The Coalescent

Introduction

I've mentioned many times by now that population geneticists often look at the world backwards. Sometimes when they do, the result is very useful. Consider genetic drift, for example. So far we've been trying to predict what will happen in a population given a particular effective population size. But when we collect data we are often more interested in understanding the processes that produced the pattern we find than in predicting what will happen in the future. So let's take a backward look at drift and see what we find.

Reconstructing the genealogy of a sample of alleles

Specifically, let's keep track of the genealogy of alleles. In a finite population, two randomly chosen alleles will be identical by descent with respect to the immediately preceding generation with probability $1/2N_e$. That means that there's a chance that two alleles in generation t are copies of the same allele in generation $t-1$. If the population size is constant, meaning that the number of alleles in the population is remaining constant, that also means that there's a chance that some alleles present in generation $t-1$ will not have descendants in generation t. Looking backward, then, the number of alleles in generation $t-1$ that have descendants in generation t is always less than or equal to the number of alleles in generation t. That means if we trace the ancestry of alleles in a sample back far enough, all of them will be descended from a single common ancestor. Figure 1 provides a simple schematic illustrating how this might happen.

Now take a look at Figure 1. Time runs from the top of the figure to the bottom, i.e., the current generation is represented by the circles in the botton row of the figure. Each circle represents an allele. The eighteen alleles in our current sample are descended from only four alleles that were present in the populations ten generations ago. The other fourteen alleles present in the population ten generations ago left no descendants. How far back in time we'd have to go before all alleles are descended from a single common ancestor depends on the effective size of the population, because how frequently two (or more) alleles are descended from the same allele in the preceding generation depends on the effective size of

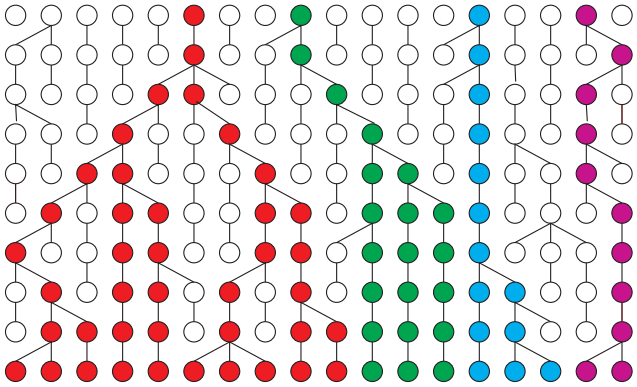

Figure 1: A schematic depiction of one possible realization of the coalescent process in a population with 18 haploid gametes. There are four coalescent events in the generation immediately preceding the last one illustrated, one involving three alleles.

the population, too. But in any finite population the pattern will look something like the one I've illustrated here.

Mathematics of the coalescent: two alleles

Thirty-five years ago J. F. C. Kingman developed a convenient and powerful way to describe how the time to common ancestry is related to effective population size [2, 3]. The process he describes is referred to as the *coalescent*, because it is based on describing the probability of *coalescent events*, i.e., those points in the genealogy of a sample of alleles where two alleles are descended from the same allele in the immediately preceding generation.[1] Let's consider a simple case, one that we've already seen, first, i.e., two alleles drawn at random from a single population.

The probability that two alleles drawn at random from a population are copies of the same allele in the preceding generation is also the probability that two alleles drawn at random from that population are identical by descent with respect to the immediately preceding

[1] An important assumption of the coalescent is that populations are large enough that we can ignore the possibility that there is more than one coalescent event in a single generation. That also means that we also only allow coalescence between a pair of alleles, not three or more. In both ways the mathematical model of the process differs from the diagram in Figure 1.

generation. We know what that probability is,[2] namely

$$\frac{1}{2N_e^{(f)}} \quad .$$

I'll just use N_e from here on out, but keep in mind that the appropriate population size for use with the coalescent is the inbreeding effective size. Of course, this means that the probability that two alleles drawn at random from a population are *not* copies of the same allele in the preceding generation is

$$1 - \frac{1}{2N_e} \quad .$$

We'd like to calculate the probability that a coalescent event happened at a particular time t, in order to figure out how far back in the ancestry of these two alleles we have to go before they have a common ancestor. How do we do that?

