The Least Coalescent Time for Two Individuals

Zhao, Zehui

1 Basic Coalescence

Suppose there are N single chromosome haploid hermaphrodites, and a new generation with the same population is generated in each time step by randomly having offsprings from the previous generation. Abstractly, each generation is a list of N integers randomly generated from 1 to N that signify the parents of the individuals of this generation. In this case, for a current generation at $t = 0$, it should also be generated from the previous generation through the same process, and two individuals should have the same parent if the two numbers representing their parents are generated to be the same. So the probability of two individuals coalescing exactly at the previous generation is

$$P(1) = \frac{1}{N},$$

and the probability of coalescing exactly at the t-th previous generations is

$$P(t) = (1 - \frac{1}{N})^{t-1}\frac{1}{N}.$$

2 Finite time steps and continuous time

Suppose the time between two generations is broken into smaller pieces, where there are M time steps between two generations. Then the reasoning above can be carried over, and

$$P(t) = (1 - \frac{1}{NM})^{Mt-1}\frac{1}{NM},$$

where t can now take some values between integers. By defining the length of every time step as

$$dt = \frac{1}{M}$$

and taking the limit of this result as M goes to infinity, one can obtain the continuous probability distribution. To do this, first note that every point in time in a continuous domain has measure 0, and $P(t)$'s previously really correspond to $P(t)\,dt$ in continuous time and may be written as $P(t-1,t)$ and $P(t-1/M,t)$ if one likes. So by taking the limit,

$$\begin{aligned} P(t)\,dt &= \lim_{M\to\infty}(1 - \frac{1}{NM})^{Mt-1}\frac{1}{NM} \\ &= \frac{1}{N}\lim_{M\to\infty}(1 - \frac{1}{NM})^{Mt-1}\,dt \\ &= \frac{1}{N}\exp(-\frac{1}{N}t)\,dt. \end{aligned}$$

3 Adding simplified recombination

Now return to integer times and suppose that there is a recombination rate π, being the number of recombinations per length per generation/unit time. Interpret this recombination rate as deterministic, so that a recombination occurs every time the recombination rate accumulates to an integer over time. Then for two individuals each with a chromosome of length L after a time t, the number of relevant recombinations occurred should be

$$2\pi Lt =: Rt,$$

where a floor function may be applied if one wants, and R here is the rate of having a recombination relevant to the problem. This means at the t-th previous generation, there are in total $2 + Rt$ ancestors for the two individuals, and the probability of coalescing in this generation (ignoring the probability of reaching this state) is

$$P = \frac{Rt + 1}{N}$$

if one considers only the "true" coalescences, that is the ones in which two bases in the two ancestors' chromosomes coalesce. An argument for this can be found in the appendix. So the probability of the two individuals coalescing at the t-th previous generation is

$$P(t) = \frac{Rt + 1}{N} \prod_{t'=1}^{t-1} (1 - \frac{Rt' + 1}{N}).$$

One can again break time into finite time steps and take the continuous time limit. This time,

$$P(t) = \frac{Rt + 1}{NM} \prod_{i=1}^{Mt-1} (1 - \frac{Ri/M + 1}{NM})$$

for discrete time, and

$$\begin{aligned} P(t)\,dt &= \lim_{M\to\infty} \frac{Rt + 1}{NM} \prod_{i=1}^{Mt-1} (1 - \frac{Ri/M + 1}{NM}) \\ &= \frac{Rt + 1}{N} \lim_{M\to\infty} \prod_{i=1}^{Mt-1} (1 - \frac{Ri/M + 1}{NM})\,dt \end{aligned}$$

for continuous time. But since the limit of the product is analytically difficult to handle, consider a different approach, where the complementary CDF $F(t)$ is considered. Assume one starts with a large number of trails, where each trial's earliest coalescent time is recorded. Then in every time step under discrete time, the number of trials reduced from the remaining trials (the trials that have not coalesced yet) is given by the probability P/N of coalescing within one time step with $2 + Rt$ lineages. Then the fraction of trials remained is

$$F(t + dt) = F(t)(1 - \frac{Rt + 1}{NM}),$$

and rearranging then taking the limit of dt going to 0 gives the differential equation

$$\frac{F(t+dt)-F(t)}{dt} = -\frac{Rt+1}{N}F(t)$$
$$\frac{dF}{dt} = -\frac{Rt+1}{N}F(t).$$

Solving this differential equation and normalizing gives

$$F(t) = \exp(-\frac{R}{2N}t^2 - \frac{1}{N}t),$$

and taking its negative derivative yields the PDF

$$P(t) = \frac{Rt+1}{N}\exp(-\frac{R}{2N}t^2 - \frac{1}{N}t).$$

Now consider nondimensionalize t by defining

$$t = T\tau.$$

Then the CCDF becomes

$$F(\tau) = \exp(-\frac{RT^2}{2N}\tau^2 - \frac{T}{N}\tau),$$

where T can be anything for every setup. Now consider choosing T to be N generations in every setup. Then by defining

$$S = RT,$$

the CCDF becomes a function of only one parameter

$$F(\tau) = \exp(-\frac{RT}{2}\tau^2 - \tau)$$
$$= \exp(-\frac{S}{2}\tau^2 - \tau).$$

The form of the CCDF makes sure that it is always normalized, and taking its negative derivative gives

$$P(\tau) = (S\tau+1)\exp(-\frac{S}{2}\tau^2 - \tau).$$

A Probability of coalescing in one generation with n lineages

To find the probability, first note that this problem is the same as counting the number of overlapping pieces of two unit intervals partitioned into a total of n pieces. Here, each individual's chromosome is represented by one unit interval, and each lineage of an individual is represented by one piece partitioning the corresponding interval. Then, the probability of coalescing in one generation is the

total probability of two overlapping pieces from different intervals choosing the same parent in the previous generation. Since the probability of choosing the same parent is

$$P = \frac{1}{N}$$

for each pair, the coalescing probability is

$$P = \frac{m}{N},$$

where m is the number of overlapping pairs.

Now consider the following proof by induction that $m = n - 1$ for every n.

- When the two intervals are intact, $n = 2$ and $m = 1$, $m = n - 1$.
- If the two intervals are broken into $n = a+1$ pieces, one can consider two adjacent pieces. Then, the total number of pairs is the number of pairs when the two adjacent pieces are combined plus the change in number of pairs due to the separation of the two. By the assumption of induction, the first number is $a - 1$. Since each piece matching with the combined piece can only match with one of the two separated pieces except the piece covering the point of separation, the second number is 1. Note here that there will definitely be a piece covering the point of separation, since it is assumed that no two recombinations can occur at the same locus. Therefore

$$m = (a - 1) + 1 = a = (a + 1) - 1 = n - 1.$$

- So $m = n - 1$ for every $n \geq 2$.

Therefore the probability of coalescing in one generation with n lineages is

$$P = \frac{m}{N} = \frac{n - 1}{N}.$$