APPHYS 237 Problem Set 1

DUE: 1/21/20

Directions: Everyone should do Problems 1, 2, and 7, and **one** other problem of your choosing.

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.[1] 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

[1] The same data can also be accessed in CSV form in `influenza_HA_dna_sequences.csv`, though the FASTA format is more commonly encountered in the wild.

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_{in}?

(b) Solve for the population size n^* as a function of c^*. In many cases of interest, we will have $c^* \ll c_{\text{in}}$ – 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 $c^* \ll c_{\text{in}}$. 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_{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,

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

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

(e) Suppose that $n_{\text{wt}}(0) + n_{\text{mut}}(0) = n^*$. What can we say about the dynamics of $N(t)$ at later times? Use this result to eliminate $n_{\text{wt}}(t)$ and obtain an effective model for the mutant frequency $f(t) = n_{\text{mut}}(t)/N(t)$. 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 $n_{\text{mut}}(t)$ and n_{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.

APPHYS 237 Problem Set 2

DUE: 2/4/20

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

Problem 1: Measuring the per-base-pair mutation rate with the Luria-Delbrück fluctuation test

In Problem 2 of Problem Set 1, you worked out the theory behind the Luria-Delbrück experiment, which is often used to estimate mutation rates in the laboratory (the ***fluctuation test***). The file `lang_murray_08_fluctuation_test.txt` contains the results of one such experiment performed by Lang and Murray.[2] Approximately $n = 720$ populations of *S. cerevisiae* (baker's yeast) were grown from an initial population size of $N_0 = 2000$ for a total of $T = 13$ generations, and then plated on Petri dishes containing the drug 5-fluoroorotic acid ("5-FOA"). Resistance to this drug is primarily caused by loss-of-function mutations in the URA3 gene.[3] Thus, the number of resistant colonies in this experiment reflects the aggregate mutation rate for loss-of-function variants in this gene ($U_{\Delta URA3}$). Lang and Murray used this fact, along with some targeted DNA sequencing, to back out an estimate of the per-base-pair mutation rate. We will work through the key steps in their analysis below.

(a) The colony counts in this experiment should follow a Luria-Delbrück distribution, which has some peculiar sampling properties due to the presence of rare "jackpot" mutations. Can you pick out a few of these jackpots by eye in the data file?

(b) Revisiting the theory in Problem 2 of Problem Set 1, calculate the probability p_0 that we observe zero resistant colonies in a particular population. We can estimate this number using the observed fraction of plates with zero colonies:

$$\overline{p}_0 = \frac{\#\text{ experiments with } M_{T,i} = 0}{n} \tag{11}$$

which satisfies $\langle \overline{p} \rangle = p_0$. This was also true for the sample mean $\overline{M}_T$ in Problem 2 of Problem Set 1. Can you explain why we expect $\overline{p}_0$ to be more robust to the presence of rare jackpot events?

(c) Rearrange your expression in (b) to solve for $U_{\Delta URA3}$ as a function of p_0, and obtain an estimator $\hat{U}_{\Delta URA3}$ by replacing p_0 with the measured value $\overline{p}_0$. What is the mean and variance of $\hat{U}_{\Delta URA3}$ in limit of many replicates ($n \gg 1$)? Estimate $U_{\Delta URA3}$ and its uncertainty using the data provided above. Based on the inferred parameters, do you think that this is a reasonable fitting procedure?

(d) To connecting the phenotypic mutation rate $\hat{U}_{\Delta URA3}$ to a per-base-pair mutation rate, Lang and Murray used Sanger sequencing to sequence the URA3 gene in 237 of the resistant colonies from different plates in their experiment. 30 of these colonies did not have any mutations in

[2] Lang, G.I. and A.W. Murray (2008), "Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae*," *Genetics* **178**:67–82.

[3] 5-FOA is nontoxic on its own, but it is converted into a toxic byproduct (5-fluoro-uracil) by the uracil biosynthesis pathway. The URA3 gene catalyzes a key step in this process, so loss-of-function variants in URA3 confer resistance when grown in media containing an external source of uracil.

