Short-term evolution: population genetics

Erol Akçay

Biol 417 Spring 2019

March 11, 2019

In this lecture, we will look at mathematical descriptions of short-term evolution, i.e., the competition between two genetic variants in the same population. The next lecture will that the long-term evolutionary perspective where the question is what happens over many such instances of short-term competition.

Natural selection

Haploid case

Haploid means that each individual possesses one copy of a given gene; diploid organisms (like humans) have two copies of each gene. Suppose you have two alleles, denoted by A and B, and the number of individuals carrying them is given by n_A and n_B. In population genetics, we are mostly interested in the frequency of different alleles (i.e., variants of the same gene). The frequency of the A allele, p is:

$$p = \frac{n_A}{n_A + n_B} ;$$

the frequency of B, q, is simply $1 - p$, or $\frac{n_B}{n_A+n_B}$. Suppose now that the allele somehow affects the per-capita reproduction of individuals, such that carriers of the A allele produce w_A offspring, and B allele w_B in one time step. Then, we have for the numbers after one time step:

$$n_A(t+1) = w_A n_A(t) \qquad n_B(t+1) = w_B n_B(t) .$$

Consequently, the new frequencies of A and B are:

$$\begin{aligned} p(t+1) &= \frac{n_A(t+1)}{n_A(t+1) + n_B(t+1)} = \frac{w_A n_A(t)}{w_A n_A(t) + w_B n_B(t)} \\ &= \frac{w_A}{w_A \frac{n_A(t)}{n_A(t)+n_B(t)} + w_B \frac{n_B(t)}{n_A(t)+n_B(t)}} \frac{n_A}{n_A + n_B} \\ &= \frac{w_A}{p(t)w_A + q(t)w_B} p(t) = \frac{w_A}{\overline{w}} p(t) , \end{aligned} \tag{1}$$

where $\overline{w} = pw_A + qw_B$ is the mean fitness. Likewise, the new frequency of the B allele, $q(t+1) = 1 - p(t+1) = \frac{w_B}{\overline{w}} q(t)$. Another useful way to write the

same information is to look at the change in p:

$$\begin{aligned}\Delta p = p(t+1) - p(t) &= \frac{w_A}{\overline{w}}p(t) - p(t) = \frac{w_A - \overline{w}}{\overline{w}}p(t) \\ &= \frac{w_A - p(t)w_A - (1-p(t))w_B}{\overline{w}}p(t) = \frac{w_A - w_B}{\overline{w}}p(t)(1-p(t)) \\ &= \frac{w_A - w_B}{\overline{w}}\mathrm{var}(p)\ , \end{aligned} \tag{2}$$

where in the last step, the variance refers to the allele frequency within individuals (which in the haploid case is 0 or 1). Notice that the variance is always non-negative (and positive whenever $0 < p < 1$); and similarly $\overline{w} > 0$. So, the sign of change in the frequency of A allele is completely determined by the sign of $w_A - w_B$: if $w_A > w_B$, A increases ($\Delta p > 0$); it decresases if $w_A < w_B$. A corollary of this is that when w_A and w_B are fixed, a mutant allele that can increase when rare, will also go to fixation, i.e., reach frequency 1.

We can also look at the change in mean fitness after selection:

$$\Delta\overline{w} = \overline{w}(t+1) - \overline{w}(t) = \Delta p(w_A - w_B) = \frac{(w_A - w_B)^2}{\overline{w}}\mathrm{var}(p)\ . \tag{3}$$

The last expression looks innocuous enough, but it's a special case of a result that caused a lot of ink to be spilled in the history of population genetics. To see why, note that the numerator, being a square, is non-negative, and so is $var(q)$, and since $\overline{W} > 0$, it follows that the change in mean fitness in the population has to be non-negative (and strictly positive whenever alleles have different fitnesses). This means that selection will always increase the mean fitness of the population, which seems to encapsulate the intuition behind Darwinism nicely.

But remember, we derived this result assuming w_A and w_B to be constant; in particular, they don't depend on the allele frequencies in the population. When this assumption is violated, it turns out one can write a similar result to (3), but its interpretation is much more subtle. We will talk about this more later when we do the Price equation.

