# Bi 1: The Great Ideas of Biology
# Homework 4
# Due Date: Thursday, May 4, 2023

## 1. Genetic Drift as a Force of Evolution

### Simulating the processes of evolution

In class this week we learned about the mathematical formalism behind population genetics, one of the centerpieces of evolutionary theory. The ideas described in class will provide a quantitative backdrop for understanding the different evolutionary forces that shape life on our planet. It is both profound and amusing how much we can learn about evolution by thinking about coin flips and similar games of chance. Indeed, the broad reach of the mathematics of coin flips is an example of what former Caltech undergrad and now Harvard professor Joe Blitzstein likes to say: "The nouns change, but the verbs remain the same."

In this problem we want you to explore different evolutionary forces by means of simulations. You will use what you learned in the *stochastic simulation* tutorial to explore the interplay of different evolutionary forces such as genetic drift and mutation. By using simulations we will sidestep more advanced mathematics of stochastic differential equations needed to study these concepts analytically while still getting clear insights into how these forces may affect the course of evolution.

### The Buri genetic drift experiment

In 1956 Peter Buri, a student of Sewall Wright published the now classic paper "*Gene Frequency in Small Populations of Mutant Drosophila*" in which he experimentally demonstrated the concept of genetic drift. The idea for this beautiful experiment is depicted in Figure 1. Briefly, Buri began with eight female and eight male flies, all heterozygotes of the *bw* locus. This means that all of the flies had 1 copy of the gene associated with white eyes, and one copy of the gene associated with red eyes. The phenotype that this combination of alleles gives is flies with orange eyes. He then allowed the flies to reproduce, and after removing the adults, he randomly chose 8 males and females from the next generation of offspring without looking at the eye color. These new

8 males and 8 females were transferred to a new flask and the procedure was repeated for 19 generations.

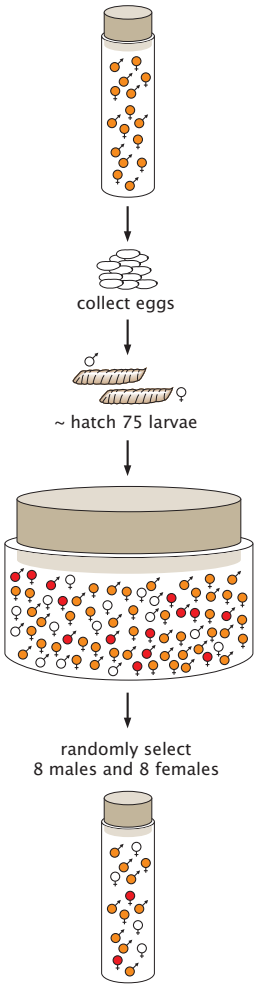


Figure 1: **Buri's experimental setup.** At time $t = 0$ eight heterozygote females and eight heterozygote males were allowed to reproduce. From their offspring, eight males and eight females were chosen at random and transferred into a new flask.

Question 1a

Work out what is the expected genotype frequency of red-eyed flies, white-eyed flies and orange-eyed flies after the first generation. (Hint: Recall that each allele is drawn from the parent's pool **at random with replacement**. This means that to compute the frequency of red-eyed flies you should calculate $f_{rr} = P(\text{red allele first draw}) \cdot P(\text{red allele second draw})$

Since the offspring that made it to the next generation were chosen at random, Buri knew that the outcome would be different if he repeated an identical experiment in different vials. As a result, for statistical power he simultaneously tracked 107 flasks as shown in Figure 1. Each generation, he counted the number of red-eyed, white-eyed and orange-eyed flies he had randomly chosen. Figure 2 shows the outcomes for these different vials after 19 generations. Because the flies are allegedly mating at random, with each generation there is an accumulation of fluctuations. As a result, after 19 generations, many vials contained only white-eyed or red-eyed flies, though some vials still contained a mixture of eye colors.

Having quantified the number of red-eyed, white-eyed and orange-eyed flies Buri was able to quantify the frequency of alleles in the population. Since none of the alleles were dominant, he could infer the genotype by looking at the phenotype of the flies.

Question 1b

Write down the formula for the genotype frequencies in terms of the eye color count. Use the notation $N_{red}$ for the number of red-eyed flies in a given vial, $N_{white}$ for the number of white-eyed flies in that same vial and finally, $N_{orange}$ for the number of orange-eyed flies in that same vial. Your task is to figure out the frequency of red ($f_r$) and white ($f_w$) alleles in a given vial given the counts of the number of red-, white- and orange-eyed flies.