URA3, and presumably reflect resistance mutations that arose in other genes. The remaining colonies had just a single mutation in URA3 (or adjacent mutations that likely arose as a complex mutational event). The distribution of mutations is broken down in the following table:

Mutation type	Number of colonies
Nonsense SNVs	64
Other SNVs	103
Indels and *etc.*	40
WT URA3	30

The length of the URA3 gene is 803bp, so there are a total of 2412 possible single nucleotide variants that could be produced. From the sequence of URA3, 123 of these variants results in a nonsense mutation (i.e. a preumature stop codon, which we assume leads to a nonfunctional URA3 protein). Use these numbers to convert the phenotypic mutation rate $U_{\Delta URA}$ to a per-base-pair estimate (assuming that all single nucleotide mutations are equally likely).

(e) The same data allow us to estimate another interesting but difficult-to-observe quantity: the probability that a random single nucleotide mutation disrupts the function of a protein. Estimate this quantity using the URA3 data above.

Problem 2: Measuring the per-base-pair mutation rate from the accumulation of synonymous mutations

The falling costs of whole-genome DNA sequence have made it possible to use more direct approaches to estimate the per-site-mutation rate. One of the simplest is based on the accumulation of neutral mutations along a single line of descent. Suppose that we propagate a given strain for T generations and sequence a single clonal isolate from the population at the final timepoint. Under very general conditions, one can show that the accumulation of neutral mutations between the ancestor and sequenced individual will occur as a collection of independent Poisson process with rate μT per site, regardless of the population size, mode of propagation, or selection elsewhere on the genome.[4]

(a) If we assume that all synonymous mutations are neutral, use the theorem above to derive an estimate for the per site mutation rate based on the total number of synonymous mutations we observe in n isolates sampled from n independently evolved populations?

(b) The file `tenaillon_etal_2012_mutations.txt` contains the results of one such experiment performed by Tenaillon and colleagues.[5] 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. Use your results in part (a), along with the total number of synonymous sites you calculated in Problem 7 of Problem Set 1,[6] to estimate the per site mutation rate μ from these data. What is the uncertainty in your estimate?

(c) A 115th population was also sequenced, but the clone was found to contain a much higher number of mutations than the others (`tenaillon_etal_2012_outlier_clone.txt`). Use the

[4] A particularly elegant proof can be made using *coalescent theory*, which we will cover later in the course.

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

[6] The ancestor of this experiment was the same REL606 strain in Problem 7 of Problem Set 1.

approach you developed in (b) to estimate the per site mutation rate from this strain as well. Based on the uncertainties, can you conclude that it is significantly different than the other 114 clones? Since the populations were evolved in identical conditions, we must conclude that the outlier clone acquired a mutation that increased its genome-wide mutation rate at some point during the experiment – we will revisit the dynamics of these ***mutator*** phentotypes later in the course.

Problem 3: The molecular diversity of adaptive convergence

In the previous problem, you examined the accumulation of synonymous mutations in a collection of *E. coli* populations that were adapted to high temperature for $T = 2000$ generations. In this problem, we will examine the remaining 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?

(b) 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.

We'll first examine signatures of convergence at the single nucleotide level. Focusing on the point mutations[7] 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*[8] 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[9] 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*[10] 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?

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

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

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

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

(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***. 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 $n = 114$?

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 ≥ 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 these 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.)

(e) A potential complication for the saturation curve analysis is part (d) is that the beneficial effect of a mutation may 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 4: Universality and non-universality among serial dilution models

(a) Let's consider a more elaborate version of the serial dilution model we discussed in class, in which the transfer processes introduces some growth rate variability across individuals. Specifically, let's assume that the fitness of each individual at the beginning of the daily cycle is drawn from a Gaussian distribution with a genotype-dependent mean and variance. We'll let r and σ^2 denote the mean and variance for wildtype individuals, while $r + s$ and $\sigma^2 + \nu$ will denote the mean and variance for mutant individuals. We'll assume that these fitness perturbations are inherited by all of an individual's descendants over the entire course of the daily cycle.[11] Calculate the mean and variance of the total mutation frequency after one cycle to leading order in $1/N$, s, and ν. Does this model lie in the same universality class as the basic serial dilution model we discussed in class? If so, what are the effective parameters s_e and N_e?

