The Least Coalescent Time for Two Individuals

Zhao, Zehui

1 The basic model

Suppose a population has $2N$ individuals, N females and N males. Each generation, males and females mate randomly to generate a new population of size $2N$, again with N individuals each sex. Suppose this species has only one pair of chromosomes with ploidy 2. Then for two individuals at this generation, one can trace back in their ancestry and see at which generation do the four chromatids have a common source or ancestor. The number of generations here is called the coalescent time. Note that this question is not the same as when do the two individuals has a common ancestor. The question of interest is related to genetics, but the similar question is not.

To avoid drawing pictures, one can label the $2N$ individuals from 1 to $2N$, and call them the 0-th generation. But since the subject of the investigation is the chromatids not the individuals, one may add another entry to the label, indicating which chromatid is the referent. Then, for example, $(3, 1)$ refers to the first chromatid of the third individual. Since the population is fixed, the previous generation can also be labeled from 1 to $2N$, and one can distinguish those from the next generation by calling this generation the (negative) 1-st generation. By repeating this procedure, every chromatid that has ever existed will have a unique label, being an ordered 3-tuple in $\{n \leq 2N\} \times \mathbb{N} \times \{1, 2\}$, and every label will refer to an individual. Using this labeling, one can generate a carrier graph for each chromatid of interest. Suppose one starts at $(0, 3, 1)$, and the third individual of the current generation inherited this chromatid from the seventh individual's second chromatid of the previous generation, then the graph becomes $(0, 3, 1) - (1, 7, 2)$. Now with the four chromatids at present, one can generate four graphs by tracing back in their ancestry. And the question becomes at which generation do the four graphs become connected.

To keep statements general, one can look at not the coalescent time of individuals with a fixed, known ancestry, but the expected coalescent time of randomly generated ancestries. Take two chromatids that belong to two different individuals. Then the probability that the coalescent time is 1 generation is

$$P_{\text{diff}}(t = 1) = \frac{(2 * N * 2) * (1 * 1 * 1)}{(2 * N * 2)^2} = \frac{1}{4N}.$$

If the two chromatids belong to a single individual, then the coalescent time is

$$P_{\text{same}}(t = 1) = \frac{(2 * N * 2) * 0}{(2 * N * 2) * (1 * N * 2)} = 0.$$

Before proceeding to $t = 2$, first compute $P_{\text{diff}\to\text{same}}$, being the probability that two chromatids are traced back to the same individual though not the same chromatid (so not coalescing). This should be

$$P_{\text{diff}\to\text{same}} = \frac{(2 * N * 2) * (1 * 1 * 1)}{(2 * N * 2)^2} = \frac{1}{4N}.$$

Similarly,

$$P_{\text{diff}\to\text{diff}} = \frac{(2*N*2)*(1*(N-1)*2+1*N*2)}{(2*N*2)^2} = \frac{4N-2}{4N}$$
$$P_{\text{same}\to\text{diff}} = \frac{(2*N*2)*(1*N*2)}{(2*N*2)*(1*N*2)} = 1$$
$$P_{\text{same}\to\text{same}} = \frac{(2*N*2)*0}{(2*N*2)*(1*N*2)} = 0.$$

Certainly, if the two chromatids are from different individuals, then they either coalesce, trace back to different individuals, or trace back to the same individual. The same is true when they are from the same individual. And indeed,

$$P_{\text{same}\to\text{same}} + P_{\text{same}\to\text{diff}} + P_{\text{same}}(t=1) = 0+1+0 = 0$$
$$P_{\text{diff}\to\text{same}} + P_{\text{diff}\to\text{diff}} + P_{\text{diff}}(t=1) = \frac{1}{4N} + \frac{4N-2}{4N} + \frac{1}{4N} = 0.$$

Now for the different individuals case, the probability that the coalescent time is 2 generations is

$$\begin{aligned} P_{\text{diff}}(t=2) &= P_{\text{diff}\to\text{same}}P_{\text{same}}(t=1) + P_{\text{diff}\to\text{diff}}P_{\text{diff}}(t=1) \\ &= \frac{1}{4N}*0 + \frac{4N-2}{4N}*\frac{1}{4N} = \frac{4N-2}{16N^2} \end{aligned}$$

and for the same individual case is

$$\begin{aligned} P_{\text{same}}(t=2) &= P_{\text{same}\to\text{same}}P_{\text{same}}(t=1) + P_{\text{same}\to\text{diff}}P_{\text{diff}}(t=1) \\ &= 0*0 + 1*\frac{1}{4N} = \frac{1}{4N}. \end{aligned}$$

Repeating the steps above gives a distribution of coalescent times for each case:

$$P_{\text{same}}(t=1) = 0$$
$$P_{\text{diff}}(t=1) = \frac{1}{4N}$$
$$P_{\text{same}}(t=i+1) = P_{\text{diff}}(t=i)$$
$$P_{\text{diff}}(t=i+1) = \frac{4N-2}{4N}P_{\text{diff}}(t=i)$$

2 Adding recombination

As a simple fix to add in recombination, let each parent when mating has a probability π to have recombination.

Suppose a recombination occurs when the n-th generation is generated. Then for the offspring which carries the chromosome affected by recombination, it has two genetic ancestry trees, one for each part of the recombined chromosome. This can be seen from the figure below:

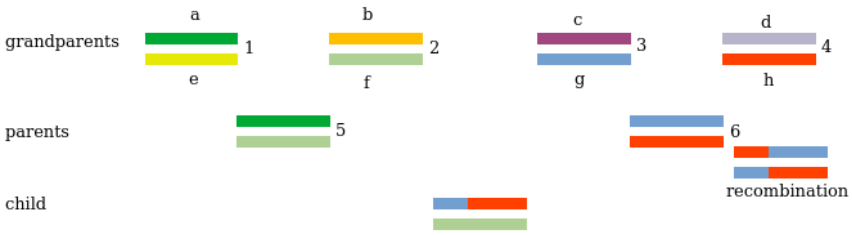

Figure 1: Each letter here represents a line of previous carriers, and each number represents an individual. Individual 5 has two carrier graphs, one being a-append-1, and the other being f-append-2. Individual 6 also has two carrier graphs, one being g-append-3, and the other being h-append-4. It should then be apparent that due to recombination, the child has three carrier graphs, one being (f-append-2)-append-5, one being (g-append-3)-append-6, and one being (h-append-4)-append-6.