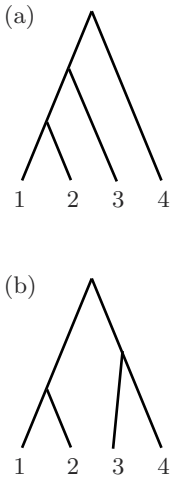
Considering the backward process of the coalescent, it is easy to see that genealogies are random-joining binary trees. We will use this notion in the next chapter to derive some predictions about DNA sequence polymorphism. Note that sometimes it is easier to consider the forward process of random bifurcation of lineages rather than the backward process of random joining. The study of tree structures has a long history in probability theory and evolutionary biology, dating back at least to the work of Cayley (1889) and Yule (1924). One property of these trees has already been mentioned: every one of them is equally likely. Figure 3.6 shows all the possible genealogical structures for a sample of size $n = 4$. We count eighteen in all: there are two different trees (a) and (b) with twelve and six possible labellings of the tips, respectively. This results from the fact that there are six possible pairs that can be the first to coalesce, and for each of these there are three possible pairs to coalesce among the three remaining lineages. Thus, coalescent genealogies are rooted (by the MRCA) bifurcating trees with labelled tips and nodes ordered in time.

The trees in figure 3.6 are distinguished by the number of tips on either side of the root: three and one in (a), and two and two in (b). Genealogies of the type (a) are twice as likely as genealogies of type (b) because, after the first coalescent event occurs, trees of type (b) require that the next coalescent event is between the two lineages that have not yet coalesced. In contrast, the other two possible coalescent events, between one of the "uncoalesced" lineages and the lineage ancestral to the first coalescent event, produces a genealogy of type (a). However, if we distinguish between the two branches which descend from the root of the tree, we can further separate the trees of type (a) that have one descedent to the left of the root from those that have three descedents to the left of the root. We can count these two kinds of trees by rotating the branches of the twelve possible labelled genealogies of type (a) so that the labels are always in the order ABCD. This shows that there are size of each type. Together with the six possible genealogies of type (b), this implies that the probability $p(i;n)$ that there are i tips to the left of the root is uniform on $i \in \{1, 2, \ldots, n-1\}$, that is $p(i;n) = 1/(n-1)$.

We can use the idea of random bifurcation of lineages to show that this is true. Consider the $n+1$ bifurcation event. The pattern $(i; n+1)$ could result either from the pattern $(i;n)$ and a bifurcation of one of the $n-i$ lineages to the right of the root, or from the pattern $(i-1;n)$ and a bifurcation of one of the $i-1$ lineages to the left of the root. This gives the following the recursion over a single bifurcation event

$$p(i;n+1) \quad = \quad \frac{n-i}{n} p(i;n) \; + \; \frac{i-1}{n} p(i-1;n). \tag{3.39}$$

It is easily verified, or proved by induction, that $p(i;n+1) = 1/n$, so that number of descendents on one side of the root of a coalescent tree is uniform on 1, 2, ..., $n-1$. This result has implications for the distribution of polymorphism in a sample, and we will put it to use in Chapter 4.



Labellings

Tree (a)

1	2	3	4
A	B	C	D
A	B	D	C
A	C	B	D
A	C	D	B
A	D	B	C
A	D	C	B
B	C	A	D
B	C	D	A
B	D	A	C
B	D	C	A
C	D	A	B
C	D	B	A

Tree (b)

1	2	3	4
A	B	C	D
A	C	B	D
A	D	B	C
B	C	A	D
B	D	A	C
C	D	A	B

Figure 3.6: The eighteen possible genealogies of a sample of size four.

It is important to keep in mind that the coalescence times do not depend at all on the branching structure of the genealogy under the standard coalescent model. For example, no reference was made to tree structures in the derivation of $f_{T_{\mathrm{MRCA}}}(t)$ and $f_{T_{\mathrm{total}}}(t)$ in Section 3.3.1 above. In Chapter 5, we will see that the branching patterns of genealogies can reflect patterns of

population subdivision and migration or track the historical association of subpopulations, and so might contain information about important biological features of the population. Here and in most of Chapter 4, however, predictions are made by "integrating" over all possible genealogies, including both the banching structure and the coalescence times. We will see that this can often be done by considering the simple process of random joining, or random bifurcation, without making explicit reference to particular trees. It is fortunate that this is so, because the number of possibles trees is enormous. In Chapter 8 we will encounter simulation based methods of inference that do explicitly consider genealogical trees, even under the standard coalescent model. We take a moment here to reflect on what a formidable task this is.

n	Random-joining trees (nodes ordered in time)	Unrooted bifurcating trees (nodes not ordered in time)
2	1	1
3	3	1
4	18	3
5	180	15
6	2700	105
7	56700	945
8	1587600	10395
9	57153600	135135
10	2571912000	2027025
100	1.37×10^{284}	1.70×10^{182}
1000	3.02×10^{4831}	1.91×10^{2860}

Table 3.2: The numbers of possible trees.

