APPHYS 237 Problem Set 1

DUE: 1/21/20

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

Problem 1: Molecular evolution and genetic diversity in the influenza virus

The text file `influenza_HA_dna_sequences.fasta` contains a list of 841 complete DNA sequences of the hemagluttinen (HA) gene from influenza virus samples collected between 1968 and 2005. Hemagluttinen is a surface protein that allows the viruses to enter host cells, making it a primary target for neutralizing antibodies. This creates a strong selection pressure for the HA gene to evolve over time to evade these immune defenses.

(a) Calculate the number of single nucleotide differences between the first sample (`A/Aichi/2/1968`) and the remaining samples, and plot the results as a function of the sampling year. How many differences have accumulated over this ∼40 year period? What fraction of the HA gene does this account for?

(b) Calculate the number of genetic differences between all pairs of strains from the same year, and plot the distribution of this quantity aggregated across all years. Estimate the genetic "turnover time" – i.e., how long would we have to wait for the population to accumulate the same number of genetic differences that typically separate co-circulating strains.

Problem 2: The Luria-Delbrück experiment

In the early 1940s, Salvador Luria was conducting experiments to understand what made bacteria resistant to viruses. Many had observed that the offspring of resistant bacteria were also resistant, but it was unknown whether the virus induced resistance (like animals that become resistant from surviving an infection) or if the bacteria acquired resistance through a spontaneous mutation prior to encountering the virus. Luria and his colleague Max Delbrück settled the question and provided a method of measuring mutation rates, known as the Fluctuation Test, for which they received a Nobel Prize.

Consider a population of bacteria that grows for T generations, reaching a final size of $N_T = N_0 2^T$ cells. The population is then exposed to the virus and the number of resistant individuals are counted (e.g. by plating and counting colonies). If resistance is induced by the virus, then the number of resistant bacteria should be Poisson distributed with mean pN_T, where $p \ll 1$ is the probability of acquiring resistance during the encounter.

However, if resistance is acquired through spontaneous mutations, we must also account for individuals that inherited their resistance phenotype from mutation events that occurred before exposure to the virus, while the population was still growing. Assume that none of the initial N_0 cells are resistant, and let $\mu \ll 1$ be the probability that each of the daughter cells acquires a mutation during division. Let M_T denote the total number of mutant cells in the population at the end of the growth phase. The distribution of M_T is known as the **Luria-Delbrück distribution**, and it remains an active area of research today.

(a) What is the expected number of new mutations that are produced in generation t? Call this number $\theta(t)$. The actual number of mutations that are produced will be a random variable

which we will denote by $m(t)$. We will assume that $m(t)$ will be Poisson distributed with mean $\theta(t)$.

(b) If a mutation arises in an individual at generation t, how many descendants does it leave at time T, assuming that it grows deterministically? Call this number $n(t)$. Write an expression for M_T as a sum over $m(t)$ and $n(t)$. For simplicity, you may assume that mutations are sufficiently rare that all mutations arise in previously non-resistant cells.

(c) Use your results to calculate the mean and variance of M_T, which we will denote by $\langle M_T \rangle$ and $\mathrm{Var}(M_T)$. It is useful to compare these quantities to their expected relationship under a Poisson distribution using using a so-called *Fano factor*,

$$F = \frac{\mathrm{Var}(M_t)}{\langle M_t \rangle} \tag{1}$$

which is equal to 1 for Poisson distributions. How much larger is F for the Luria-Delbrück distribution? How might you use this information to distinguish between the induction vs mutation hypotheses above?

(d) Unfortunately, the same properties of the Luria-Delbrück distribution that allow us to distinguish the two hypotheses make it difficult to measure $\langle M_t \rangle$ and $\mathrm{Var}(M_t)$ in practice. To see this, suppose we ran n independent replicates of this experiment and calculated the sample average,

$$\overline{M}_T = \frac{1}{n} \sum_{i=1}^{n} M_{T,i} \tag{2}$$

For sufficiently large n, the sample average $\overline{M}_T$ will approach the theoretical mean $\langle M_T \rangle$. The relative error can be estimated by the cofficient of variation,

$$CV = \frac{\sqrt{\mathrm{Var}(\overline{M}_T)}}{\langle \overline{M}_T \rangle} \tag{3}$$

Calculate the coefficient of variation for $\overline{M}_T$. How many independent experiments would you need to run to ensure a relative error of order ϵ? What happens in the limit that $N_0\mu \ll 1$?