(b) Now let's consider a slightly different scenario, in which fitness perturbations are created by environmental fluctuations that are shared across all individuals. Specifically, let's assume that the fitness difference between mutant and wildtype in a given cycle is normally distributed with mean s and variance ν. Calculate the mean and variance of the mutation frequency after one cycle to leading order in $1/N$, s, and ν. Does this model lie in the same universality class as the serial dilution model we discussed in class?

[11] In practice, one might imagine that these fitness perturbations will be lost over a few divisions. Our calculation therefore represents an upper bound on the magnitude of these effects.

Problem 5: Mutation accumulation in individuals vs populations

Suppose we found a population from a clonal ancestor and allow it to evolve for t generations.

1. Suppose that you know the population frequency of mutations (f_ℓ) at each site ℓ in the genome ($\ell = 1, \ldots, L$). Write a sampling formula for the average number of mutations in a randomly sampled individual from the population ($M(t)$) as a function of f_ℓ.

2. Write a sampling formula for the average number of mutations in a randomly sampled *pair* of individuals in the population. What about a random sample of size n?

3. Write a stochastic differential equation for f_ℓ assuming neutral evolution. Use this model to derive a deterministic differential equation for the average frequency, $\langle f_\ell(t) \rangle$. Solve this equation and show how $M(t)$ grows with time.

4. Now use the stochastic model to derive a deterministic equation for the second moment $\langle f_\ell(t)^2 \rangle$. Solve this equation and show how $M_2(t)$ grows with time. How long do we have to wait for the two expressions to give similar results? How can we explain the discrepancy at short times?

Problem 6: Sweep times vs fixation times

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, but estimates range from $\sim 1 - 10$ generations per day).

(c) We can contrast the sweep timescale $T_{\rm sw}$ 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 (e.g., 50%). (i) estimate the fixation timescale for the same mutation in Lenski's experiment ($N_e \approx 3 \times 10^7$) and compare it with the sweep time above. (ii) Repeat for the population of gut bacteria, assuming that the effective population size 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 ~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 7: Sequence conservation and broadly neutralizing antibodies in influenza

One of the most powerful and widely used principles in evolutionary biology is the use of sequence conservation to infer function. The basic idea is related to the concept of survivorship bias, as illustrated by the following example. (Text and image reproduced from Wikipedia.)

> *During World War II, researchers from the Center for Naval Analyses had conducted a study of the damage done to aircraft that had returned from missions, and had recommended that armor be added to the areas that showed the most damage. Statistician Abraham Wald noted that the study only considered the aircraft that had survived their missions—the bombers that had been shot down were not present for the damage assessment. The holes in the returning aircraft, then, represented areas where a bomber could take damage and still return home safely. Wald proposed that the Navy reinforce areas where the returning aircraft were unscathed, since those were the areas that, if hit, would cause the plane to be lost. His work is considered seminal in the then-nascent discipline of operational research.*

Just like the bullet holes in the aircraft above – mutations are constantly creating "holes" in the genomes of living organisms. The mutations that disrupt critical biological functions will rarely be observed in a sample of living individuals. Turning this argument around, we might expect that regions of the genome that are preferentially depleted for genetic variation might have important biological function.
The text file `influenza_HA_protein_sequences.fasta` contains and alignment of the amino-acid sequences from the HA gene in several different influenza strains.