The number of possible tree structures can be obtained by considering the number of possible coalescent events at each step towards the MRCA. Beginning with the present-day sample of n items, whenever there are i lineages present there are $i(i-1)/2$ possible pairs of lineages to coalesce. Therefore, the total number of these random-joining trees is given by

$$\prod_{i=2}^{n} \binom{i}{2} = \frac{n!(n-1)!}{2^{n-1}}. \tag{3.40}$$

These are given in the left-hand column of Table 3.2. The table also shows the number of unrooted bifurcating trees with labelled tips, which is the number of phylogenetic trees typically considered in the systematic literature; see Penny *et al.* (1982). The latter is given by $(2n-5)!!$, or the product of the odd numbers from 1 to $(2n-5)$ (Felsenstein, 1978). The techniques for navigating this space of trees will be described in Chapter 8.

3.4 Human-Neanderthal Couples?

Following the publication of a mitochondrial DNA sequence recovered from a Neanderthal skeleton (Krings *et al.*, 1997), Nordborg (1998) presented a simple and elegant application of the theory covered in this chapter to an important question about the genetic ancestry of humans, namely whether or not there is evidence of a Neanderthal contribution to the current human

gene pool. Neanderthals, an extinct group of archaic hominids, are known to have coexisted with humans in Europe and western Asia until as recently as 30,000 years ago. There is long-standing debate about the relationship between Neanderthals and humans, and about the possibility that human genes may show Neanderthal ancestry (Stringer and Gamble, 1993). A genetic locus that had Neanderthal ancestry could, for example, show a pattern in which copies from some humans living today have a more recent common ancestor with a sequence from a Neanderthal than they do with copies of the same locus from other humans. This was not observed when Krings *et al.* (1997) compared a sequence from the control region of mitochondrial DNA from the Neanderthal type specimen to sequences from 986 modern humans. Instead, the pattern shown in figure 3.7 was observed in which all modern human sequences share a common ancestor to the exclusion of the Neanderthal sequence, with a long branch connecting the Neanderthal to humans (Krings *et al.*, 1997).

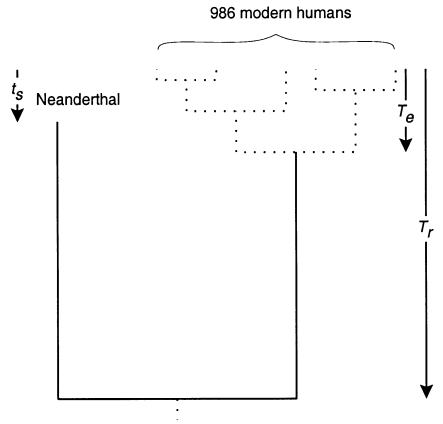


Figure 3.7: Schematic genealogy for human and Neanderthal samples, reproduced from Nordborg (1998).

Nordborg (1998) used coalescent theory to investigate whether random mating between Neanderthals and humans could be rejected based upon the observations of Krings *et al.* (1997). The 986 human mtDNA sequences share a common ancestor at an unknown time T_e in the past. This may greater than or it may be smaller than the date t_s of the Neanderthal mtDNA sequence, which is assumed to be between 30,000 and 100,000 years before present. On the coalescent time scale, this corresponds to t_s between 0.44 to 1.47, assuming a generation time of 20 years and an effective population size of 3,400 females, since human mitochondria are maternally inherited. The time T_r of the most recent common ancestor of the entire sample (humans plus Neanderthal) is also unknown. From the numbers of polymorphisms among humans and between humans and Neanderthal, Krings *et al.* (1997) claimed that T_r was at least four times T_e, and we follow Nordborg (1998) in taking this as given. The null model of random mating between archaic humans and Neanderthals can be tested by computing the chance of observing data as extreme or more extreme than those of Krings *et al.* (1997). We compute $P\{\text{tree and } T_r \geq 4T_e\}$ under the null model with t_s ranging between 0.44 and 1.47 (30,000 and 100,000). Of course, the genealogical tree is already the most extreme that could have been observed in terms of distinguishing between humans and Neanderthals, but it turns out that this alone does not rule out random mating between Neanderthals and archaic humans.