Well, in order for a coalescent event to occur at time t, the two alleles must have *not* have coalesced in the generations preceding that.[3] The probability that they did not coalesce in the first $t-1$ generations is simply

$$\left(1 - \frac{1}{2N_e}\right)^{t-1} \quad .$$

Then after having remained distinct for $t-1$ generations, they have to coalesce in generation t, which they do with probability $1/2N_e$. So the probability that two alleles chosen at random coalesced t generations ago is

$$P(T = t) = \left(1 - \frac{1}{2N_e}\right)^{t-1} \left(\frac{1}{2N_e}\right) \quad . \tag{1}$$

It's not too hard to show, once we know the probability distribution in equation (1), that the average time to coalescence for two randomly chosen alleles is $2N_e$.[4]

Mathematics of the coalescent: multiple alleles

It's quite easy to extend this approach to multiple alleles.[5] We're interested in seeing how far back in time we have to go before all alleles are descended from a single common ancestor.

[2] Though you may not remember it.

[3] Remember that we're counting generations backward in time, so when I say that a coalescent event occurred at time t I mean that it occurred t generations ago.

[4] If you've had a little bit of probability theory, you'll notice that equation 1 shows that the coalescence time is a geometric random variable.

[5] Okay, okay. What I should really have said is "It's not *too* hard to extend this approach to multiple alleles, if you are comfortable with probability thinking." Rembmer: I don't expect you to be able to derive

We'll assume that we have m alleles in our sample. The first thing we have to calculate is the probability that any two of the alleles in our sample are identical by descent from the immediately preceding generation. To make the calculation easier, we assume that the effective size of the population is large enough that the probability of two coalescent events in a single generation is vanishingly small. We already know that the probability of a coalescence in the immediately preceding generation for two randomly chosen alleles is $1/2N_e$. But there are $m(m-1)/2$ different pairs of alleles in our sample.[6] So the probability that one pair of these alleles is involved in a coalescent event in the immediately preceding generation is

$$\left(\frac{1}{2N_e}\right)\left(\frac{m(m-1)}{2}\right) \quad .$$

From this it follows[7] that the probability that the first coalescent event involving this sample of alleles occurred t generations ago is

$$P(T=t) = \left(1-\left(\frac{1}{2N_e}\right)\left(\frac{m(m-1)}{2}\right)\right)^{t-1}\left(\frac{1}{2N_e}\right)\left(\frac{m(m-1)}{2}\right) \quad . \tag{2}$$

So the mean time back to the first coalescent event is

$$\frac{2N_e}{m(m-1)/2} = \frac{4N_e}{m(m-1)} \text{ generations} \quad .$$

But this is, of course, only the first coalescent event. We were interested in how long we have to wait until *all* alleles are descended from a single common ancestor. Now this is where Kingman's sneaky trick comes in. After the first coalescent event, we have $m-1$ alleles in our sample, instead of m. So the whole process starts over again with $m-1$ alleles instead of m.[8] Since the time to the first coalescence depends only on the number of alleles in the sample and not on how long the first coalescence event took, we can calculate the average time until all coalescences have happened as

$$\bar{t} \;=\; \sum_{k=2}^{m} \bar{t}_k$$

these results on your own. Don't worry if you can't see how you could have come up with the mathematics that follow, don't worry about it. Unless you want to make contributions to developing new theory in population genetics, you don't need to do derivations like these.

[6]Where did I get that $m(m-1)/2$? You can either take my word for it as "a well known fact," or you can ask me about it, and I'll show you where it comes from.

[7]Using logic just like what we used in the two allele case.

[8]For anyone who cares, this is another example of the Markov property of genetic drift.

$$
\begin{aligned}
&= \sum_{k=2}^{m} \frac{4N_e}{k(k-1)} \\
&\quad \text{TAMO} \\
&= 4N_e \left(1 - \frac{1}{m}\right) \\
&\approx 4N_e
\end{aligned}
$$

An example: Mitochondrial Eve

Cann et al. [1] sampled mitochondrial DNA from 147 humans of diverse racial and geographic origins. Based on the amount of sequence divergence they found among genomes in their sample and independent estimates of the rate of sequence evolution, they inferred that the mitochondria in their sample had their most recent common ancestor about 200,000 years ago. Because all of the most ancient lineages in their sample were from individuals of African ancestry, they also suggested that mitochondrial Eve lived in Africa. They used these arguments as evidence for the "Out of Africa" hypothesis for modern human origins, i.e., the hypothesis that anatomically modern humans arose in Africa about 200,000 years ago and displaced other members of the genus *Homo* in Europe and Asia as they spread. What does the coalescent tell us about their conclusion?