Figure 3 summarizes the results of the experiment. By tracking alleles over time with these 107 populations exposed to the same conditions, Buri was able to observe evolution driven entirely by genetic drift! He saw how in some of the populations one of the alleles went extinct, arising from nothing more than

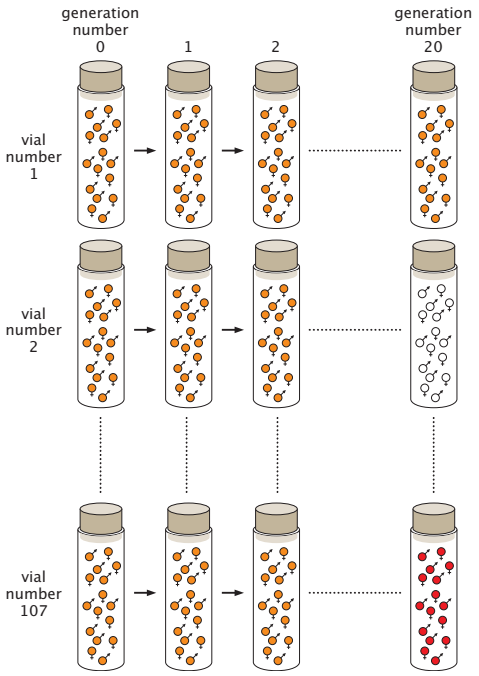


Figure 2: **Multiple replicates of the Buri experiment.** Buri repeated his experiment in 107 separate vials, with the evolutionary trajectory different each time as a result of genetic drift. Note that in the long time limit, many of the vials have gone to fixation with all flies having either white or red eyes.

the fluctuations inherent in small populations.

It is now time for us to use our computational prowess to simulate and explore the Buri experiment.

### Reproducing the Buri experiment *in-silico*

Your first task will be to reproduce Figure **??**figBuriExperiment] by means of stochastic simulations. The key elements of the code you need to do this analysis you already worked out in the stochastic simulation tutorial.

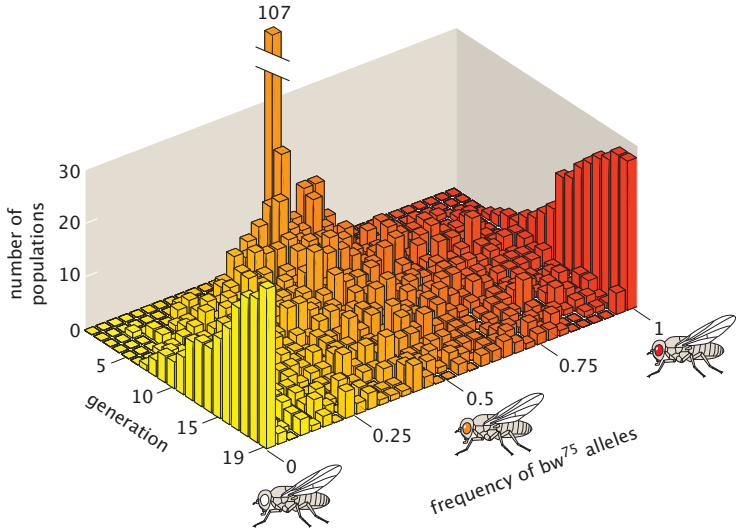


Figure 3: **Results of the Buri experiment.** By tracking the phenotypes of the flies, Buri was able to infer the allele frequencies for each population. The allele frequencies change as a result of genetic drift and after 19 generations, many of the vials contain flies all with the same eye color, implying fixation of alleles and evolution due to genetic drift.

**Question 1c**

Perform stochastic simulations of genetic drift for 107 populations over 19 generations using the same population size as Buri, i.e. 16 flies total (32 alleles). Plot histograms of the allele frequency for generation numbers 0, 1, 10, and 19. (Assemble your code in a Python notebook like that structured in the template linked on the course webpage.)

### The effect of the population size

Using these exact same tools we will now explore the effect of the population size.

Question 1d

Repeat the stochastic simulations for 107 populations during 1000 generations using the same population size as Buri. Quantify the time it takes for each of these populations to have one of the alleles fixed, i.e. find the time point for each population at which the allele frequency becomes either zero or one, and save the generation number at which this happened. Now repeat the simulation for varying population size ($N = 4$, 8, 16, 32, and 64). Plot the mean time to fixation as a function of the population size and comment on the results. (Hint: to find which generation one of the alleles was fixed in the population, the function `numpy.where` might become handy. Basically you just need to find a way for Python to tell you at which entry of the array the frequency became `f == 1` **or** `f == 0`).