It is straightforward to compute the desired probability, Nordborg (1998) noted, by condi-

tioning on the number of human mtDNA lineages that existed at time t_s when the Neanderthal sequence joins these remaining lineages in the coalescent process. Following Tavaré (1984), Nordborg (1998) used $A_n(t)$ to denote the number of ancestral lineages that exist at time t in the past of a present-day sample of size n. The integers 1 through n are the possible values of the random variable $A_n(t)$. Computing the P-value is simplified by the fact that for given a value of $A_n(t)$ the probability of the tree and the the probability that $T_r \geq 4T_e$ are independent of one another. Thus, we have

$$P\{\text{tree and } T_r \geq 4T_e\} = \sum_{k=1}^{986} P\{\text{tree}|k\}P\{T_r \geq 4T_e|k\}P\{A_n(t_s) = k\}, \tag{3.41}$$

and we proceed by calculating all of the terms on the right side of this equation.

First, let $g_{n,k}(t) = P\{A_n(t) = k\}$, and note that this is equal to the probability that exactly $n-k$ coalescent events occur before time t in the past. Slightly different derivations are required depending on whether $k = 1$ or $k \geq 2$. We can obtain $g_{n,1}(t)$ directly from the distribution of the time to the MRCA of the sample:

$$\begin{aligned} g_{n,1}(t) &= \int_0^t f_{T_{MRCA}}(x)dx \\ &= \int_0^t \sum_{i=2}^{n} \binom{i}{2} e^{-\binom{i}{2}x} \prod_{\substack{j=2 \\ j\neq i}}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} dx \\ &= \sum_{i=2}^{n} \left[1 - e^{-\binom{i}{2}t}\right] \prod_{\substack{j=2 \\ j\neq i}}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} \\ &= 1 - \sum_{i=2}^{n} e^{-\binom{i}{2}t} \prod_{\substack{j=2 \\ j\neq i}}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}}. \end{aligned} \tag{3.42}$$

The last step above uses the fact that

$$\sum_{i=2}^{n} \prod_{\substack{j=2 \\ j\neq i}}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} = 1,$$

which we know to be true because the distribution of T_{MRCA} given in equation 3.30 must integrate to one over all t (a more general relation, *i.e.* for any rates λ_i, may be obtained in the same way from equation 2.66).

The derivation of $g_{n,k}(t)$ for $k \geq 2$ is a bit more complicated because it is necessary to consider that the $(n-k)$th coalescent event occurs before time t but that the $(n-k+1)$th coalescent event occurs after time t. Let $T_{n,k} = \sum_{i=k}^{n} T_i$ be the time to the $(n-k)$th coalescent event. Using the same rule for the convolution of independent exponential random variables that we used to derive $f_{T_{MRCA}}(t) = f_{T_{n,1}}(t)$, we can obtain

$$f_{T_{n,k}}(t) = \sum_{i=k+1}^{n} \binom{i}{2} e^{-\binom{i}{2}t} \prod_{\substack{j=k+1 \\ j\neq i}}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}}. \tag{3.43}$$

Using considerably more algebra than equation 3.42 required, we find

$$\begin{aligned} g_{n,k}(t) &= \int_0^t f_{T_{n,k}}(x) \int_{t-x}^{\infty} f_{T_k}(y) dy dx \\ &= \frac{1}{\binom{k}{2}} \sum_{i=k}^{n} \binom{i}{2} e^{-\binom{i}{2}t} \prod_{\substack{j=k \\ j \neq i}}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}} \qquad k \geq 2. \end{aligned} \tag{3.44}$$

This derivation of equation 3.44 requires the use of

$$\sum_{i=n'}^{n} \prod_{\substack{j=n' \\ j \neq i}}^{n} \frac{1}{\binom{j}{2} - \binom{i}{2}} = 0,$$

which can been seen to be true from equations 2.65 and 2.66 (for any rates λ_i). Thus, equations 3.42 and 3.44 comprise the probability function for $A_n(t)$. As before, Tavaré (1984) provides a slightly different form,

$$g_{n,k}(t) = \begin{cases} 1 - \sum_{i=2}^{n} \dfrac{e^{-\binom{i}{2}t}(2i-1)(-1)^i n_{[i]}}{n_{(i)}} & \text{if } k = 1 \\ \sum_{i=k}^{n} \dfrac{e^{-\binom{i}{2}t}(2i-1)(-1)^{i-k} k_{(i-1)} n_{[i]}}{i!(i-k)! n_{(i)}} & \text{if } k \geq 2 \end{cases} \tag{3.45}$$

which is identical to equations 3.42 and 3.44.

