3 Random genetic drift

So far in describing evolutionary dynamics we have been using a deterministic framework allowing for no stochastic fluctuations of any kind. Using this framework, however, requires some additional assumptions. In particular, our description of different populations in terms of allele frequencies implies that the population size is very large. In this and the next two sections, we will consider how allowing for stochastic effects resulting from the finiteness of biological populations affects evolutionary dynamics.

An important notion in evolutionary population genetics is that of random genetic drift by which one usually means random variation in genotype frequencies resulting from the finiteness of populations. There are two important effects of random genetic drift that we will consider. Genetic drift (i) removes genetic variation from populations and (ii) affects the probability of survival of new mutations.

We consider a finite population of size $N$ assuming discreet and non-overlapping generations. There is a single locus with $k$ alleles which are identical with respect to viability and fertility (i.e. there is no selection). We assume that each adult produces a very large number of offspring of which only $N$ (randomly chosen from a common offspring pool) organisms will develop into the adults of the next generation. Let us consider an allele $A$ that has frequency $p$ (so that on average there are $Np$ alleles $A$ out of $N$ alleles present in the population). Our assumptions about the way the next generation is formed translates into assuming that the probability that $i$ alleles $A$ made it to the next generation is given by the binomial probability

$$\text{Prob}\{\text{there are } i \text{ alleles } A\} = \frac{N!}{i!(N-i)!} p^i (1-p)^{N-i},$$

where $i$ can be 0, 1, ..., $N$. This is a so called Fisher-Wright binomial scheme.

Of course, changes in allele frequencies in any individual population resulting from random genetic drift cannot be predicted. But the average behavior in a large number of populations can be predicted. Let $p$ be the frequency of allele $A$ in this generation. We consider the expected frequency, $E\{p'\}$, of this allele in the next generation. The mean of a binomial variable with parameters $N$ and $p$ is $Np$ so that

$$E\{p'\} = E\{i/N\} = p.$$  

This shows that the expected change in allele frequency is zero, $E\{\Delta p\} = 0$. The variance of a binomial variable with parameters $N$ and $p$ is $Np(1-p)$ so that the variance in $p'$,

$$\text{var}\{p'\} = \text{var}\{i/N\} = \frac{pq}{N},$$

where $q = 1 - p$. This shows that the variance of the allele frequency change is $\text{var}\{\Delta p\} = \frac{pq}{N}$ and is positive. Thus, the allele frequencies are expected to change.

[Random genetic drift: fluctuations in the genetic structure of a population (which is characterized by allele/genotype frequencies) resulting from inherent stochasticity in the process of reproduction (number of offspring, segregation, recombination, mutation etc) and the finiteness of the population size.

One locus multiallele haploid system: if each individual leaves the same number of offspring, no drift even if the population size is finite.]
3.1 Identity in state

We consider a one-locus haploid population. Let $f_t$ be the probability that two randomly chosen distinct individuals in generation $t$ are identical in state (that is carry alleles of the same type). Note that $h_t = 1 - f_t$ can be interpreted as a measure of genetic variability of the population. (With only two alleles with frequencies $p$ and $q$, $f_t = p_t^2 + q_t^2$, $h_t = 2p_tq_t$).

Let us consider the next generation. Two randomly chosen distinct individuals are identical in state if (i) they have the same “parent” (which happens with probability $1/N$) or (ii) if they have different parents (which happens with probability $1 - 1/N$) but these parents already were identical in state (which happens with probability $f_t$). Thus, the dynamics of $f_t$ can be described by a single recurrence equations

$$f_{t+1} = \frac{1}{N} + (1 - \frac{1}{N}) f_t.$$

The corresponding equation for $h_t$ is even simpler:

$$h_{t+1} = 1 - f_{t+1} = (1 - \frac{1}{N}) h_t.$$

The solution of this equation is

$$h_t = h_0(1 - \frac{1}{N})^t \approx h_0 e^{-t/N}$$

where the approximation is good if $N$ is sufficiently large. Thus, as $t \to \infty$, $h_t \to 0$ and only one allele remains in the population. The characteristic time for approaching a monomorphic state is approximately $N$ generations.

Note that $h \to 0$ independently of how many alleles are present initially. Even if initially all $N$ individuals have different alleles ($k = N$), in $N$ generations $h \approx 0$. Another way to put it is to say that all alleles present right now can be traced to a single ancestor allele $\approx N$ generations ago! This also means that with no differences in fitness, the probability of fixation of a specific allele is $1/N$.