(e) This pathological behavior arises from the fact that the theoretical mean and variance are averaging over rare events: "jackpot" mutations that occur early in the growup phase and have an outsized impact on M_t. When $N_0\mu \gg 1$, large numbers of jackpots occur in the first generation of growth, and the behavior of M_t is relatively predictable. However, when $N_0\mu \ll 1$, jackpots are so rare that they will not occur in a typical experiment, though they continue to influence the theoretical mean. This suggests that there is reason for hope: if jackpots are causing all the problems, then the behavior across replicate experiments should be *more* predictable if we know that some jackpots have definitely *not* occured – we just have to come up with a way to predict the *typical* values of M_t in a set of n experiments. We will explore one such scheme here.

Calculate the total number of mutations that are expected to arise before generation t, across all n populations. Call this number $\theta_<(t|n)$. Find the critical value t^* where $\theta_<(t^*|n) \sim 1$. For

$t \ll t^*$, there will typically not be any mutations earlier than t in any of our replicates. This suggests that one might be able to predict the *typical* behavior by repeating our calculations above with a modified version of $\theta(t)$:

$$\hat{\theta}(t|n) = \begin{cases} 0 & \text{if } t < t^*(n) \\ \theta(t) & \text{if } t \geq t^*(n) \end{cases} \tag{4}$$

which enforces this typicality constraint. Use this expression to calculate the typical mean and variance across the replicates as a function of n, as well as the coefficient of variation. Is there still any pathological behavior in the $N_0\mu \ll 1$ limit? Does the coefficient of variation scale in the way we expect from the central limit theorem?

(f) **Bonus:** At this point you might be worried, because we did not use the full Luria Delbrück distribution for our serial dilution model in class. However, the serial dilution model contains one extra step that we have not considered here: the dilution of the final culture into a new flask with the initial size N_0. The number that matters is the number of mutants in the new flask, M_T', which will be Poisson distributed with a *random* mean, $N_0(M_T/N_T)$. Calculate the mean and variance of M_T', as well as the Fano factor. How large is the deviation from the Poisson approximation we used in class?

Problem 3: Single Locus Simulations

(a) Write a computer program that simulates the frequency trajectory of a mutation in the serial dilution model described in the lecture notes. Plot a few example trajectories starting at an initial frequency of $f(0) = 0.5$, with different values of N and s ($N = 10^2, 10^3, 10^6$, $s = 0, 10^{-2}, 10^{-3}$.

(b) Modify your simulation to include mutations, using the simple approximation described in the Lecture notes. Plot a few example trajectories with $N = 10^4$, $f(0) = 0$, and $\mu = 10^{-5}$, both for (i) a deleterious mutation with $s = -10^{-3}$ and (ii) a beneficial mutation with $s = 10^{-2}$.

Problem 4: Competitive fitness in a long term evolution experiment in E. coli

One of the longest running laboratory evolution experiments was started by Richard Lenski in 1988 and is still in progress today. Lenski founded 12 independent populations of *E. coli* from a common ancestor strain, and he and his team have been propagating these 12 populations in glucose-limited media using a serial dilution protocol similar to what we discussed in class. This experiment uses a 1:100 dilution factor, so that the populations experience about $\log_2(100) \approx 7$ generations a day with a daily bottleneck size of $N_b \approx 5 \times 10^6$. Every 500 generations, a copy of each population is cryogenically preserved for future study. After more than 30 years, Lenski's experiment has produced >1500 archived samples covering >70,000 generations of evolution in the same controlled conditions.

Among other applications, these frozen population samples are used to measure the fitness of the evolved populations using a variant of the fitness assay we described in class. Variants of the ancestral strains were created that produce different colored colonies when grown on a special media in Petri dishes. These modified ancestral strains are mixed with the evolved populations (usually at a 50:50 ratio) and are competed for Δt generations (typically one daily cycle). The relative frequencies at the beginning and end of the cycle are measured by plating the cultures and

counting the number of colonies of each type. The relative fitness of the evolved population (S) can be estimated by the plug-in estimator,

$$S \equiv \frac{1}{\Delta t} \log \left(\frac{N_{\text{pop}}(\Delta t)}{N_{\text{anc}}(\Delta t)} \frac{N_{\text{anc}}(0)}{N_{\text{pop}}(0)} \right) . \tag{5}$$

where $N_i(t)$ denotes the number of colonies of each type at a given timepoint.

(a) The file `LTEE_ancestor_fitness_assays.txt` contains the results of ∼500 fitness assays performed by Wiser *et al* (*Science*, 2012). Approximately 250 samples were assayed across 6 populations, with 2 biological replicates for most of the samples. The difference between the fitness estimates from these biological replicates provides an estimate of the uncertainty in the fitness measurements. Plot the distribution of these errors across all timepoints. What is the typical uncertainty in an individual fitness measurement using this approach?

