13 Coalescent Processes

Imagine a population of \( N \) individuals, each of which produces a random number of offspring in every generation. Each offspring has one parent. This process could represent asexual reproduction, or the reproduction of individual genes. The population size is assumed to stay constant, which implies that the reproduction of the individuals is not independent. However, in a large population the interaction between individuals may be very weak. Working forwards in time, the reproduction of a single individual leads to a branching process.

The Wright-Fisher model. We are interested in the ancestry of a sample of individuals, traced backwards in time; see Figure 20. Ultimately, the whole sample traces back to a single most recent common ancestor, through a tree-like genealogy. We will make the key assumption that all the individuals are equivalent to each other, so that the process is neutral. A convenient model for such a population is the Wright-Fisher model. We imagine that, tracing back in time, each child chooses its single parent at random, independently of the other children. This resembles reality in the case in which every parent produced a very large number of offspring (much larger than \( N \)), which are then randomly culled so that the total population size remains constant at \( N \). For large \( N \), the number of offspring from each individual is close to a Poisson distribution with mean 1, and therefore variance 1.

Note. In population genetics, we often count diploid individuals; here, we count the actual genes. As a consequence, the formulae in textbooks will differ by a factor of 2.

We are interested in the shape of genealogies, and specifically, in how closely a random pair of genes are related. Let \( F_t \) be the probability that two genes are identical by descent, coming from the same ancestor \( t \) generations ago. (This is of practical importance, for familiar organisms, which typically carry two copies of each gene. Inbred individuals may inherit the same defective gene from mother and father.) There is a simple recursion for \( F_t \): starting with \( F_0 = 0 \), we have

\[
F_t = \frac{1}{N} + \left(1 - \frac{1}{N}\right) F_{t-1}
\]

for \( t \geq 1 \), where the first term on the right represents the possibility that the two genes have the same ancestor already one generation earlier. We rewrite this to

\[
1 - F_t = \left(1 - \frac{1}{N}\right) (1 - F_{t-1})
\]

so it is easy to iterate, giving

\[
F_t = 1 - \left(1 - \frac{1}{N}\right)^t \approx 1 - e^{-t/N}.
\]

This is the cumulative probability of an exponential distribution:

\[
\int \lambda e^{-\lambda t} dt = -e^{-\lambda t}, \text{ with parameter } \lambda = 1/N.
\]

In this context, we call the parameter the rate of coalescence. Note that for \( t = N \), the probability of having a common ancestor is already about \( 1 - \frac{1}{e} = 0.632 \ldots \). We see that populations become inbred over a timescale of \( t \) on the order of \( N \) generations.

Collection of genes. If we trace a sample of \( k \) genes back through time, then coalescence of any pair of genes is rare. As a consequence, simultaneous coalescence of two pairs of genes is even rarer and can be neglected. The same holds true for the coalescence of three or more genes in one common parent. To see that this is really the case, we work out the probability of a few events in a large population, in which \( N \) is much larger than \( k \). There is no coalescence if all \( k \) genes have distinct predecessors one generation earlier. The probability of this event is that of hashing \( k \) items into \( N \) slots without collision:

\[
P_0 = \frac{N - 1}{N} \cdot \frac{N - 2}{N} \cdot \ldots \cdot \frac{N - k + 1}{N}
\]

\[
\approx \exp \left[ -1 - 2 - \ldots - (k - 1) \right] \frac{1}{N}
\]

\[
\approx \exp \left[ -\frac{k(k-1)}{2N} \right]
\]

\[
\approx 1 - \frac{k(k-1)}{2N}.
\]

Similarly, the probability that there is one coalescence is that of having one collision:

\[
P_1 = \binom{k}{2} \cdot \frac{1}{N} \cdot \frac{N - 1}{N} \cdot \frac{N - 2}{N} \cdot \ldots \cdot \frac{N - k + 2}{N}
\]

\[
\approx \frac{k(k-1)}{2N} \cdot \exp \left[ -(k - 1)(k - 2) \right] \frac{1}{N}
\]

\[
\approx \frac{k(k-1)}{2N}.
\]
where we drop the exponential term since its contribution is of a lower order than the leading term. To first order approximation, \( P_0 \) and \( P_2 \) add up to 1. For comparison, we also compute the probability that there are two pairwise coalescences:

\[
P_{22} = 3 \binom{k}{4} \frac{N - 1}{N^3} \frac{N - 2}{N} \frac{N - 3}{N} \cdots \frac{N - k + 4}{N}
\approx \frac{k(k-1)(k-2)(k-3)}{8N^2},
\]

where the first term counts the 4-tuplets times the ways to partition a 4-tuplet into two pairs. Contrast this with the probability that there is one three-way coalescence, we notice that the latter is less likely that two pairwise coalescence, at least if \( k \) is not too small:

\[
P_3 = \binom{k}{3} \cdot \frac{1}{N^2} \cdot \frac{N - 1}{N} \cdots \frac{N - k + 3}{N}
\approx \frac{k(k-1)(k-2)}{6N^2}.
\]

In summary, \( P_{22} \) and \( P_3 \) are of lower order than \( P_2 \), which justifies that we ignore events beyond the coalescence of a pair.

