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 Adding recombination and some coalescing restrictions

Now suppose each new generation is not only random offsprings from the previous generation, but due to recombination, each offspring may also inherit a piece of genetic information from some individual other than its parent. Abstractly, when generating every new individual in the next generation, the individual may have an ordered 3-tuple instead of an integer with chance p, where the first entry in the 3-tuple signifies the ordinary parent, the second entry is also randomly generated between 1 and N and signifies the recombining parent, and the last entry signifies the location of recombination. Let the individual inherit the first half (before the location of recombination) from its ordinary parent and the second half from the recombining parent. Here, the total length L of the chromosome is actually relevant, since

$$p = L\pi$$

where π is the recombination rate per length in a generation, and the third entry is between 1 and L. Then at any previous generation, a current individual may have zero, one, or multiple ancestors, where each ancestor may only contribute a small piece to the individual's genome.

The computation will quickly become inefficient if one considers all possible geneologies. So consider adding restrictions on the random number generators in the model. First of all, forbid any two individuals in any previous generation from having the same parent, if they are both ancestors of one individual in the current generation. This means each current individual's ancestors cannot coalesce, and its number of ancestors can only grow or stay the same. Second, forbid any two ancestors of different individuals in the current generation from having the same parent. As an

example, suppose individuals A and B in the t-th previous generation both contributed to the genome of the individual C in the current generation, such that one segment of the chromosome of C is from A and a different piece is from B. Then the second restriction would forbid A and B from having the same parent.

Then by counting and drawing some diagrams of the split chromosomes, if the two individuals in the current generation have m ancestors (so $m-2$ recombinations) in the $(t-1)$-th generation, then the probability of them coalescing at the t-th generation is exactly

$$P_m = \frac{m-1}{N}$$

without the restrictions and approximately (?) the same with restrictions.

Since the restrictions are in place, the probabilities of coalescence are not difficult to compute. Take the probability of (two individuals) coalescing exactly at the 2-nd previous generation $P(2)$. Since there has been two generations traced back, including the 2-nd one, there are at most 2 recombinations. So

$$P(2) = p^2 P_{11}(2) + p(1-p)P_{10}(2) + (1-p)pP_{01}(2) + (1-p)^2 P_{00}(2),$$

where the sequences of 0 and 1 signify if a recombination occurred at the corresponding place. So $P_{01}(2)$ is the probability of coalescing at the 2-nd previous generation if there isn't a recombination in the 1-st previous generation and there is one in the 2-nd. By using the result for P_m,

$$\begin{aligned}
P_{11}(2) &= (1-P_3)P_4 = (1-\frac{2}{N})\frac{3}{N} \\
P_{10}(2) &= (1-P_3)P_3 = (1-\frac{2}{N})\frac{2}{N} \\
P_{01}(2) &= (1-P_2)P_3 = (1-\frac{1}{N})\frac{2}{N} \\
P_{00}(2) &= (1-P_2)P_2 = (1-\frac{1}{N})\frac{1}{N},
\end{aligned}$$

and

$$\begin{aligned}
P(2) &= p^2(1-\frac{2}{N})\frac{3}{N} + p(1-p)(1-\frac{2}{N})\frac{2}{N} + (1-p)p(1-\frac{1}{N})\frac{2}{N} + (1-p)^2(1-\frac{1}{N})\frac{1}{N} \\
&\approx p^2(1-\frac{3}{N})\frac{3}{N} + 2p(1-p)(1-\frac{2}{N})\frac{2}{N} + (1-p)^2(1-\frac{1}{N})\frac{1}{N}.
\end{aligned}$$

This approximation should have a small effect because when N is large, $(1-\frac{a}{N}) \approx (1-\frac{b}{N})$ when