### The effect of mutations

Let's now explore the effect of another evolutionary force – mutation. In our toy model, rather than thinking about tracking the complexity of single base pair mutations, we will think of a "reaction" of the following form

$$A \underset{\mu_2}{\overset{\mu_1}{\rightleftharpoons}} a \tag{1}$$

where $A$ and $a$ are the two versions of the allele (for example red and white), and $\mu_1$ and $\mu_2$ are the mutation rates that take you from one allele to the other. To simplify things even further we will assume $\mu_1 = \mu_2 \equiv \mu$.

Question 1e

Implement a stochastic simulation to include the effect of mutation for a single population and plot the allele frequency over time. (Hint: The mating still happens at random in this scenario, but now each allele after being selected for the next generation must flip a second coin to decide if it remains as the same allele, or it mutates into the other allele). Use the value $\mu \approx 0.001$ for your simulations.

Question 1f

Extend the algorithm you just wrote and simulate 100 populations. Plot 10 of these trajectories, as well as histograms of allele frequency at representative time points such as $t = 0, 5, 10, 50, 100, 500$ generations. Compare this to the null model where the mutation rate is equal to zero and comment on the differences if any between the distributions over time.

Question 1g:

You will now explore the effect of the magnitude of the mutation rate. Run the simulation for 100 generations for $\mu$ =0, 0.001, 0.01, 0.1 and plot the histogram of allele frequencies of the final time point for each of these mutation rates.

## 2. Experimental evolution in the era of genome sequencing

Within the past two decades, sequencing an organisms' entire genome has become a nearly trivial procedure. This affords us the ability to observe evolution at the genetic level in real time. This has opened up an exciting new field in which the technology of next-generation sequencing is combined with the experimental advantage of microbial systems making it possible to test quantitative evolutionary theories.

A particularly interesting long-term experiment involves Professor Richard Lenski at Michigan State University. On February 24, 1988, Lenski began growing twelve *E. coli* cultures in parallel, similar to Buri's experiment from problem 1 but in a haploid world with organisms with generation times of around one hour. Twenty-nine years and almost 70,000 generations later this experiment has watched more generations of evolution than any other experiment ever done.

These bacterial cultures have been adapting to a very simple environment with a fixed media composition. The advantage of working with these microbes is that every certain number of generations a sample can be frozen and brought back to life at will. In this sense, Lenski's $-80°$C freezers act as an evolutionary time-machine, allowing him to recover organisms from the "fossil record"!

One surprising outcome of this experiment is the appearance of a bacterial strain capable of metabolizing a new carbon source. For historical reasons (most likely to avoid phage infection) the cultures have always been grown in the presence of citrate. Back in the day, before the sequencing revolution, one of the ways to identify bacterial species was by their metabolic repertoire. Scientists would classify a bacterium as *E. coli* for example based on its ability to ferment arabinose, lactose, mannitol, and the **lack of ability** to ferment citrate, among other things (look at this site for a complete list of the features). So in principle if you were to collect a sample from the soil that was able to ferment citrate you would immediately conclude it was not a wild-type *E. coli* strain.

In fact, *E. coli* does contain the machinery to ferment citrate encoded in its genome, but this set of genes is only expressed under anaerobic conditions. Lenski found that in one of his 12 replicate populations bacteria were able to metabolize citrate under aerobic conditions. This means that once the glucose that is found initially in the media runs out, this mutant strain can still grow further before the culture is diluted the next morning, giving it a clear fitness advantage over its competitors!

In this problem we will work out a very simple equation to analyze how long it would take for this mutant to overtake the culture.

### Toy model for two competing bacteria strains

Consider the case in which two alleles, $A_1$ and $A_2$, are present in a population with initial frequency $p$ and $q = 1 - p$, respectively. For example, these alleles could be those associated with the ability to metabolize citrate or not. Let us assume that cells harboring allele $A_1$ have a growth rate $m_1$ and those harboring $A_2$ have a growth rate $m_2$. When thinking about microbial organisms the growth rate is often taken as a metric for fitness and it's given the name of Malthusian parameter. For natural selection to act on organisms there must be a difference in fitness, otherwise if all organisms had the same fitness, no Darwinian evolution would occur.

Let us further assume that $A_1$ represents the allele that allows bacteria to metabolize citrate, and as a consequence $m_1 > m_2$. In particular we will say that $m_2 = m_1(1 - s)$, where $s$ is a small parameter $s \ll 1$. If $N_1$ represents the number of cells with allele $A_1$, and $N_2$ the number of cells with allele $A_2$, the equation that describes the growth curve is given by

$$\frac{dN_i}{dt} = m_i N_i, \tag{2}$$

for $i \in \{1, 2\}$. The solution to this differential equation results in an exponential growth profile, namely,

$$N_i(t) = N_i(0)e^{m_i t}, \tag{3}$$

where $N_i(0)$ is the initial number of cells with allele $A_i$.

