Coalescent theory

Richel Bilderbeek

March 20, 2015

1 The model

The model used is a Wright-Fisher model (also depicted in figure 1) that follows the inheritance of an allele in a constant-size population of random-mating haploid individuals. The term 'random mating' does not fully apply, as the individuals are haploid. What is the case, is that paernts create offsping, of which only a certain constant number survive to reproduce. The generations are strictly separated.

2 Derivation of the likelihood

Assume four individuals sampled ($n = 4$)(see also table 1) from a much larger population ($N \gg n$) being at present time ($t = 0$). These four haploid individuals have a parent, yet, together, they may have one single to four different parents. Assuming a large population ($N \gg n$) we can expect either zero or one coalescent event (that is: the four individuals have either four or three different parents). When assigning parents to the focal individuals, there are three ways a coalescence can take place:

- when assigning the second parent, which might be the first individual's parent as well. This equals the chance of the second coalescing with the

Symbol	Description
$\mathcal{L}$	Likelihood
l	Number of lineages, $l = n$ at $t = 0$
n	Sample size, number of individuals sampled
N	(constant) effective population size, individuals
t	Time, the generation number
$t = 0$	Current, present-day generation
$t = 1$	Previous generation, parents of present-day generation
ts	Vector of branching times, $ts = \{t_0, t_1, t_n\}$

Table 1: Symbol descriptions

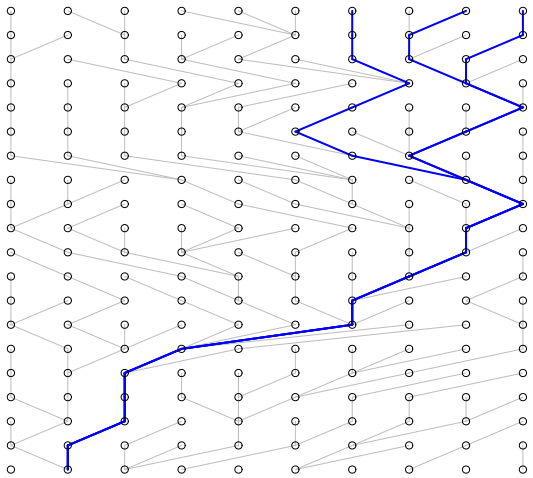

Figure 1: Wright-Fisher model. The circles represent haploid individuals of multiple generation. Each (horizontal) generation consists out of ten individuals. In this example there are twenty generations, going from the current time (top) into the past (bottom). The grey lines show the inheritance of genes from parents to offspring.

first $(1/N)$ multiplied by the chance the third individual does not coalesce $(1-(1/N))$ multiplied by the chance the fourth individual does not coalesce $(1-(2/N))$:

$$p_{i=2}(n=4,N,t=1)=\left(\frac{1}{N}\right)\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)$$

- when assigning the third parent, which might be the first or second individual's parent as well

$$p_{i=3}(n=4,N,t=1)=\left(1-\frac{1}{N}\right)\left(\frac{2}{N}\right)\left(1-\frac{2}{N}\right)$$

- when assigning the fourth parent, which might be the first,second or third individual's parent as well

$$p_{i=4}(n=4,N,t=1)=\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\left(\frac{3}{N}\right)$$

Summing these results in the chance there will be a coalescent event between now and the previous generation:

$$p(n=4,N,t=1)=\left(\frac{1}{N}\right)\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)+\left(1-\frac{1}{N}\right)\left(\frac{2}{N}\right)\left(1-\frac{2}{N}\right)+\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\left(\frac{3}{N}\right)$$

This equation is for 4 lineages. This can be generalized for n lineages. First extract the common pattern from the equation above:

$$p(n=4,N,t=1)=\left(\frac{1}{N}\right)\prod_{i=1}^{n-2}\left(1-\frac{i}{N}\right)+\left(\frac{2}{N}\right)\prod_{i=1}^{n-2}\left(1-\frac{i}{N}\right)+\left(\frac{3}{N}\right)\prod_{i=1}^{n-2}\left(1-\frac{i}{N}\right)$$

$$p(n,N)=\sum_{j=1}^{n-1}\left[\left(\frac{j}{N}\right)\prod_{i=1}^{n-2}\left(1-\frac{i}{N}\right)\right]$$

Assuming a large population $(N \gg n)$, the product term can be left out $(1-(i/N)=1)$, resulting in

$$p(n,N,t=1)=\sum_{j=1}^{n-1}\frac{j}{N}=\frac{1}{N}\sum_{j=1}^{n-1}j=\frac{1}{N}.\frac{n(n-1)}{2}=\frac{1}{N}\left(\begin{array}{c}n\\2\end{array}\right)$$

If there is a coalencent event t timesteps back in time, this generalizes to:

$$p(n,N,t)=\frac{1}{N}\left(\begin{array}{c}n\\2\end{array}\right)\left[1-\frac{1}{N}\left(\begin{array}{c}n\\2\end{array}\right)\right]^{t-1}$$

The discrete time equation at the left-hand side can be approximated by the continuous equation at the right-hand side of equation 1:

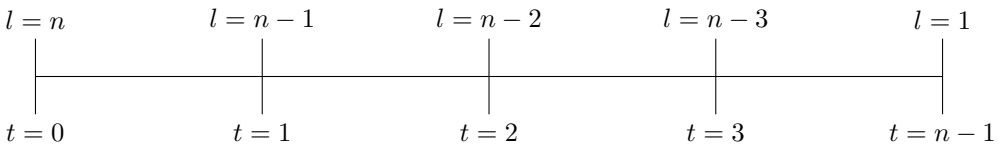

Figure 2: Timeline from current time (left) to the past (right)

$$\left[1 - \frac{1}{N}\begin{pmatrix} n \\ 2 \end{pmatrix}\right]^{t-1} \approx e^{-\begin{pmatrix} n \\ 2 \end{pmatrix}\frac{t_i}{N}} \tag{1}$$

A tree structure consists out of branch lengths, where the different branches have a different number of lineages (l) present (see figure 2).

This results in the likelihood of finding this structure (i represents the number of lineages, l):

$$\mathcal{L}(n, N, ts) = \prod_{i=2}^{n}\left[\frac{1}{N}\begin{pmatrix} i \\ 2 \end{pmatrix} e^{-\begin{pmatrix} i \\ 2 \end{pmatrix}\frac{t_i}{N}}\right]$$

3 Adding mutation

The change to find a certain number of mutations after some generations follows a Poisson distribution

$$Poisson(k, \lambda) = \frac{\lambda^k . e^{-\lambda}}{k!} = \left[e^{-\lambda}\right]\frac{\lambda^k}{k!}$$

The change P to find k mutations after t generations is (equation [12])(which is the Poisson distribution with λ replaced by μt)

$$P(k|t) = \left[e^{-\mu t}\right]\frac{(\mu t)^k}{k!}$$

The expected number of mutation

4 Acknowledgements

Thanks to César Martinez for demonstrating the derivation. Thanks to Guillaume Achaz for his document 'Introduction a la coalesce'.