3.1.1 Major points

- random genetic drift removes genetic variability;
- the characteristic time of fixation is order $N$ generations.

---

Homework:

What is the probability that a particular allele gets at least a copy in the next generation? Hint: consider the probability of the complementary event (= “no copies of this allele in the next generation”).
3.2 Drift and mutation

With mutation the permanent loss of genetic variation is, obviously, impossible. Here we consider how including mutation affects the dynamics of genetic variation. We assume that there are $k$ alleles $A_i$ ($i = 1, \ldots, k$) and that each allele can mutate (after reproduction) to any other $k-1$ alleles with probability $\mu/(k-1)$:

$A_i \overset{\mu}{\rightarrow} A_j$.

Note that the overall probability that an allele mutates is $\mu$.

The probability of identity in state after reproduction but before mutation satisfies to an expression derived in the previous section:

$$f_{t+1} = \frac{1}{N} + (1 - \frac{1}{N})f_t.$$  \hfill (42a)

The probability of identity in state in the next generation (after reproduction and mutation) is

$$f_{t+1} = f_{t+1}' [\mu^2 + (1 - \mu)^2] + (1 - f_{t+1}') \frac{\mu}{k-1} (1 - \mu).$$  \hfill (42b)

Here the first term in the right-hand side specifies the probability that two randomly chosen allele were identical before mutation and did not mutate or did mutate but to the same allele. The second term specifies the probability that the allele that were different in state after reproduction become identical after mutation in one of them.

Equations (42) can be used to show that as time increases, $f_t$ approaches an equilibrium value which can be approximated (assuming that $\mu << 1$) as

$$f = \frac{2N\mu}{2N\mu + 1} + 1.$$  \hfill (43a)

This simplifies to

$$f = \frac{2N\mu + 1}{4N\mu + 1}$$  \hfill (43b)

for $k = 2$, and to

$$f = \frac{1}{2N\mu + 1}$$  \hfill (43c)

for $k \to \infty$. Note that $N\mu$ is the number of mutants per generation. With $N\mu = 1$, $f = 3/5$ for $k = 2$ and $f = 1/3$ for large $k$. Thus, a small number of mutants per generation is sufficient to maintain large levels of genetic variability in a finite population (with no selection). If $N\mu << 1$ (which is the case if the population size is very small), then $f \approx 1$ (that is the level of variability is very small).

3.2.1 Major points

- mutation maintains genetic variation in finite populations;
- an important parameter controlling the level of genetic variation maintained is the number of mutants per generation, $N\mu$. 

49
3.3 Coalescence

- Two genes are *identical by descent* if they descend from the same gene in some ancestral population.

Consider a randomly chosen pair of alleles from a diploid population of size $N$. The probability of coalescence in the previous generation is

$$P^{(2)}_c = \frac{1}{2N}.$$  

The probability of no coalescence in the previous generation is

$$P^{(2)}_{nc} = 1 - \frac{1}{2N}.$$  

The probability of coalescence in generation $t + 1$ is

$$P^{(2)}_c (t + 1) = \left(1 - \frac{1}{2N}\right)^t \frac{1}{2N} \approx \frac{1}{2N} \exp\left(-\frac{t}{2N}\right).$$

This is an exponential distribution. Let $T$ the the time until coalescence. Then

$$E(T^{(2)}) = 2N, \quad var(T^{(2)}) = 4N^2. \quad (44)$$

Next, consider $k$ randomly chosen alleles. The probability that they all came from different copies is

$$P^{(k)}_{nc} = \prod_{i=1}^{k-1} \left(1 - \frac{i}{2N}\right) \approx 1 - \binom{k}{2} \frac{1}{2N},$$

where the approximation assumes that $k(k - 1) \ll N$. The probability of coalescence in the previous generation is

$$P^{(k)}_c = \frac{k}{2N} = \frac{k(k-1)}{4N}.$$  

The mean waiting time until the next coalescence is the inverse of $P^{(k)}_c$:

$$E(\tau) = \frac{4N}{k(k-1)}.$$  

The distribution of $\tau$ is approximately exponential.

The expected total time until all $k$ alleles to coalesce is

$$E(T^{(k)}) = \frac{4N}{k(k-1)} + \frac{4N}{(k-1)(k-2)} + \frac{4N}{(k-2)(k-3)} + \cdots + \frac{4N}{2} = 4N \left(1 - \frac{1}{k}\right). \quad (45)$$

Comparing eq. (44) with eq. (45) one can see that about a half of the depth of the coalescent tree is spent for the last two copies to coalesce. So lineages coalesce rapidly at first and then the process gradually slows down.