APPHYS 237 Problem Set 3

DUE: 2/20/20

Directions: Everyone should do Problems 1, 2, and 3, and **one** other problem of your choosing.

Data files available at: `https://bgoodlab.github.io/courses/apphys237/data_files.zip`

Problem 1: Heuristics for recessive mutations

The goal of this problem is to have you practice using the heuristic approach we discussed in class to work out the dynamics of recessive mutations. In the course so far, we have primarily focused on evolution in ***haploid*** organisms like bacteria, which carry just a single copy of each of their chromosomes. Humans and other ***diploid*** organisms carry two copies of each chromosome, and this requires additional assumptions about how mutations on different copies of each chromosome combine to determine the individual's phenotype (a phenomenon known as ***dominance***). An extreme limit of dominance is a ***recessive mutation***, which must be present in all chromosomes within an individual before it exerts its cost or benefit. Some of the most well known genetic diseases in humans (e.g. sickle cell disease) are caused by recessive mutations, so they play an important role in the field of human genetics.

We'll consider a very simple model of diploid reproduction, in which individuals are formed by randomly choosing 2 chromosomes that exist in the current population. In the diffusion limit, the population frequency of a recessive mutation will satisfy

$$\frac{\partial f}{\partial t} = sf^2(1-f) + \sqrt{\frac{f(1-f)}{2N}}\eta(t)\,. \tag{15}$$

where N is the number of individuals in the population. Unlike the single-locus models we have been considering so far, the low-frequency limit,

$$\frac{\partial f}{\partial t} = sf^2 + \sqrt{\frac{f}{2N}}\eta(t),, \tag{16}$$

now includes a nonlinear selection term, so we can no longer derive an exact solution for the dynamics using the method of characteristics. However, as you will see below, the heuristic approaches we discussed in class will continue to work perfectly well for this case.

(a) Repeat the heuristic derivation from class to partition frequency space into drift-dominated and selection-dominated regimes. For which values of N and s will selection be effective in at least some part of frequency space?

(b) Use your results in (a) to approximate the fixation probability and fixation time of a strongly beneficial recessive mutation. How does compare to the haploid case that we analyzed before?

(c) Use the same approach to analyze mutation-selection balance for a strongly deleterious recessive mutation. What is the maximum typical frequency of a recessive mutation with a near lethal effect ($s \approx -1$) in a population of size $N = 10^6$? What is the typical age of such a mutation?

Problem 2: The molecular diversity of adaptive convergence

The combination of laboratory evolution and whole-genome sequencing provides an opportunity to measure the *distribution* of possible evolutionary responses to a given selection pressure. The power of this approach was first demonstrated in a landmark study by Tenaillon and colleagues[7] in 2012. A total of $n = 114$ populations of *E. coli* were evolved in high temperature for $T = 2000$ generations, and a single clone was isolated and sequenced at the final timepoint. The mutations that were detected in each clone are listed in the data file `tenaillon_etal_2012_mutations.txt`. In this problem, we will treat the evolutionary dynamics of this experiment as a black box, and try to use the statistics of the observed mutations to see what we can learn about the targets of natural selection in this environment.

(a) Most tests for natural selection are based on a comparison between putatively neutral regions of the genome and those that might be subject to selection. A classic approach is to compare the relative divergence (i.e., the number of observed mutations per site) at synonymous vs nonsynonymous sites – also known as a **dN/dS** ratio. If synonymous mutations evolve neutrally, then a dN/dS ratio greater than 1 indicates that some of the nonsynonymous mutations must have been positively selected. Calculate separate dN/dS ratios for the missense and nonsense mutations in the Tenaillon et al data (`tenaillon_etal_2012_mutations.txt`). Is there enough evidence to conclude that mutations in both classes are positively selected?

The dN/dS test is a relatively coarse measurement, since relies on very general *a priori* considerations to partition mutations into putatively neutral and functional categories. In replicated experimental designs like this one, repeated observations of the same (or similar) genetic change in different populations provide a powerful alternative for identifying fine-grained targets of selection. This is an example of a more general concept known as ***parallel*** or ***convergent*** evolution.

(b) We'll first examine signatures of convergence at the single nucleotide level. Focusing on the point mutations[8] in the Tenaillon et al dataset, calculate the total number of sites that were mutated m or more times across the $n = 114$ replicates, and plot this function for different values of m. How many sites would we expect to see at a given value of m if the same number of mutations were distributed evenly across all the sites in the *E. coli*[9] genome? Is there a value of m above which you would conclude that the mutations are probably beneficial? What fraction of the observed point mutations do these sites account for?