(b) After averaging over the two replicates at each timepoint, plot the fitness trajectories for each population as a function of time.

(c) Previous studies have argued that these fitness trajectories can be fit by the logarithmic function,

$$X(t) = X_c \log \left(\frac{v_0 t}{X_c} \right) , \tag{6}$$

with $X_c \approx 4.6 \times 10^{-2}$ and $v_0 = 7.7 \times 10^{-4}$. Plot this function against your data. Does it look consistent? What is the predicted fitness gain between generation 40,000 and 50,000? How does this compare to the measurement uncertainty estimated above?

(d) The file `LTEE_40k_fitness_assays.txt` contains another ∼800 fitness assays performed by Lenski *et al* (*Proc R Soc B*, 2012). Unlike the previous experiments, where the evolved populations are competed against the ancestor, these experiments compete the evolved population against a reference strain that was isolated from one of the populations at generation 40,000. They also use a longer competition period (3 daily cycles, rather than 1) and perform more biological replicates for each sample. Use these data to calculate the gain in fitness between generation 40,000 and 50,000, and between generation 50,000 and 60,000, along with the uncertainties in these estimates. Is there evidence that fitness is still increasing in Lenski's experiment at these later timepoints?

Problem 5: Pooled fitness assay

Suppose that you have a population that contains a mixture of K different strains, each with its own fitness difference s_k ($k = 1, \ldots, K$) relative to a reference strain.

(a) Using the serial dilution model we discussed in class, calculate the relative frequencies of each strain after one cycle of growth (i.e., just before the dilution step), assuming that each strain starts at a relative frequency $f_k(0)$.

(b) Neglecting noise, solve for the relative frequencies of each strain after another cycle. Can you write a formula for the relative frequencies after an arbitrary number of cycles? or an arbitrary time t?

(c) Suppose that the fitnesses s_k were all shifted by a constant amount c. What happens to the dynamics of the strain frequencies $f_k(t)$? What does this mean for our ability to measure s_k by tracking strain frequencies over time?

(d) One way to avoid this issue is to ensure that one of the strains in the pool (e.g. $k = 0$) is the common reference against which fitness is measured (e.g. the wildtype), so that $s_k = 0$. What is the value of $f_k(t)/f_0(t)$ in this case? Use this result to generalize the formula in Eq. 5 to the multi-strain case, where you have counts N_k of each strain k at a pair of timepoints. This is known as a ***pooled fitness assay***. With the advent of DNA sequencing, pooled fitness assays have become a common tool to measure the fitness of large collections of mutants simultaneously (e.g. all single gene deletions of a given strain) – we will see examples of these in a later problem.

(e) For the moment, let's stay in theory land. Let's assume that K is very large, and that the fitnesses of the non-wildtype strains are normally distributed with mean $\mu = 0$ and variance σ^2. Assuming that the wildtype strain starts at frequency f_0, and all the remaining strains are evenly distributed, write a formula for the frequency trajectory of a focal strain k with fitness $s_k > 0$ as a function of time. The following property of normal distributions may be useful:

$$\langle e^{zx} \rangle = e^{\frac{1}{2}\sigma^2 z^2} \tag{7}$$

Is the frequency trajectory monotonic? If not, when does it reach its maximum?

Problem 6: Experimental evolution in a chemostat (in theory)

In addition to the serial dilution model we discussed in class, another common protocol for experimental evolution makes use of a continuous-culture device known as a ***chemostat***. A chemostat is a well-mixed vessel (volume V) in which nutrients are fed in at a fixed rate via an input tube, and cells and nutrients are continually removed through an output tube. In this setup, the number of cells in the vessel, $n(t)$, as well as the nutrient concentration, $c(t)$ (in units of cell biomass equivalents per unit volume), are both dynamical variables that adjust according to the internal dynamics of the system. For a single strain growing in isolation, these dynamics can be written in the form

$$\frac{\partial n}{\partial t} = \underbrace{r(c)n}_{\text{growth}} - \underbrace{\delta n}_{\text{dilution}} \tag{8}$$

$$\frac{\partial c}{\partial t} = \underbrace{\delta c_{\text{in}}}_{\text{input}} - \underbrace{\delta c}_{\text{dilution}} - \underbrace{\frac{r(c)n}{V}}_{\text{growth}} \tag{9}$$

where δ is the ***dilution factor*** (i.e., the fraction of the total vessel volume that flows in and out per unit time), c_{in} is the concetration of nutrients in the input tube, and $r(c)$ is the growth rate of the microbe as a function of the resource concentration. In the absence of evolution, the system will eventually approach a (non-equilibrium) steady state characterized by a constant values of $n(t)$ and $c(t)$. Let's call them $n(t) = n^*$ and $c(t) = c^*$, respectively.