Problem 8: The Kolmogorov backward equation

In class, we derived the Fokker-Planck equation (or ***forward equation***) for the probability density $p(f,t|f_0)$ of observing a mutation at frequency f at time t, given that it started at frequency f_0 at time $t=0$. In this problem, you will derive a related partial differential equation known as the ***backward equation***, which is particularly useful for calculating fixation probabilities and fixation times. For concreteness, we will consider the standard single locus diffusion process,

$$\frac{\partial f}{\partial t} = sf(1-f) + \sqrt{\frac{f(1-f)}{N}}\eta(t) \tag{12}$$

though the same derivation will apply to a large class of Markov models in the same diffusion limit.

(a) Start with the probability density $p(f,t|f_0)$ and consider what happens in the very first timestep $(0,dt)$. Write a recursion relation for $p(f,t|f_0)$ by integrating over the intermediate frequency f' at time dt.

(b) Taylor expand the integral equation to linear order in dt to obtain a partial differential equation for $p(f,t|f_0)$. This is known as the ***backward equation***.

(c) When $f=1$, the backward equation becomes a partial differential equation for the fixation probability, $p_{\text{fix}}(f_0,t)$. What boundary conditions must this function satisfy at $f_0=0$ and $f_0=1$?

(d) Using your answers in (b) and (c), solve for the long-term fixation probability, assuming that it approaches a constant value $p_{\text{fix}}(f_0)$ at long times.

(e) Repeat your derivation in parts (a)-(c) above to calculate the probability that a mutation reaches some other frequency $f_{\max}$ before it goes extinct, given that it starts at frequency f_0.

APPHYS 237 Problem Set 3

DUE: 2/18

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

Problem 1: Continuous-time branching process

Another classic 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. 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 $H(z,t) = \langle e^{-zn(t)} \rangle$, and the extinction probability $p_{\mathrm{ext}}(t)$

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 1, 2, or 0 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, one can average both sides and expand to lowest order in dt to arrive at a differential equation for $H(z,t)$.

(b) Solve your differential equation in part (a) subject to the initial condition $n(0) = 1$. Compare your results to diffusion model, $\partial_t f = s_e f + \sqrt{f/N_e}\eta(t)$, that we discussed in class:

$$H_n(z) \equiv \langle e^{-zNf} \rangle = \exp\left[\frac{z e^{s_e t}}{1 + \frac{zN}{2N_e s}\left(e^{s_e t} - 1\right)}\right] \tag{13}$$

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. Is there a timescale where you expect the convergence to break down?

(c) Use the same reasoning as in part (a) to derive differential equations for the mean $\langle n(t) \rangle$ and the non-extinction probability $p_{\mathrm{survival}}(t)$.

(d) Solve your differential equations in (a) and (b) subject to the initial condition $n(0) = 1$. Compare your results to diffusion model, $\partial_t f = s_e f + \sqrt{f/N_e}\eta(t)$, that we discussed in class:

$$\langle n(t) \rangle = e^{s_e t} \tag{14}$$

$$p_{\mathrm{survival}}(t) = \frac{2N_e s}{1 - e^{-st}}\left(\frac{1}{N}\right) \tag{15}$$

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.

Problem 2: Continuous-time branching process with bursty reproduction

In this problem, we'll consider a variant of the continuous time-branching process from Problem 1 of Problem Set 3, in which birth events now result in a "burst" of exactly K offspring.[12]

(a) Repeat your derivation in Problem 1 of Problem Set 3 to calculate the mean value $\langle n(t) \rangle$. If $D = 1$, solve for the value of B such that the long-term average growth rate is still $\langle n(t) \rangle = n(0)e^{st}$.

(b) Now calculate the long-term survival probability p_{survival} in the limit that $s \ll 1$ and compare this result to the ordinary branching process with net growth rate s? Does increasing the burst size make it more or less likely for a lineage to survive?

Problem 3: Mutation-selection-drift balance

The goal of this problem is to give you some practice using the method-of-characteristics approach we discussed in class.

(a) Calculate the generating function, $H(z, t)$, for mutation-selection-drift model,

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

subject to the initial condition $f(0) = 0$.

(b) Use your answer in (a) to show that for $t \ll 1/s$, the distribution of f is indistinguishable from a neutral mutation, even if $Ns \gg 1$.