(c) Now repeat part (b) at the gene level. Calculate the total number of genes in which we observed m or more mutations[10] across the $n = 114$ datasets, and plot this function for different values of m. How many genes would we expect to see at a given value of m if the same number of mutations were distributed evenly across the genes in the *E. coli*[11] genome? Is there a value of m above which you would conclude that some mutations in the gene are probably beneficial? What fraction of the observed mutations do these genes acount for?

(d) Part (c) shows that some genes acquire mutations at significantly higher rates than expected by chance, presumably because they are targeted by positive selection. We can try to estimate the total number of genes that are targeted in this way with the help of a ***saturation curve***.

[7] O. Tenaillon, *et al* (2012), "The Molecular Diversity of Adaptive Convergence," *Science* **335**:457–461.

[8] i.e., exclude `indel` or `structural` mutations

[9] Recall that you calculated the genome length for this strain of *E. coli* in Problem 7 of Problem Set 1.

[10] Include all `nonsense` and `missense` mutations, as well as `indel` mutations that occurred in a gene.

[11] Recall that you calculated the number of genes for this strain of *E. coli* in Problem 7 of Problem Set 1.

By choosing random subsets of the replicate populations, plot the average number of genes that were mutated in 3 or more populations in subsamples of size $n = 3, \ldots, 114$. Does this function look like it has saturated at $n = 114$?

(e) To gain some theoretical intuition for these saturation curves, let p_i be the probability that we observe a mutation in gene i in a given population. What is the probability of observing mutations in this gene in $\geq$3 populations in an experiment with n replicate populations? Plot this quantity as a function of n for $p_i = 3/114$, $5/114$, and $10/114$. For each value of p_i, what fraction of genes are likely to be detected in an experiment with $n = 114$ replicates? Based on your theoretical and empirical curves, what is your best guess for the total number of genes that are likely to be beneficial in this environment? (There is no right or wrong answer for this part.)

(f) **Bonus:** A potential complication for the saturation curve analysis is part (d) is ***epistasis***, which could cause the beneficial effect of a mutation to depend on other mutations that have accumulated in the same genetic background. If true, this could potentially show up in the co-occurence patterns of mutations in different replicate populations. As an example, consider mutations in the *rho* and *iclR* genes. How many populations have mutations in both genes simultaneously? Is this more or less than we expect by chance, given the same number of total mutations in both genes? Based on your findings, do you think this example is consistent with a simple model where mutations in *iclR* are *only* beneficial in a genetic background with a *rho* mutation?

Problem 3: Measuring the DFE for *de novo* beneficial mutations, Part I

A common criticism of DFE estimates obtained from deletion screens (e.g. Problem 4 of Problem Set 2) is that they only provide information about a narrow spectrum of mutations. One would really like to estimate the fitness effects of the beneficial mutations that actually occur in a given environment. Levy, Blundell, and colleagues[12] devised a clever method to do this in a high throughput way, using a variation of the standard pooled fitness assay.

The basic idea is to start with a large pool of strains, each labeled with a unique DNA barcode. This time, however, the barcodes are inserted in a common location in the genome, so that the strains are initially neutral with respect to each other. After a few cycles of evolution, some fraction of the lineages will acquire a beneficial mutation, and this can be detected by a sudden increase in frequency of their respective barcode as measured by PCR amplification and sequencing.

While the basic idea is simple, implementing this approach requires a careful integration between theory and experiment, involving many of the theoretical concepts we have covered in this course. We will work through the key steps in their analysis in the next two problem sets.

(a) The first step is to determine the parameters of the experiment. In particular, we get to choose:

- The total number of generations that the lineages are monitored over, T.
- The total number of cells in the population that are transferred at the dilution step, N_b.
- The total number of barcoded lineages, B.

[12] Levy, Blundell, *et al*, (2015), "Quantitative evolutionary dynamics using high-resolution lineage tracking," *Nature* **519**:181–186.

- The total number of sequencing reads, D, to generate for each timepoint.