Question 2a:

Write an expression for $N_{tot}(t)$ the total number of cells as a function of time. (Hint: Remember we have two competing cell types and we are assuming they don't interfere with each other).

Having this expression for $N_{tot}(t)$ is interesting. But what we really care about is the frequency of alleles in the population given that one of the alleles has a fitness advantage over the other. This means that the quantity we care about is the **normalized frequency** $p(t)$.

Question 2b:

Write an expression for $p(t)$, the frequency of the mutant allele $A_1$ and another expression for $q(t)$, the frequency of the wild-type allele $A_2$ as a function of time. This should be a function of the initial cell count $N_1(0)$ and $N_2(0)$, as well as the selection coefficient $s$.

Hopefully you ended up with a nice, compact expression that looks like a logistic function. Let's now explore the consequences of this expression.

Question 2c:

Let $p(0)$, the initial frequency of $A_1$, be $10^{-9}$. Assume the doubling time of the mutant is 1 hour, then plot $p(t)$ and $q(t)$ for different selection coefficients, $s = 10^{-4}, 10^{-3}, 10^{-2}$. Label the $x$ axis as years rather than hours to have a better sense of how long it would take for the mutants to overtake the population.

In one of Lenski's experiments, since his fridges contain samples at many time points of this long-term experiment, he was able to go back in time and measure the relative fitness of strains compared to the parental strain. Figure 2 shows some of his results. The blue curve shows how for the first 20,000 generations the rate of mutation accumulation remains pretty constant over time. The green curve shows the relative fitness of further time points compared to the parental strain at the beginning of the experiment. There we can see that at the beginning there was a sharp increment in the relative fitness, to then transition to a less steep rate of fitness increment.

Question 2d:

Provide a qualitative explanation for why these two phases of fitness increment might exist. (Hint: Think of the number of sites in a genome where a mutation might be beneficial as finite.)

Throughout the first 20,000 generations of a specific *E. coli* culture, the mutation rate was estimated to be approximately $1.6 \times 10^{-10}$ per bp per generation. However, between generations 25,000 and 40,000, eighty-three new synonymous mutations were detected. Additionally, most of these mutations

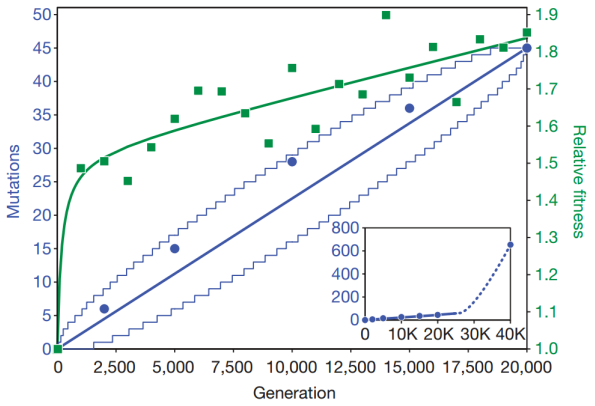


Figure 4: **Rates of genomic evolution and fitness improvement**. Blue circles show the total number of genomic changes relative to the ancestor in each sampled clone. The blue line represents a model where mutations accumulate uniformly over time. The light blue curves define the 95% confidence interval for this linear model. Green squares show the improvement of this populationâĂŹs mean fitness relative to the ancestor over time, and the green curve is a hyperbolic plus linear fit of this trajectory. Each fitness estimate is the mean of three assays; most of the spread of points around the fitness trajectory reflects statistical uncertainty inherent to the assays. The inset shows the number of mutations in the 40,000-generation clone. Reproduced from [1]

.

involved A:T pairs mutating to C:G, a mutation that has an 11.3% chance of being synonymous.

**Question 2e:**

With a genome size of $4.57 \times 10^6$ bp, what was the mutation rate for this timeframe? How does this compare to the original mutation rate of $1.6 \times 10^{-10}$ per bp per generation? Given your knowledge of how mutations arise, what might have happened near the 25,000th generation that caused this discrepancy?

Question 2f (EXTRA CREDIT):

Typically, an allele with a frequency within a population as low as 0.01 would not be detectable with sequencing. However, the increasing ease of sequencing entire genomes has made the technique of "deep sequencing" possible, in which a population is sequenced many (possibly hundreds) of times over so that even rare alleles have a high probability of being detected. In the sequencing techniques used by Lenski, the genome received 50× coverage, meaning that each nucleotide was read at least fifty times. What is a reasonable lower bound of allele frequency that you would expect to be able to detect? Remember that sequencing techniques are not perfect, with possible error rates around 1%.