Well, we expect all mitochondrial genomes in the sample to have had a common ancestor about $2N_e$ generations ago. Why $2N_e$ rather than $4N_e$? Because mitochondrial genomes are haploid. Furthermore, since we all got our mitochondria from our mothers, N_e in this case refers to the effective number of *females*.

Given that a human generation is about 20 years, a coalescence time of 200,000 years implies that the mitochondrial genomes in the Cann et al. sample have their most recent common ancestor about 10,000 generations ago. If the effective number of females in the human populations is 5000, that's exactly what we'd expect. While 5000 may sound awfully small, given that there are more than 3 billion women on the planet now, remember that until the recent historical past (no more than 500 generations ago) the human population was small and humans lived in small hunter-gatherer groups, so an effective number of females of 5000 and a total effective size of 10,000 may not be unreasonable. If that's true, then the geographical location of mitochondrial Eve need not tell us anything about the origin of modern human populations, because there had to be a coalescence somewhere. There's no guarantee, from this evidence alone, that the Y-chromosome Adam would have lived in Africa, too. Having said that, my limited reading of the literature suggests that other dara are consistent with the "Out of Africa" scenario. Y-chromosome polymorphisms, for example, are also consistent with the "Out of Africa" hypothesis [5]. Interestingly, dating of those polymorphisms suggetsts that Y-chromosome Adam left Africa 35,000 – 89,000 years

ago.

The coalescent and F-statistics

Suppose we have a sample of alleles from a structured population. For alleles chosen randomly within populations let the average time to coalescence be $\bar{t}_0$. For alleles chosen randomly from different populations let the average time to coalescence be $\bar{t}_1$. If there are k populations in our sample, the average time to coalescence for two alleles drawn at random without respect to population is[9]

$$\bar{t} = \frac{k(k-1)\bar{t}_1 + k\bar{t}_0}{k^2} \quad .$$

Slatkin [4] pointed out that F_{st} bears a simple relationship to average coalescence times within and among populations. Given these definitions of $\bar{t}$ and $\bar{t}_0$,

$$F_{st} = \frac{\bar{t} - \bar{t}_0}{\bar{t}} \quad .$$

So another way to think about F_{st} is as a measure of the proportional increase in coalescence time that is due to populations being separate from one another. One way to think about that relationship is this: the longer it has been, on average, since alleles in different populations diverged from a common ancestor, the greater the chances that they have become different. An implication of this relationship is that F-statistics, by themselves, can tell us something about how recently populations have been connected, relative to the within-population coalescence time, but they can't distinguish between recent common ancestry that is due to lots of migration among populations and recent common ancestry that is due to a recent split between populations.

A given pattern of among-population relationships might reflect a migration-drift equilibrium, a sequence of population splits followed by genetic isolation, or any combination of the two. If we are willing to assume that populations in our sample have been exchanging genes long enough to reach stationarity in the drift-migration process, then F_{st} may tell us something about migration. If we are willing to assume that there's been no gene exchange among our populations, we can infer something about how recently they've diverged from one another. But unless we're willing to make one of those assumptions, we can't really say anything.

[9] If you don't see why, don't worry about it. You can ask if you really care. We only care about $\bar{t}$ for what follows anyway.

References

[1] R L Cann, M Stoneking, and A C Wilson. Mitochondrial DNA and human evolution. *Nature*, 325:31–36, 1987.

[2] J F C Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19A:27–43, 1982.

[3] J F C Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

[4] Montgomery Slatkin. Inbreeding coefficients and coalescence times. *Genetical Research*, 58:167–175, 1991.

[5] Peter A Underhill, Peidong Shen, Alice A Lin, Li Jin, Giuseppe Passarino, Wei H Yang, Erin Kauffman, Batsheva Bonne-Tamir, Jaume Bertranpetit, Paolo Francalacci, Muntaser Ibrahim, Trefor Jenkins, Judith R Kidd, S Qasim Mehdi, Mark T Seielstad, R Spencer Wells, Alberto Piazza, Ronald W Davis, Marcus W Feldman, L Luca Cavalli-Sforza, and Peter J Oefner. Y chromosome sequence variation and the history of human populations. *Nature Genetics*, 26(3):358–361, 2000.

Creative Commons License