For the experiment to work as planned, we'll need to choose these parameters so that the following criteria are met:

- A large number of barcoded lineages (e.g., ~ 1000) acquire a beneficial mutation during the T generations of the experiment.
- Only a small fraction of these acquire ***multiple*** beneficial mutations over this time period.
- Beneficial mutations ***do*** noticeably perturb the frequency of the lineage that they occur in (so that we can actually observe them).
- Genetic drift ***does not*** substantially perturb the lineage frequencies on the same timescale (i.e., if we see a several-fold change in frequency, we want to be able to attribute it to selection rather than random genetic drift).

Of course, these criteria themselves depend on the fitness effects and mutation rates of new beneficial mutations – precisely what this experiment is trying to measure. Previous experiments suggested laboratory evolution experiments in yeast were consistent with a typical beneficial mutation rate of order $U_b \sim 10^{-5}$ and a typical fitness effect of order $s_b \sim 10^{-2}$. Using these estimates, what values of T, N_e, and B would you suggest to your experimental collaborators? How many sequencing reads would you need to generate for each timepoint? How many lanes of sequencing would you need for the experiment?

Choose just **one** of the following problems.

Problem 4: Continuous-time branching process with bursty reproduction

A classic microscopic population model is the continuous-time branching process. This is a discrete-individual model, in which every individual has an independent probability of giving birth or dying in an infinitesimal time interval dt. We'll denote the birth rate and death rate by B and D respectively. When an individual gives birth, we'll assume that it replaces itself with a "burst" of exactly K offspring. The continuous-time branching process has numerous applications outside of evolution, e.g. the production of muons from chain reactions seeded by cosmic rays in the atmosphere. Here, we will use it as a model of the number of mutant individuals in a large population. To that end, we'll measure time in (wildtype) generations by taking $B = 1 + b$ and $D = 1 + d$.

(a) Let $n(t)$ denote the (random) number of descendants of a single individual after t generations. Derive a differential equation for the generating function $H(z,t) = \langle e^{-zn(t)} \rangle$.

Hint: This is easiest to do using a recursion argument. Start by writing $e^{-zn(t+dt)}$ on the left hand side, and consider the very first time slice $(0, dt)$. At the end of this time slice, we will either have 0, 1, or K individuals. What are the relative probabilities of these three events? Conditioned on each outcome, can you write $e^{-zn(t+dt)}$ using one or more independent copies of the original process $n(t)$? If so, you can then average both sides to arrive at an ordinary differential equation for $H(z,t)$.

(b) Solve your differential equation in the special case where $K = 2$, using the initial condition $n(0) = 1$. Compare your results to the diffusion model we discussed in class. Based on this result, do you think the continuous-time branching process belongs to the same universality class in the limit that $b, d \ll 1$? If so, what are the effective parameters? Use this result to comment on relevance of discreteness of individuals or birth rate vs death rate differences in the diffusion limit.

(c) When the burst size is greater than 2, the generating function no longer has a closed form solution. This mode of reproduction is relevant for some viruses, which often produce many multiple new viral particles per infected cell. Using the property that $H(z,t) \approx 1 - z\langle f(t) \rangle$, expand your differential equation to lowest order in z to derive a differential equation for $\langle f(t) \rangle$. What is the long-term growth rate, s_e, of the average frequency as a function of b, d, and K?

(d) In the limit of long times, we expect the generating function to approach a constant value, $H(z,t) \approx e^{-z \cdot 0}(1-p)$, where p is the survival probability. Solve for the survival probability in the limit that $pK \ll 1$, and compare this to the $K = 2$ case at the same long-term growth rate, s_e. For what values of s_e and K do you expect this expression to break down? What happens to the survival probability in this case? Can you give an intuitive explanation for this behavior?

Problem 5: Sweep times vs fixation times from natural selection

The goal of this problem is to give you a numerical feeling for some of the relevant timescales of natural selection.

(a) How many generations are required for a beneficial mutation with fitness effect s to go from 10% to 90% frequency? From 1% to 99%? We will call this the ***sweep timescale***, $T_{\rm sw}$, since it is the time required for a mutation to visibly sweep through a population (e.g. in metagenomic data).

(b) Estimate the sweep timescale (in days) for a mutation with a 1% fitness benefit in Lenski's long-term evolution experiment in *E. coli* (Problem 4 of Problem Set 1). Then estimate the same quantity for a population of bacteria in an individual's gut microbiome. (We don't know what the generation time is in this case, but estimates range from $\sim 1-10$ generations per day.)

