HST.508/Biophysics **170 Evolutionary Genomics**

Problem Set 1

Due September 22, 2023, 11:59pm

*** You are free to discuss problems and work together,*
*but you must hand in your own unique copy of the answers. ***

1. Mutation and Drift for the Two-Allele Model (6 points)

Consider the evolution of heterozygosity H_t (and homozygosity Gt) in a population subject to a random drift and mutations. Use the approach developed in class to consider a two-allele model, which has the same rate of mutation, μ, for alleles A2 -> A1 and A1 -> A2.

a. Obtain an expression for G_{t+1} as a function of G_t, the mutation rate, and population size. To get this expression, consider all possible ways for getting two identical alleles in the t+1 generation, given that homozygosity (probability of two identical alleles) in generation t was Gt.

b. Expand the obtained expression while dropping the terms higher than the first power of or 1/N. Obtain an expression for the steady state heterozygosity H_{ss} in the two-allele model. How does the change in heterozygosity in one generation subject to mutations and drift compare to the case due to drift only?

c. Extra credit [+2]: Now consider a k-allele model with the same rate for all possible mutations Ai->Aj. How does the number of alleles affect the steady state heterozygosity? Consider limits of k=2 and k-> inf**.**

---

2. Simulations of Genetic Drift (8 points)

Start with a population of N=100 diploid individuals, each of which has two homologous chromosomes, giving you a total of 2N chromosomes to track. Each chromosome has a single polymorphic locus (SNP) that can be in one of two states. To code this, create an array of integers, of length 2N, and set the value of each element to either 1 or 2 depending on the allele of this chromosome. In the initial population, alleles 1 and 2 are equally abundant, i.e. p=q=0.5. Please take a look at the included starter code if you need guidance (pset1_sample_code.py).

a. Simulate drift in one population. Generate the next generation by drawing individuals at random from the current generation with replacement. Compute heterozygosity for each generation. Continue until all individuals become of one type -- this is called fixation. Show a plot of heterozygosity (p) over time for 1 trajectory.

b. Simulate 1000 trajectories, keeping a record of the fixation time for each simulation. Compute the mean time to fixation and compare it to N. You can run your simulations at 3-5 different values of N to establish this dependence better.

c. Plot 10 individual trajectories of heterozygosity vs. time on a single graph. In addition, for each time point, compute the average heterozygosity by averaging over all 1000 trajectories and plot the average heterozygosity vs time. To better see the exponential decay of heterozygosity you can plot the log of heterozygosity vs time. Measure the rate of decay of heterozygosity (the slope on the log-linear plot). Compare it to N.

d. Consider an effect of changing population size on the time to fixation. Run simulations first at some N=N1 and then change N to N=N2 for t2 number of generations then back to N1. Try N2<<N1, e.g. N2=N1/10. Make 1000 runs and compute the mean time to fixation. Try several N1 and N2. How does decrease in the population size change the mean time to fixation? Compute Neff for each parameters (N1,N2, t2), and correlated mean time to fixation with N1, N2, and Neff. Interpret your findings.

e. [**Extra credit**: +2 points] Plot the distribution of the fixation time for each N. How broad is the distribution? Think of various ways to quantify this. Study the mean time to fixation as a function of the number K of alleles (types of individuals), starting with 1/K initial frequency.

---

3. Parameters and population genetics (3 points)

A researcher is trying to understand the genetic mutations that lead to resistance to a new antibiotic they have discovered, called techacillin. To do this, they decide to grow many replicate populations of E. coli in a morbidostat (an experimental device that keeps a constant population size and growth rate, while increasing antibiotic concentration as bacteria become more resistant) until they have reached a high level of techacillin resistance. They will then sequence the final and intermediate mutants and identify the mutations that emerged and their order.

a. The researcher wants to observe the full diversity of possible resistance mutations, including those with relatively minor, but significant, effects on resistance (>10% increase in growth rate in presence of low levels of drug). To do so, they will need to minimize clonal interference between mutants. They estimate that during every replication, 10^{-5} new mutations emerge that provide significant resistance (as by the aforementioned criteria). What is the maximum population size they can use without interference occurring?

b. Why would the researcher want to use the highest possible population size within this regime?

c. How will the types of observed mutations (in terms of fitness effect) change if they use a higher population
size?

---

4. Effective population size. (3 points)
Derive a formula for the effective population size, Ne, if the number of males (Nm) and number of females (Nf) are not equal. (Hint: use the same method we used to derive Ne for the case of a population that is fluctuating in size.)

To fix a concrete case, consider the following example. Imagine a zoo population of primates with 20 males and 20 females. Due to the dominance hierarchy, only one of the males actually breeds. What is the relevant population size that informs us about the strength of drift in this system? Is it 40? 21? To get the answer, compute the probability that two genes drawn at random are alike (identical by descent) in this new situation, depending on random draws of genes from both the males (actually just one male) and 20 females

5. Extra Credit: Mean time to either fixation or lose (10 points)

The mean time of either fixation or loss $\bar{t}(p)$ can be derived similar to the way probability of fixation was derived in the class. Consider a two-allele model, an allele with initial frequency *p* in a population of *N* diploids without selection or mutations. Derive a recursive relationship for the mean time to fixation-or-loss, which only slightly differs from the recursive relationship from the probability of fixation

(a) To obtain ODE for $\bar{t}(p)$ do Taylor expansion while keeping only two leading terms.

(b) Write down boundary conditions for $\bar{t}(p)$ and solve the equation.

(c) Discuss genetic implications of your result.

Consider $\bar{t}(p)$ for a new mutation. For how long does a polymorphism live a population? Discuss the role of the population size, effect of bottlenecks. Make numeric estimates.