Diploid case

Now, each individual carries two copies of a gene, so we have three types of individuals AA, AB, and BB. Almost all diploid organisms go through a lifecycle that takes them through a haploid stage, the gametes. So, there is a choice to be made about when to censusing the population. When the gametes unite at random (e.g., as in many broadcast spawners that simply release their gametes into the ocean), then knowing the frequencies of gametes carrying different alleles allows us to write the diploid genotype frequencies from the allele frequencies. This means that if we census the population at the gametic stage, we only have to keep track of one variable (the frequency of one of the alleles), as opposed to two. If the frequency of A at the gametic stage at generation t is p again, the adult frequencies f_{XY} are given by: $f_{AA} = p^2$, $f_{AB} = 2p(1-p)$, $f_{BB} = (1-p)^2$. Now, suppose that the number of gametes produced by each genotype is given by w_{AA}, w_{AB}, and w_{BA} , and fair meiosis, such that each copy of the allele has equal chance to go into each gamete produced, the AA

To cut down on notation, I will dispense with the (t) argument and just write p, using a prime to denote the value of the same variable in the next time step, as in p'.

Variation in the ws can represent either differential reproduction, or differential survival, or a combination of both.

(BB) individual produces w_{AA} (w_{BB}) A (B) gametes, and the AB individual produces $w_{AB}/2$ each of A and B gametes. Thus, the frequency of A gametes after the adults have reproduced is given by:

$$p' = \frac{w_{AA}p^2 + w_{AB}p(1-p)}{w_{AA}p^2 + 2w_{AB}p(1-p) + w_{BB}(1-p)^2} = \frac{1}{\overline{w}}(w_{AA}p^2 + w_{AB}p(1-p)) \tag{4}$$

Similar to the haploid case, we can write Δp:

$$\begin{aligned}\Delta p = p' - p &= \frac{1}{\overline{w}}(w_{AA}p^2 + w_{AB}p(1-p)) - p \\ &= \frac{1}{\overline{w}}(w_{AA}p^2 + w_{AB}p(1-p) - w_{AA}p^3 - 2w_{AB}p^2(1-p) - w_{BB}p(1-p)^2) \\ &= \frac{1}{\overline{w}}p(1-p)(pw_{AA} + w_{AB}(1-2p) - (1-p)w_{BB}) \\ &= \frac{1}{\overline{w}}p(1-p)\left[p(w_{AA} - w_{AB}) + (1-p)(w_{AB} - w_{BB})\right] \ . \end{aligned} \tag{5}$$

To see what equation (5) entails, consider the following table of fitness values:

	AA	AB	BB
w_{XY}	1	$1-hs$	$1-s$

In this model, with $s > 0$, the B allele is a deleterious mutation, and h is the dominance coefficient. If $h = 0$, then genotype AB has the same fitness as AA, so A is dominant; $h = 1$ means B is dominant. If $1 > h > 0$, we have incomplete dominance, and $h > 1$ and $0 > h$ are termed under- and over-dominance, respectively.

Now, we can explore what the dynamics of an allele A look like under different scenarios for h. Plugging the fitness values into equation (5), we get:

$$\Delta p = \frac{1}{\overline{w}}p(1-p)s\left[ph + (1-p)(1-h)\right] \tag{6}$$

Thus, the sign of Δp is the same as the term in the square brackets. That term is guaranteed to be positive when $0 < h < 1$ (incomplete dominance), so with incomplete dominance, we can be sure that allele A will always increase.

What happens with over-dominance ($h < 0$)? To see what happens we can take the derivative of the square brackets with respect to p, yielding $2h - 1$. If $h < 0$, then this derivative is negative. Furthermore, we can solve for the value of p that makes the square brackets vanish:

$$p^* = \frac{1-h}{1-2h} \ , \tag{7}$$

which is between 0 and 1 for $h < 0$. That means that for $p < p^*$, the sign of the square bracket is positive, and for $p > p^*$ it's positive. Thus, $\Delta p > 0$ for $p < p^*$ and $\Delta p < 0$ for $p > p^*$. This is called balancing selection, where for low values of p, selection acts to increase its value, and for high values it decreases. It is

one way in which a polymorphism (i.e., multiple alleles of the same locus) can be maintained in a population.

With underdominance $h > 1$, the converse happens: i.e., at low values of p, $\Delta p < 0$, and at high values $\Delta p > 0$, yielding two equilibria, where one or the allele is fixed in the population.

This will be assigned as an exercise.