(c) Assuming that the mutation is deleterious ($s < 0$), calculate the long-term mutation-selection-drift balance.

(d) Assuming that the mutation is beneficial, calculate the distribution of the establishment prefactor $\nu(t) = f(t)e^{-st}$ for $t \gg 1/s$, and compare this to the deterministic solution of . When is the deterministic solution a good approximation, and when does it break down?

Problem 4: Typical paths of non-extinct mutations

A convenient property of the linear branching process model $[\partial_t f = sf + \sqrt{f/N}\eta(t)]$ is that its generating function $H(z, t)$ has a simple exponential dependence on the initial condition f_0. As we will see below, this offers a convenient route for calculating multi-timepoint statistics without solving any additional differential equations.

[12] This mode of reproduction is relevant for some viruses, which often produce many multiple new viral particles per infected cell.

1. Calculate the multi-time generating function $H(z1, z2) \equiv \langle e^{-z_1 f(t_1) - z_2 f(t_2)}$ for the mutation frequency at timepoints t_1 and t_2, conditioned on an initial frequency f_0 at time $t = 0$. (**Hint:** first conditer the conditional generating function $\langle e^{-z_2 f(t_2)} | f_1 \rangle$, and then use the laws of conditional expectation to calculate $H(z_1, z_2)$.)

2. In class, we showed how the generating function can be used to calculate the typical frequency of a mutation at time t, conditioned on non-extinction. In a similar way, we can use the multi-time generating function to learn about the typical paths that mutations take to get to those typical frequencies. To do so, first calculate the marginal generating function $H^*(z_1)$, conditioned on non-extinction at time t_2.

3. Consider a beneficial mutation ($s > 0$) and two timepoints $t_1 = t \gg 1/s$ and $t_2 = 2s$. What are the typical frequencies at time t_1, conditioned on non-extinction at time t_2? How does this relate to the typical frequency conditioned on non-extinction at time t_1?

4. Repeat your calculation in (c) for a deleterious mutation ($s < 0$). How does this compare with the typical frequency conditioned on non-extinction at time t_1?

5. Now consider a neutral mutation ($s = 0$) and timepoints $t_1 = t \gg N f_0$ and $t_2 \gg t_1$. Do the typical paths still look like $f(t) \sim t/N$?

Problem 5: 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 diploid (or more general polyploid) organisms. In the course so far, we have primarily focused on evolution in haploid organisms (i.e., those with just a single copy of each chromosome). Organisms with more than one copy of each chromosome open up the possibility for ***recessive mutations***, i.e., mutations that must be present in all chromosomes within an individual before they can exert their 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 polyploid reproduction, in which individuals are formed by randomly choosing C chromosomes that exist in the current population. In the diffusion limit, the population frequency of a recessive mutation will satsify

$$\frac{\partial f}{\partial t} = s f^C (1 - f) + \sqrt{\frac{f(1-f)}{CN}} \eta(t) \,. \tag{17}$$

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,

$$\frac{\partial f}{\partial t} = s f^C + \sqrt{\frac{f}{CN}} \eta(t), , \tag{18}$$

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

(a) Repeat our heuristic derivation to partition frequency space into drift-dominated and selection-dominated regimes. For which values of N and s will natural selection be efficient?

(b) Use these results to calculate 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 6: Adaptive walks on uncorrelated fitness landscapes

An extreme limit of epistasis can be obtained by an uncorrelated fitness landscape, in which each genotype $\vec{g}$ in a genome of length L is assigned a random fitness $X(\vec{g})$ from a common distribution $p(X)$. We will consider an adaptive walk on such a landscape in the SSWM limit. For simplicity, we will take $p(X)$ to be an exponential distribution with mean σ, though our approach can be extended to other distributions as well.

(a) Starting a random genotype with fitness $X(\vec{g}) = X$, what is the probability that a given neighboring genotype has fitness benefit $s = X(\vec{g} + \hat{e}_i) - X(\vec{g}) > 0$? What is the distribution of s, conditioned on $s > 0$?