(c) Compare these sweep timescales with the ***fixation timescale*** $T_{\rm fix} \sim \frac{1}{s}\log(N_e s)$, which is the time required for a newly produced variant to reach observable frequencies in the population. Estimate the fixation timescale for the same 1% mutation in Lenski's experiment and in the gut microbiome. (We don't know what the effective population size is, but for this problem let's assume that it is similar to the census population size, $\sim 10^{12}$ cells.)

(d) Use your answer in (b) to speculate about the following scenario: let's imagine that a host starts a new diet that renders a particular metabolic pathway unnecessary for the gut bacteria, and that a $\sim$1% benefit could be gained by eliminating the resources that are currently devoted to it. How long would the individual have to adhere to the new diet before we could hope to observe a new loss-of-function variant at appreciable frequencies in the within-host population? How does this compare this to the case where a strain with the loss-of-function mutation was already present in the host at 1% frequency.

Problem 6: Correlated evolution and protein-protein interactions

We previously saw how sequence conservation can signal functionally important regions of proteins (or genomes). An extension of this idea is that slightly less constrained but *correlated* evolution at different sites in a genome can signal interactions between the corresponding genomic regions. In this problem, we will explore a classic example of correlated evolution in signal transduction pathways.

In order to respond to changes in the environment, bacteria employ a family of proteins known as the **two-component signal transduction system**. Each pathway in this family typically contains a transmembrane protein known as the **histidine-kinase (HK)**, which senses some condition outside the cell, and a corresponding **response regulator (RR)**, which can receive signals from its partner HK and then go on to effect changes in cellular physiology or behavior. These HK-RR signaling systems are found throughout the bacterial kingdom, with most species containing 20 to 30 HK-RR pairs. However, there is little crosstalk *between* different HK-RR pairs, despite a large degree of sequence similarity within the HK and RR families. This suggests that the sequences of the HK and RR proteins are tuned to interact with their specific partner. In this problem, you will use information theory to explore the molecular basis of this specificty.

The file `skerker_etal_hk_alignment.txt` contains a multiple alignment of the amino acid sequences of a portion of the HK protein across 1,297 different signaling pathways.[13] Each row contains the protein sequence of a different HK protein, and each column gives the amino-acid at that position in the sequence (with gaps denoted by '`-`'). The files `skerker_etal_rr_alignment_1.txt` and `skerker_etal_rr_alignment_2.txt` contain an analogous alignment for a portion of the RR protein. One of the two files (we don't know which) is sorted so that the each RR protein lines up with its partner in `hk_alignment.txt`. The other file lists the RR proteins in a random order.

(a) For each file, calculate the mutual information,

$$\mathrm{MI}(a_i, a_j) = -\sum_{a,a'} \Pr(a_i = a,\, a_j = a') \log \left[\frac{\Pr(a_i = a,\, a_j = a')}{\Pr(a_i = a)\Pr(a_j = a')} \right] \tag{17}$$

between each site i in the HK protein and each site j in the RR protein. Plot the distribution of MI values for each file as a histogram. Based on this information, which file do you think corresponds to the proper pairing of HK and RR proteins? Explain your reasoning.

(b) If you wanted to "rewire" an HK protein to interact with a different RR protein by switching a single amino acid residue, which position would you want to mutate? Explain your reasoning. (Amazingly, Skerker *et al* tried this and it actually worked!)

(c) Fitness valley crossing is often cited as a potential mechanism for creating the high mutual information at the sites that control interaction specificity. The idea is that a deleterious mutation that destablizes the interaction can be rescued by a compensatory mutation in the interaction partner that restores the function of the interaction. Let's try explore the feasibility of this process using order-of-magnitude estimation. Consider a pair of sites. What is the substitution rate of valley crossing mutations if the valley has a fitness cost s_d, the sequences on either side have the same fitness as the wildtype, and the mutation rate at both

[13] Data from Skerker *et al* (2008) "Rewiring the Specificity of Two-Component Signal Transduction Systems," *Cell* **133**, 1043–1054.

sites is on the order of the per site mutation rate, μ. Substituting reasonable values for these parameters, how many such mutations would you expect to see at a pair of sites in $\sim$1000 gene families over the total number of generations that have elapsed since the origin of life ($\sim$4 billion years ago). Compare this to the number of double mutations you see at your informative site in part (b) above. Do you think this simple valley crossing explanation is reasonable?