**The coalescent process.** Beyond the Wright-Fisher model, we can imagine a wide variety of ways to formalize reproduction, in either discrete of continuous time. Remarkably, provided that all individuals are equivalent, and the variance of the distribution of offspring number is finite, then in the limit of large population size, all such models converge to the same, continuous coalescent process. This process depends on a single parameter, \( N_e \), which we refer to as the effective population size. This limiting model can be stated very simply:

- tracing back in time, any pair of lineages coalesce at a rate \( 1/N_e \),
- simultaneous coalescence events as well as coalescence amongst more than two lineages are negligible.

To understand the importance of the variance of offspring number, we generalize the Wright-Fisher model to allow for an arbitrary distribution of offspring number. Suppose that individual \( i \) has \( j_i \) offsprings; necessarily, \( \sum_{i=1}^{N} j_i = N \). If we now pick a parent at random, the expected number of offsprings it has is \( E(j) = \sum_{i=1}^{N} j_i/N = 1 \). This implies that the variance is

\[
V(j) = \frac{1}{N} \cdot \sum_{i=1}^{N} (j_i - 1)^2
= \frac{1}{N} \cdot \sum_{i=1}^{N} j_i(j_i - 1),
\]

because \( \sum(1-j_i)/N = 1 - E(j) = 0 \). Dividing by \( N - 1 \) gives

\[
\frac{1}{N_e} = \sum_{i=1}^{N} \frac{j_i}{2} \frac{N}{2},
\]

which is also the probability that two given lineages coalesce in one generation. As mentioned earlier, we call this probability the rate of coalescence. Note that with a Poisson distribution, \( V(j) = E(j) = 1 \), so that we get \( N_e = N - 1 \). In the limit of large populations, this becomes \( N_e = N \), which is the reference case stated in textbooks.

**Properties.** Using \( f(t) = (1/N_e) \exp[-t/N_e] \) as the density function, we note that the expected time of coalescence for a given pair is

\[
E(T_2) = \int_0^{\infty} \frac{t}{N_e} e^{-t/N_e} dt
= \left[ -t e^{-t/N_e} - N e^{-t/N_e} \right]_0^{\infty},
\]

which evaluates to \( N_e \). Generalizing this to the time \( T_k \) it takes for \( k \) lineages to coalesce down to \( k - 1 \), we get \( E(T_k) = N_e / \binom{k}{2} \). The expected time back to the most recent common ancestor – the root of the tree – is therefore

\[
E(T) = E(T_2 + T_3 + \ldots + T_k)
= 2N_e \left( \frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \ldots + \frac{1}{k(k-1)} \right)
= 2N_e \left( 1 - \frac{1}{k} \right). \tag{26}
\]

For a very large number of genes (\( N \) much larger than \( k \), which is much larger than 1), it takes on average \( N_e \) generations to trace back to two lineages, and then the same time (on average) to trace back to the common ancestor. This implies that when studying a single gene, it is not worth collecting large samples.

Another interesting measure is the length of the tree, from the current population of \( k \) genes back to the root of the tree. The expected length is

\[
E(L) = E(2T_2 + 3T_3 + \ldots + kT_k)
= 2N_e \left( \frac{2}{2} + \frac{3}{6} + \frac{4}{12} + \ldots + \frac{k}{k(k-1)} \right)
= 2N_e \sum_{j=1}^{k-1} \frac{1}{j} \tag{30}
\]

which is approximately \( 2N_e (\gamma + \log k) \), with \( \gamma = 0.577 \ldots \). In order to double the expected length of the tree, we therefore need to take the square of the number of genes, so we get progressively less information per gene as we increase their number.

**Mutation.** We cannot observe genealogies directly, but we can see which mutations are carried by each individual. Assume the infinite sites model of mutation: each individual carries a DNA sequence with a very large number of sites, so that every mutation is unique, and can be identified. Assuming mutations occur at a steady rate \( \mu \) per individual per generation, the number
of mutations on each branch of the tree follows a Poisson distribution with expectation $\mu t$, where $t$ is the length of the branch.

Suppose that we sample two individuals, and recall that the expected time back to a common ancestor is $E(T_2) = N_e$. The expected number of mutations that distinguish them is therefore $2\mu N_e$. The probability that they are distinguished by $j$ mutations is:

$$\int_0^\infty \frac{e^{-t/N_e}}{N_e} e^{-2\mu t} \frac{(2\mu t)^j}{j!} \, dt = \frac{(2N_e\mu)^j}{(1 + 2N_e\mu)^{j+1}}.$$  

This is a geometric distribution with mean $N_e\mu$ and variance $2N_e\mu(1 + 2N_e\mu)$.

We can also ask how many mutations are found on the whole tree: this is the number of distinct mutations that will be seen in the sample of $k$ individuals. The expected number is just $\mu E(L) \approx 2N_e\mu(\gamma + \log k)$. The ratio between the number of distinct mutations, and the average number of mutations that distinguish two individuals, is approximately $\gamma + \log k$, which provides a simple test of the null model of neutral evolution.

**Extensions.** The standard coalescent process can be extended in several ways:

- If the variance of offspring number is infinite, then there is a different limiting process, in which there may be multiple simultaneous coalescence events.
- We can consider a set of genes that recombine with each other, through sexual reproduction, or even, a long linear chromosome. This is the natural model for analyzing DNA sequence data, and is very easy to define. However, it is very hard to analyze or simulate.
- The coalescent assumes that individuals are indistinguishable. If individuals that carry different mutations reproduce at different rates, there is selection. This is even harder to analyze than recombination.