For a given value of t, $g_{n,k}(t)$ is a probability function (*i.e.* sums to one) over $k = 1, 2, \ldots, n$ since the sample must have an ancestor or ancestors. Figure 3.8 plots $g_{n,k}(t)$ for a sample of size $n = 20$ over a range of possible values of t and for every possible value of k. The value of $g_{n,k}(t)$ is given at discrete time points between zero and three, in thirty steps of size 0.1. Figure 3.8 shows the dramatic dependence of $g_{n,k}(t)$ on t when t is small which reflects the very high rate of recent coalescent events. For example, while $g_{20,20}(0.0) = 1$, the figure shows that $g_{20,20}(0.1) \approx 0$ and the mode of $g_{20,k}(0.1)$ occurs at about $k = 10$. That is we expect about ten coalescent events in a sample of size $n = 20$ over only 0.1 units of time. When $n = 986$, as with the human mtDNA sample, the rate of recent coalescence is extremely high since there are $n(n-1)/2 = 986 \times 985/2 = 485,605$ possible pairs of sequences to coalesce! This is evident in the equation 3.9 for the distribution of the coalescence time, but the consequence is seen clearly in figure 3.8. As we trace the history of a large sample back in time, the number of ancestors $A_n(t)$ collapses quickly to just a handful.

In order to calculate the P-value, equation 3.41, we next compute the probability of the tree given that $A_{986}(t_s) = k$. Again, we are interested in the chance that the entire set of human samples share a common ancestor to the exclusion of the Neanderthal sample. This is simply the probability that one particular lineage, here the Neanderthal lineage, in a sample of size $k+1$ does not coalesce with any others until the final $(2 \to 1)$ coalescent event. This is obtained easily by considering the process of random joining of lineages, with one lineage labelled to distinguished it from the others. When there are j lineages, there are $j(j-1)/2$ possible pairs to coalesce and $j-1$ of these would involve the labelled lineage since there are $j-1$ lineages that could coalesce with the labelled lineage. The desired probability is simply the product, from $j = k+1$ down to $k = 3$, of the fraction of coalescent events that do not involve the labelled lineage. Once there are just $k = 2$ lineages, one of which is the labelled lineage, the tree of

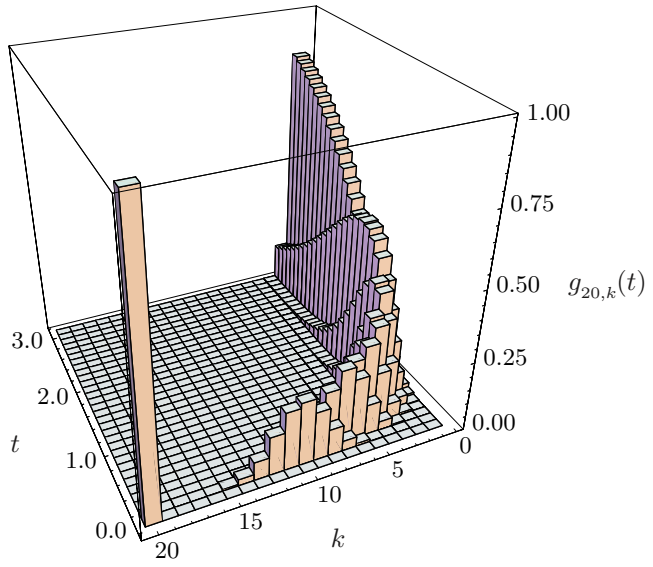


Figure 3.8: Bar chart of the probability $g_{20,k}(t)$ as a function of t and k.

interest is guaranteed. Therefore, we have

$$\begin{aligned} P\{\text{tree}|k\} &= \prod_{j=3}^{k+1}\left(1-\frac{j-1}{\binom{j}{2}}\right) = \prod_{j=3}^{k+1}\left(\frac{j-2}{j}\right) \\ &= \frac{(k-1)(k-2)(k-3)\cdots 3\cdot 2\cdot 1}{(k+1)k(k-1)(k-2)\cdots 5\cdot 4\cdot 3} \\ &= \frac{2}{k(k+1)}. \end{aligned} \tag{3.46}$$

Nordborg (1998) notes that this is a special case of a more general result; see Watterson (1982) and Saunders *et al.* (1984), as well as Chapter 4, Section 4.1.3.

