The coalescent – coalescence theory

Peter Beerli

September 1, 2009

Historical note

Up to 1982 most development in population genetics was prospective and developed expectations based on situations of today. Most work did provide expectations about the future. With the easy availability of genetic data retrospective analyses did catch up only in phylogenetics (starting in the sixties). Only Malécot, who pioneered "looking backwards in time" in 1948, developed "backwards looking" results in population genetics. Kingman expressed this "looking backwards in time" approach as the coalescence of sampled lineages. He was not the only one working on such problem at the the time as with many great solutions it was in the air, see Hudson (1983) and Tajima (1983).

The coalescent

A sample of n gene copies is taken at the present time and we are interested in the ancestral relationship of these gene copies. We express time τ increasing the further back in real time we go: $\tau_1 < \tau_2$ means that τ_2 is further in the past than τ_1. Kingman (1982) and Ewens (2004) describe this backwards in time process with equivalence classes. Two copies are in the same equivalence class at time τ when they have a common ancestor at that time. At time $\tau = 0$ each individual gene can be considered in its own equivalence class and we could express this for a sample of $n = 8$ as

$$\phi_0 = \{(a), (b), (c), (d), (e), (f), (g), (h)\}$$

Kingman's n-coalescent describes the moves from ϕ_0 to a single equivalence class

$$\phi_n = \{(a, b, c, d, e, f, g, h)\}.$$

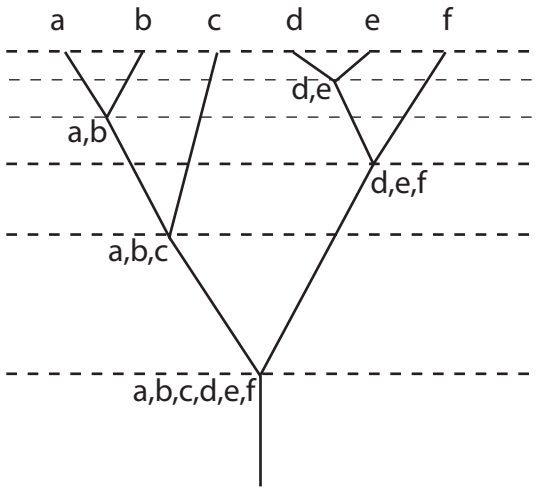

Figure 1: Example of the coalescence process

All individuals are in some equivalence relation ξ and we can find a new equivalence relation η by joining two of the equivalence classes in ξ. This joining process is called a *coalescence*, and a series of such joinings is called the *coalescent* or *coalescence process*. Figure 2 gives an example of the relationship of a sample and the equivalence classes describing the process. It is assumed that the probability of a coalescence depends on the waiting time $\delta\tau$

$$\text{Prob(process in } \eta \text{ at time } \tau + \delta\tau \mid \text{process in } \xi \text{ at time } \tau) = \delta\tau$$

(ignoring higher order terms), and if k is the number of equivalence classes in ξ then

$$\text{Prob(process in } \xi \text{ at time } \tau + \delta\tau \mid \text{process in } \xi \text{ at time } \tau) = 1 - \frac{k(k-1)}{2}\delta\tau = 1 - \binom{k}{2}\delta\tau$$

We will see that if we apply the right time scale to τ then we will end up in the more familiar terms that are common in the applied population genetics literature.

Kingman focused on the Canning model, and since the Wright-Fisher model is a special case of the results carry over easily. The coalescent is an approximation to these models because it was developed on a continuous time scale whereas the Canning and Wright-Fisher population models have discrete time. Any findings using this coalescent machinery needs to be rescaled to the time scale of these discrete time models. In the coalescent framework one has only a single coalescent per infinitesimal time period. This forces us to restrict the use of the coalescent to discrete time models were we can guarantee that there is not more than one coalescent event occurring per time period. For example, for a Wright-Fisher population we can allow only one coalescent event per generation. this sound rather restrictive but as long as the sample n is much smaller than the

population N this situation rarely occurs. Fu (2006) calculates that a sample needs to be less than the square-root of N

$$n < \sqrt{N},$$

when we use the coalescent for a model such as the Wright-Fisher or some version of the Canning model. Joe Felsenstein shows this for the discrete case in the Wright-Fisher model in his course "population genetics" (http://evolution.genetics.washington.edu/pgbook/pgbook.html).