(a) Solve for the growth rate, $r^* \equiv r(c^*)$ that is achieved at this steady state – this gives a measure

of the effective generation time, $\tau_g = 1/r^*$. How do r^* and τ_g depend on the concentration of the input nutrients, $c_{\rm in}$?

(b) Solve for the population size n^* as a function of c^*. In many cases of interest, we will have $c^* \ll c_{\rm in}$ – i.e., the microbes are eating most of the nutrients that we put it in the chemostat. Calculate the lowest order contribution to n^* in the limit that $c^* \ll c_{\rm in}$. How does this quantity depend on the growth function $r(c)$? How do you explain this result?

(c) Now we will consider how the system relaxes to this steady state. It is often useful to consider an adiabatic limit, where the dynamics of the nutrient concentration relax more rapidly than the dynamics of population size. Formally, this is equivalent to neglecting both the $\partial_t c$ and $-\delta c$ terms in Eq. 9. Use this approximation to eliminate the resource concentration from Eq. 8 and obtain a differential equation that depends only on the current value of $n(t)$. Solve this equation for $n(t)$ as a function of $n(0)$. Assuming that $n(0) \neq n^*$, how quickly does the system relax to the steady state?

(d) We will now consider competition dynamics between two strains. Suppose we have a wildtype strain $n_{\rm wt}$ with growth function $r(c)$. The functional form of $r(c)$ is not typically known, but in simple cases, it takes on a Monod-like form,

$$r(c) = r_{\rm max}\left(\frac{c}{c+K}\right), \tag{10}$$

where $r_{\rm max}$ is proportional to the expression of an enzyme in the limiting metabolic pathway. Suppose have a mutant $n_{\rm mut}$ that increases the expression of this enzyme by a factor $(1+s)$. Write down joint model for $n_{\rm wt}$, and $n_{\rm mut}$, and c. Use the adiabatic approximation above to eliminate $c(t)$ and obtain an effective model that depends only on $n_{\rm wt}$ and $n_{\rm mut}$. Use this model to obtain a corresponding equation for the total population size, $N(t) = n_{\rm wt}(t) + n_{\rm mut}(t)$.

(e) Suppose that $n_{\rm wt}(0) + n_{\rm mut}(0) = n^*$. What can we say about the dynamics of $N(t)$ at later times? Use this result to eliminate $n_{\rm wt}(t)$ and obtain an effective model for the mutant frequency $f(t) = n_{\rm mut}(t)/N(t)$. To lowest order in s, how does the form of this model compare to the serial dilution model discussed in class?

(f) Finally, we will briefly consider the effects of stochasticity. This can be tricky to define in continuous time, so let's imagine that all of the input and output flow in our device occurs in discrete timesteps of length Δt. What is the probability that a single cell falls in the dilution volume $\delta V \Delta t$? If there are currently $n_{\rm mut}(t)$ and $n_{\rm wt}$ cells of the mutant and wildtype, respectively, what is the typical variation in the number of cells of each type that are diluted out in each timestep? How does this compare to the noise term in the serial dilution model from class?

Problem 7: The *E. coli* genome

The text file `ecoli_reference_genome.fasta` contains the genome sequence of the bacterium *Escherichia coli* – specifically, the genome sequence of a lab strain named REL606, which we will encounter several times throughout this course.

(a) How long is this *E. coli* genome? What is the relative fraction of `A`'s, `T`'s, `C`'s, and `G`'s?

(b) Calculate the distribution of 20-mer's in the genome (i.e., the number of times you see each sequence of 20 bases, allowing for overlaps). What fraction of the 20-mer's occur only once? What does this tell us about fraction of sites in the *E. coli* genome that can be uniquely identified by a 20bp sequence?

(c) The text file `ecoli_genes.txt` contains a list of all the genes in this strain, along with their locations (in 1-based coordinates), and whether they are transcribed in the forward or reverse orientation. Plot the distribution of gene lengths. How many genes are there? What fraction of the genome do they account for? What fraction of genes are transcribed in the reverse orientation?

(d) Using the genetic code (`https://en.wikipedia.org/wiki/DNA_codon_table`), calculate the total number of possible synonymous mutations (those that don't change the amino acid sequence of the protein), the number of nonsense mutations (those that change one of the existing amino acids to a stop codon), and the number of missense mutations (those that change one of the amino acids without introducing a stop codon). You may assume that all base pair transitions are equally likely.