The last quantity require to compute the P-value, equation 3.41, for the human and Neanderthal data is $P\{T_r \geq 4T_e|k\}$. From the definitions of these times and of t_s, depicted in figure 3.7, we have

$$\begin{aligned} P\{T_r \geq 4T_e|k\} &= P\{T_r - 4T_e \geq 0|k\} \\ &= P\{(T_r - t_s) - 4(T_e - t_s) \geq 3t_s|k\} \\ &= P\{T_{k+1,1} - 4T_{k+1,2} \geq 3t_s\} \\ &= P\{T_2 - 3T_{k+1,2} \geq 3t_s\} \\ &= P\{T_2/3 - T_{k+1,2} \geq t_s\} \end{aligned}$$

The distribution of $T_{k+1,2}$ is given by equation 3.43 and, from equation 3.9 and changing variables

as in equation 2.52, the distribution of $T_2/3$ is exponential with parameter $\lambda = 3$. Further $T_{k+1,2}$ and $T_2/3$ are independent of one another because they involve non-overlapping coalescence time intervals. In all, we have

$$P\{T_r \geq 4T_e|k\} \quad = \quad \int_0^\infty f_{T_{k+1,2}}(x) f_{T_2/3}(t_s + x)dx$$

and this can be used together with $P\{A_n(t_s) = k\} = g_{n,k}(t_s)$ and $P\{\text{tree}|k\} = 2/(k(k+1))$ in equation 3.41 to compute the probability of observing data as extreme or more extreme than what Krings *et al.* (1997) observed, under the null hypothesis of random mating between Neanderthals and archaic humans.

	t_s (in years)	
	30,000	100,000
$E[A_{986}(t_s)]$	4.86	1.75
$P\{\text{tree}\}$	0.085	0.56
$P\{\text{tree and } T_r \geq 4T_e\}$	0.0063	0.035

Table 3.3: Results for human-Neanderthal test, redrawn from Nordborg's (1998) table 1.

Table 3.3 shows the results of Nordborg's (1998) analysis. We can see that the number of human ancestral lineages expected to exist is low even for $t_s = 30,000$ years, or 0.44 coalescence time units. At first sight this may be surprising, but not in light of the results plotted in figure 3.8 which suggest very rapid change in $E[A_{986}(t)]$ for small t. Just a few lineages are expected to remain at the time the Neanderthal sequence existed. This causes the banching pattern of the tree, which may have seemed significant to the eye, to in fact be fairly likely when its probability is averaged over the distribution of $A_{986}(t)$. Finally, the hypothesis of random mating between Neanderthals and humans can be rejected at the 5% significance level over the entire plausible range of t_s. Note that, as a matter of convenience, Nordborg (1998) computed the values in table 3.3 using simulations. Then, noting that panmixia represents an unnecessarily extreme case of Neanderthal-human exchange, Nordborg (1998) used these simulations to test some less extreme scenarios in which Neanderthals could still have contributed to the human gene pool. Some of these could not be rejected, and interested readers should consult Nordborg (1998).

3.5 Exercises

1. Consider a nest-site model in which there are a finite number K of nest sites, so that the N individuals in the next generation are the descendents of just the K individuals who secure nests. What is the variance of offspring number, $\mathrm{Var}[Y_i]$, in this model?

2. Would the coalescent be the appropriate description of the limiting ($N \to \infty$) ancestral process for the population in excercise 1? Why or why not?

3. What is the expected value of the sum of the lengths of all branches on the left side of the root in a genealogy? Assume that the sample size n is even.

4. If $n = 3$ in excercise 3, what is the variance of the same quantity?

5. A sample of $n = 10$ has been taken from a population for which the coalescent holds. What is the expected length of the time during which there are an even number of lineages in the history of the sample?

6. What is the distribution of the length of time during which there are an odd number of lineages in the history of the sample?

7. What is the probability of a fully pectinate, or comb-shaped, genealogy of a sample of n sequences? Consider the sequences to be unlabelled.

8. Define a pairwise distance in a genealogy to be the sum of branch lengths as we trace from one tip of the tree to another. How many distinct pairwise distances will there be in a genealogy of a samples of size n?

9. Blah.

10. Blah.