The "discrete" coalescent and the Wright-Fisher population model

In the Wright-Fisher model we can calculate the probability that two random individuals have a common ancestor one generation earlier, very simply. Picking two individuals at random, and then decide the chance of having a common ancestor: the first individual has with probability 1 an ancestor in the last generation and the other had with probability $1/(2N)$ the same ancestor. In turn we also know the probability that the two individuals have no common ancestor in the last generation: $1 - 1/(2N)$. When we have n individuals in the sample then we can have $\binom{n}{2}$ pairs of individuals that coalesce with a chance of $1/(2N)$, therefore we have ever generation chance to see a coalescence of two individuals is

$$\frac{\binom{n}{2}}{2N} = \frac{n(n-1)}{4N} \tag{1}$$

This rate of coalescence in the discrete Wright-Fisher model is similar to a coin toss where we wait with some rate (that the head appears), this looks like a geometric series, but in the case of the coalescence the series does not use a fixed rate, but the rate changes because the coalescent event reduces the number in the sample by 1. The expected value of a geometric is approximately 1/rate, therefore the waiting time to reduce from n lineages to $n-1$ lineages is

$$\mathbb{E}(T_{(n-1|n)}) = \frac{4N}{n(n-1)} = 4N(1 - \frac{1}{n}) \tag{2}$$

To reach the final coalescent (the most recent common ancestor) we need to wait until all lineages have coalesced we simply add up all the waiting times

$$\mathbb{E}(T_{\mathrm{MRCA}}) = 4N \sum_{k=2}^{n} \frac{1}{k(k-1)} \tag{3}$$

The (continous) coalescent and the Wright-Fisher population model

The coalescent process is in effect a sequence of $n-1$ Poisson processes[1], with rates

$$r_k = \frac{k(k-1)}{2}, k = n, n-1, n-2, ..., 2$$

describing the Poisson process at which two of the equivalence classes merge when there are k equivalence classes.

Since these events are coming from a Poisson distribution we can calculate the expectation for each interval, which is 1/rate, here $2/(k(k-1))$. All mergers of the equivalence classes are independent of each other so the expectation of the whole coalescence process is

$$\mathbb{E}(T_{\text{MRCA}}) = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n} \frac{1}{k(k-1)}$$

Comparison of the content of the sum with the coalescence rate makes clear that the Wright-Fisher population model is $2\times$ the standard coalescent units and we need to multiply by $2N$ to arrive at the more familiar generation time scale.

The coalescent process results in a tree of a sample of n individuals. We call this often a genealogy as the the individuals are typically from the same species or population (hence population size).

Often the probability of the coalescent process for the Wright-Fisher population model is expressed as

$$p(G|N) = \prod_{k=2}^{n} e^{-u_k \frac{k(k-1)}{4N}} \frac{2}{4N},$$

where u is expressed in generations. The expected T_{MRCA} is the same as above. Often we will not use a time scale in generations but generations $\times$ mutation rate and then we would express the above formula as

$$p(G|\Theta) = \prod_{k=2}^{n} e^{-u_k \frac{k(k-1)}{\Theta}} \frac{2}{\Theta},$$

where Θ is $4 \times N_e \times \mu$, with μ as the mutation rate per generation and site (when using sequence data), and N_e as the effective population size. Under a strict Wright-Fisher population model

[1] From Mathworld: A Poisson process is a process satisfying the following properties:

1. The numbers of changes in non-overlapping intervals are independent for all intervals.
2. The probability of exactly one change in a sufficiently small interval is , where is the probability of one change and is the number of trials.
3. The probability of two or more changes in a sufficiently small interval is essentially 0.
   In the limit of the number of trials becoming large, the resulting distribution is called a Poisson distribution.