Now, the term in the square brackets in equation (5) seems to play a big role in determining the type of dynamics but it's not immediately obvious what that term means biologically. Sewall Wright, in 1920s was bugged about this, and observed a tantalizing fact:

$$\begin{aligned}\frac{d\overline{w}}{dp} &= 2w_{AA}p + 2w_{AB}(1-p) - 2w_{AB}p - 2w_{BB}(1-p) \\ &= 2(p(w_{AA} - w_{AB}) + (1-p)(w_{AB} - w_{BB}))\,, \end{aligned} \tag{8}$$

in other words, we can write the change in p as:

$$\Delta p = \frac{p(1-p)}{2\overline{w}}\frac{d\overline{w}}{dt}\,. \tag{9}$$

This equation says that the allele frequency will always change in the direction that increases the mean fitness of the population. This result prompted Wright to formulate his "fitness landscape" idea, where he visualized populations climbing hills defined by the mean fitness. This again seemed to justify the Darwinian intuition that natural selection leads to gradual "improvement" of mean fitness of populations, making them more adapted to their environment. However, as we will see, this conclusion does not hold in general. Also worth noting is that this result is obtained in a "short-term" context: it's about the change in allele frequency in a population with a fixed set of alleles. It does not (yet) say anything about long-term evolution, where many different alleles can arise through mutation, be maintained, fixed, or lost over time.

Mutation and Drift

All of the above assumed that dynamics of alleles are completely deterministic: individuals' reproduction is exactly the value determined by their genotypes, and there is no new variants entering into the population through mutation. These are good approximations in large populations (where average fitnesses matter) and when mutation rate is low. But in small populations, or with high mutation rate (which can happen, for example, if the mutational "target" is large), one has to consider stochasticity. The type of questions one asks about stochastic dynamics is again different than deterministic dynamics above.

Consider a population of size N. Ignore mutation and selection for the moment so that everyone is equally likely to reproduce. One of the facts of a finite population is that eventually genetic variation will disappear. To prove this, consider the probability that two randomly sampled alleles in a population

(without replacement) are identical by state (i.e., they are both A or B). Denote the current value of this probability by $\mathcal{G}$, and consider what will happen to it after one generation (during which all that happens is reproduction, and random pairing of gametes)

$$\mathcal{G}' = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\mathcal{G}\,, \tag{10}$$

The first term gives the probability that the two randomly sampled alleles in the descendant generation are copies of the same allele in the parent generation. In that case, the two sampled alleles are identical by state for sure. If the parents are not the same (which happens with probability 1-1/2N), then the probability that the alleles are identical by state is simply $\mathcal{G}$, the same probability in the previous generation. What this equation shows is that $\Delta\mathcal{G}$ is always positive, so $\mathcal{G}$ will increase until it reaches 1, at which point all the alleles will be of the same time (but this equation doesn't say what type it is). Sometimes it's easier to work with $\mathcal{H} = 1 - \mathcal{G}$, the probability that two alleles are not identical by state, for which we have:

$$\Delta\mathcal{H} = -\Delta\mathcal{G} = -\frac{1}{2N}(1 - \mathcal{G}) = -\frac{1}{2N}\mathcal{H}\,, \tag{11}$$

so $\mathcal{H}$ will decay to zero in the long term. The decay rate, however, is inversely proportional to N, showing that drift as a destroyer of variation is much less effective in large populations.

Of course, mutation is one way you can rescue variation in the population. Suppose mutations occur with probability u per reproduction event. Then, for $\mathcal{G}'$, we have:

$$\mathcal{G}' = (1-u)^2\left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\mathcal{G}\right]\,,$$

which is almost the same equation as before, but now we have to "scale" the probability of identity by state with the fact that it will only happen if neither copy experienced a mutation from their previous states (which is $(1-u)^2$). Typically, mutation probabilities are small ($u \ll 0$), so we can approximate $(1-u)^2$ with $1-2u$, and ignore terms that involve u/N (which with sizable N will be very small) write:

$$\mathcal{G}' \approx \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\mathcal{G} - 2u\mathcal{G}\,. \tag{12}$$

Now $\mathcal{G}$ is not guaranteed to increase, because of the last term. This suggests that ther eis an equilibrium value of $\mathcal{G}$, which we can solve for from equation but setting $\mathcal{G}' = \mathcal{G}$, to obtain:

$$\mathcal{G}^* = \frac{1}{1 + 4Nu}\,. \tag{13}$$

So, the larger N or u is, the lower the probability of two randomly sampled alleles being identical by state, and conversely, the higher the genetic diversity of the population.

One corollary of the neutral drift model is that without mutation, eventually all alleles will be identical by state, which means that one of the alleles will go to fixation. However, looking forward, we don't know which one. Since all alleles are equivalent, it has to be the case that any given copy of an allele will be the ancestor of the entire population in the future with probability $1/2N$. Thus, a new mutation wihich is initially is at frequency $1/2N$ will fix with that probability. Since there are $2N$ individuals, the expected number of mutations that arise in a given generation is $2Nu$, which multiplied with the probability that any given one will fix, gives the expected rate of substitution of new mutations: u.