(b) Assuming that there are a large number of possible uphill steps, calculate the mean fitness trajectory $X(t)$ and the number of mutational steps $M(t)$ taken by the population as a function of time.

(c) On a finite genome, we will eventually come to a point where there are no uphill steps in the immediate neighborhood of the population. This constitutes a local optimum. There will be a high probability of a local optimum once the population fitness approaches a critical value, $L \int_{X_{\text{local}}}^{\infty} p(X)dx \sim 1$. Solve for X_{local} as a function of L. Use your results from (c) to calculate the average number of steps until reaching a local optimum.

(d) The logarithmic fitness trajectories produced by the uncorrelated fitness landscape are at least qualitatively similar to the fitness trajectories observed in Richard Lenski's long term experiment in *E. coli*[13]. However, the uncorrelated model makes very strong predictions about the fitness effects of mutations when they are transpanted to a different genetic background. In particular, for all mutations but the first, the fitness effect on the ancestral background is distributed as $p(s + X_0)$. For $X_0 \gg \sigma$, the vast majority of these mutations will be strongly deleterious.

The file...

Problem 7: LD of sweeping mutation

Problem 8: Drift barrier hypothesis for mutation rate evolution

The fact that the deterministic mutation-selection balance imposes a fitness load of order U_d suggests that it might be beneficial for a well-adapted organism to lower its mutation rate as much as possible. Consider an asexual population at mutation-selection balance with $s \gg U_d$, and suppose that we engineer a new mutation repair pathway in this organism that lowers its mutation rate to zero.

[13] Wiser *et al*, *Science* 2013; Good and Desai *Genetics* 2015

(a) If the engineered strain and the wildtype are introduced at a 50-50 ratio, what is the frequency trajectory of the engineered lineage over time? Based on this result, would you conclude that selection would a lower mutation rate in this organism?

(b) Suppose that the engineered strain takes over the population, but now loss-of-function mutations in the engineered pathway restore the wildtype mutation rate. Calculate the probability of observing the population in the optimized or non-optimized state if the mutation rates in both directions are of order $\mu \to 0$. How large must U_d before there is more than 90% chance of observing the population in the optimized state? This is an example of the ***drift barrier*** hypothesis, and it is one potential explanation for the fact that we don't see mutation rates that are closer to the limits imposed by physics.

(c) In practice, there are probably many more ways to break a biological pathway than there are to restore it once it is broken. We can formalize this intuition in a model where there are L potential sites that could result in a loss-of-function mutation, and these sites acquire forward and backward mutations at rate μ. Repeat your calculation in part (b) to calculate the probability of observing the population in the optimized vs one of the $2^L - 1$ non-optimized states. In the limit of large L, how low can U_d be before there is an appreciable probability of losing the optimized pathway?

Problem 9: Time to the most recent common ancestor of the entire population

Suppose that we have a sample of size n from a neutral population. Calculate the mean and variance of the time to the most recent common ancestor of the entire sample, assuming that n is large. What value do we get when n is the size of the entire population?

Problem 10: Measuring the DFE of single gene knockouts

In Problem 5 of Problem Set 1, you worked out the mathematics of the *pooled fitness assay.* These experiments are often performed in the context of large *knockout screens.* Several gene-editing methods now exist for creating a pools of mutant strains, in which each strain has a particular gene disrupted and replaced with a known sequence containing a random DNA barcode. By PCR ampifiying and sequencing just the barcode region, this approach provides an easy and cost-effective way to track the frequencies of thousands of gene deletion mutants together in a single experiment.