$|a-b| \ll N$. As another example of the approximation,

$$
\begin{aligned}
P_{111}(3) &= (1-P_3)(1-P_4)P_5 = (1-\frac{2}{N})(1-\frac{3}{N})\frac{4}{N} \approx (1-\frac{4}{N})^2\frac{4}{N} \\
P_{110}(3) &= (1-P_3)(1-P_4)P_4 = (1-\frac{2}{N})(1-\frac{3}{N})\frac{3}{N} \approx (1-\frac{3}{N})^2\frac{3}{N} \\
P_{101}(3) &= (1-P_3)(1-P_3)P_4 = (1-\frac{2}{N})(1-\frac{2}{N})\frac{3}{N} \approx (1-\frac{3}{N})^2\frac{3}{N} \\
P_{011}(3) &= (1-P_2)(1-P_3)P_4 = (1-\frac{1}{N})(1-\frac{2}{N})\frac{3}{N} \approx (1-\frac{3}{N})^2\frac{3}{N} \\
P_{100}(3) &= (1-P_3)(1-P_3)P_3 = (1-\frac{2}{N})(1-\frac{2}{N})\frac{2}{N} \approx (1-\frac{2}{N})^2\frac{2}{N} \\
P_{010}(3) &= (1-P_2)(1-P_3)P_3 = (1-\frac{1}{N})(1-\frac{2}{N})\frac{2}{N} \approx (1-\frac{2}{N})^2\frac{2}{N} \\
P_{001}(3) &= (1-P_2)(1-P_2)P_3 = (1-\frac{1}{N})(1-\frac{1}{N})\frac{2}{N} \approx (1-\frac{2}{N})^2\frac{2}{N} \\
P_{000}(3) &= (1-P_2)(1-P_2)P_2 = (1-\frac{1}{N})(1-\frac{1}{N})\frac{1}{N} \approx (1-\frac{1}{N})^2\frac{1}{N}.
\end{aligned}
$$

So the pattern can be used to find in general the probability of coalescing exactly at the t-th previous generation

$$
P(t) \approx \sum_{i=0}^{t} \binom{t}{i} (1-p)^{t-i} p^i (1-\frac{i+1}{N})^{t-1} \frac{i}{N}.
$$

Mathematica shows that the graph looks like:

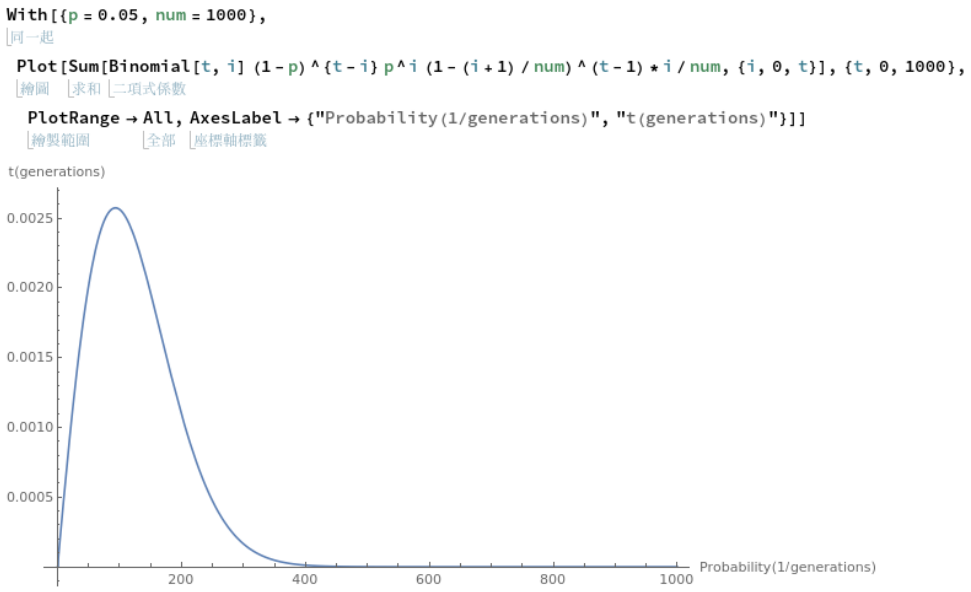

Figure 1: The code used to plot the curve and the curve are shown above.