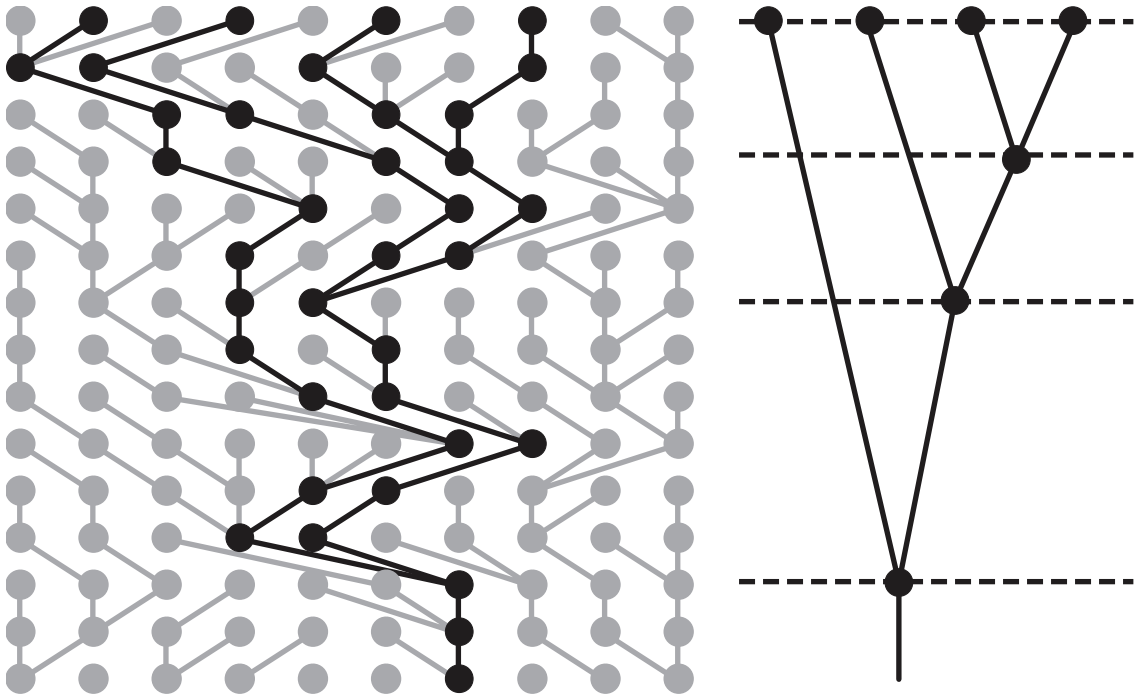

Figure 2: Example of a coalescence structure in the Wright-Fisher model

$N = N_e$, but under more biological scenarios one needs to know more about the life history of the species to translate the N_e into real numbers.

The coalescent and the Moran population model

The coalescent is an *exact* representation of the Moran model because the problems with the multiple coalescent events in one generation do not occur. The Moran model allows only one lineage to change at a given time. Therefore the limitation to small sample size as we have seen for the Wright-Fisher model is not needed.

Using our findings of the discussion of the Moran model earlier, but instead of thinking forward in time, think backward in time. Looking backwards we see that the Moran process is similar structured like the coalescence process. We have n individuals that are reduced in their ancestry to $n-1, n-2, ...$ and eventually to one gene, the most common recent ancestor of the n sampled individuals. Assume that we are at a time where we have a sample of k individuals, these are descendants of $k-1$ parents of one of these parents was chosen to reproduce and the offspring is

in ancestry of the sample of n genes. The probability of this event is

$$\frac{k(k-1)}{(2N)^2},$$

and with probability

$$1 - \frac{k(k-1)}{(2N)^2},$$

the ancestors remain at j. Tracing back this ancestry the number of death and birth events between the times when there are j and $j-1$ ancestors follows a geometric distribution with parameters $k(k-1)/(2N)^2$ and thus has a mean of

$$\mathbb{E}(u_j) = \frac{(2N)^2}{k(k-1)}$$

now we can assemble the expectation for the time to the most recent common ancestor

$$\begin{aligned}\mathbb{E}(T_{\text{MRCA}}) &= \sum_{k=2}^{n} \frac{(2N)^2}{k(k-1)} = (2N)^2(1-\frac{1}{n}) \\ &= (2N)^2 \sum_{k=2}^{n} \frac{1}{k(k-1)} \qquad (4)\end{aligned}$$

If we assume that we sampled the whole population, where $n = 2N$ than we derive the same result as with standard (forward) theory.

$$\mathbb{E}(T_{\text{MRCA}}) = (2N)^2(1-\frac{1}{n}) = (2N)^2(1-\frac{1}{2N}) = 2N(2N-1)$$

We also can make the same observation as we made with the Wright-Fisher population model. The coalescence time scale is by a factor $(2N)^2$ different from the Moran model time scale (formula 4).

We can express the probability of the genealogy under the Moran model using the fact that the exponential distribution is a good approximation to the geometric distribution

$$p(G|N) = \prod_{k=2}^{n} e^{-u_k \frac{k(k-1)}{2(2N)^2}} \frac{1}{(2N)^2},$$

where u is expressed in generations. The expected T_{MRCA} is the same as above. Often we will not use a time scale in generations but generations $\times$ mutation rate and then we would express the above formula as

$$p(G|\Theta) = \prod_{k=2}^{n} e^{-u_k \frac{2k(k-1)}{\Theta^2}} \frac{4}{\Theta^2},$$

where Θ is $4 \times N_e \times \mu$, with μ as the mutation rate per generation and site (when using sequence data), and N_e as the effective population size. Under a strict Wright-Fisher population model $N = N_e$, but under more biological scenarios one needs to know more about the life history of the species to translate the N_e into real numbers.