The text file `qian_etal_2012_deletion_fitnesses.txt` contains results from one such experiment in yeast.[14] A library of ∼4600 strains (each with a single gene deletion) was competed in rich media for 26 generations and sequenced at the initial and final timepoints. The entire process was then repeated again in a second biological replicate. The estimated fitnesses of each deletion strain (relative to the ancestor) are listed in the text file for each biological replicate. We'll write these numbers as

$$\begin{aligned} \hat{s}_{i,1} &= s_i + \epsilon_{i,1} \, , \\ \hat{s}_{i,2} &= s_i + \epsilon_{i,2} \, , \end{aligned} \tag{19}$$

where s_i is the "true" value and $\epsilon_{i,j}$ is a random error term with distribution $p(\epsilon)$. Without loss of

[14] Qian *et al* (2012), "The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast," *Cell Reports* **2**: 1399–1410.

generality, we can rewrite these as an average and a difference:

$$\begin{aligned}\overline{s}_i &\equiv \frac{\hat{s}_{i,1}+\hat{s}_{i,2}}{2}\,, \\ \Delta_i &\equiv s_{i,2}-s_{i,1}\,.\end{aligned} \tag{20}$$

(a) Suppose that the error distribution is symmetric around zero $[p(\epsilon) = p(-\epsilon)]$. Derive a relationship between the distribution of Δ_i and the residual error around the average, $\overline{\epsilon}_i \equiv \overline{s}_i - s_i$.

(b) Using your result in (a), plot the number of genes you expect to see with $|\overline{s}_i| \geq s$ if all the gene deletions were neutral ($s_i = 0$). Compare this to the observed number of genes with $|s_i| \geq s$. What fraction of gene deletions have significant fitness effects? and what are their typical fitness effects?

(c) Repeat part (b), this time focusing only on beneficial mutations ($\overline{s}_i \geq s$). What fraction of gene deletions are beneficial in this environment? What are their typical fitness effects?

(d) In Problem 1 of Problem Set 2, we estimated the fraction of spontaneous mutations that disrupt the function of a gene. If we assume that all beneficial mutations that occur in laboratory evolution experiments are effectively loss-of-function mutations, use your answer from Problem 1 of Problem Set 2, along with your results in (c), to estimate the distribution of fitness effects of spontaneous beneficial mutations for yeast grown in this environment:

$$U\rho(s)ds \equiv \text{per generation rate of producing a mutation with fitness effect } s \pm ds \tag{21}$$

We will consider a more direct way of measuring the DFE in a later problem.

APPHYS 237 Problem Set 4

DUE: 3/10/18

Problem 1: Measuring the DFE for de novo beneficial mutations

A common criticism of DFE estimates obtained from deletion screens (e.g. Problem 10 of Problem Set 3) 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[15] recently devised a clever method to do this in a high throughput way, using a variation of the standard pooled fitness assay setup.

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. Although the basic idea is simple, actually 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 the analysis below.

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

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

(a) The total number of generations that the lineages are monitored over, T.
(b) The total number of cells in the population at the serial transfer step, N_b.
(c) The total number of barcoded lineages, B.
(d) The total number of 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) A substantial number of the barcoded lineages (e.g., ~ 1000) acquire a beneficial mutation during the experiment.
(b) Beneficial mutations noticeably perturb the frequency of the lineage that they occur in (so that we can actually observe them).
(c) Genetic drift does not substantially perturb the frequency of the lineages on the same timescale (i.e., if we see a large 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[16] 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?

The file `levy_blundell_etal_2015_barcode_trajectories.txt` contains the raw read count trajectories obtained from one such experiment. Half a million barcoded lineages were serially transferred in glucose limited media for 14 days, with bottleneck size of a 256-fold dilution rate ($\Delta t = 8$ generations/day) and a bottleneck size of $N_b \approx 7 \times 10^8$. We'll denote the read count trajectory for an arbitrary barcode i by $R_{i,t}$, and we'll let $D_t = \sum_i R_i$ denote the total sequencing coverage in each timepoint. This defines a corresponding set of read count frequencies

$$\hat{f}_{i,\tau} \equiv \frac{R_{i,\tau}}{D_\tau} \,. \tag{22}$$

Noise in these read count trajectories reflects both the stochastic growth dynamics of the experiment, as well as noise in the data generation process (PCR amplification and sequencing). Levy, Blundell, *et al* argued that this compound process is well approximated by an effective branching process model that connects the read count frequencies at successive sequenced timepoints. In particular, given that we observe a lineage at frequency $\hat{f}_{i,\tau}$, the conditional probability at the next timepoint, $p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau})$, can be approximated by a branching-process-like generating function:

$$H(z|\hat{f}_{i,\tau}) \equiv \int e^{-zf} p(f|\hat{f}_{i,\tau})\, df \approx \exp\left[-\frac{z\hat{f}_{i,\tau}[1 + (X_{i,\tau} - \overline{X}_\tau)\Delta t_\tau]}{1 + z\kappa_\tau/D_\tau}\right] , \tag{23}$$

where Δt_τ is the number of generations between the timepoints, $X_{i,\tau}$ is the fitness of lineage i at timepoint τ, $\overline{X}_\tau$ is the mean fitness of the population at that timepoint ($\overline{X}_\tau \approx \sum_i X_{i,\tau}\hat{f}_{i,\tau}$), and

[16] Desai et al *Current Biology* 2007

κ_τ is an effective parameter capturing the net effects of genetic drift and measurement noise. As we noted in class, the inverse of this generating function has a convenient asymptotic expansion,

$$p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau}) \sim \frac{\left[(1+(X_{i,\tau}-\overline{X}_\tau)\Delta t_\tau)\hat{f}_{i,\tau}\right]^{1/4}}{(4\pi\kappa_\tau)^{1/2} f_{i,\tau+1}^{3/4}} \exp\left[-\frac{\left(\sqrt{\hat{f}_{i,\tau+1}}-\sqrt{(1+(X_{i,\tau}-\overline{X}_\tau)\Delta t_\tau)\hat{f}_{i,\tau}}\right)^2}{\kappa_\tau}\right] \tag{24}$$

which is valid for large $R_{i,\tau+1}$.

(b) We'll first use the measured data to verify that Eq. 23 is a good approximation. Consider the first timepoint ($\tau = 0$), where few of the lineages will have any beneficial mutations, and we can assume that $X_{i,\tau} \approx \overline{X}_\tau \approx 0$. Then consider the set of all lineages with exactly 50 reads in the first timepoint – by construction, these should all have the same conditional distribution, $p(\hat{f}_{i,1}|\hat{f}_{i,0})$. Use the observed frequencies of these lineages at the next timepoint ($\hat{f}_{i,1}$) to show that the conditional distribution is consistent with the approximation in Eq. 23.

Hint: consider the empirical generating function, $\hat{H}(z) = \frac{1}{n}\sum_i \exp\left(-z\hat{f}_{i,1}\right)$, evaluated for z near "typical" values of $1/\hat{f}_{i,1}$. (Can you explain why this should be a robust moment to estimate for a positive random variable in a finite sample?) Rearrange Eq. 23 as a linear function of $1/z$, so that you can use linear regression[17] to estimate the slope and intercept.

(c) If we continue to focus on rare mutations (e.g,, $20 \leq R_{i,\tau} \leq 40$), then the vast majority should remain neutral even for $\tau > 0$. We can therefore use the statistics of these neutral lineages to estimate κ_τ and $\overline{X}_\tau$ using the same approach you outlined in (b). Specifically, estimate a separate value of κ_τ and $\overline{X}_\tau$ for lineages with $R_{i,\tau} = 20, \ldots, 60$, and average them together to obtain a single estimate of κ_τ and $\overline{X}_\tau$ for each timepoint. Plot your estimated values as a function of time. What is the estimated fold change in frequency of a neutral lineage over the course of the experiment?

(d) We can now use the fitted values of κ_τ and $\overline{X}_\tau$ (measured for the bulk population) to scan for outlier lineages that acquired a beneficial mutation. To do so, let's imagine that a beneficial mutation with effect s occured in lineage i some timepoint $t < t_\tau$. The lineage frequency at later timepoints can then be split into neutral and beneficial components,

$$\hat{f}_{i,\tau} = \hat{f}^0_{i,\tau} + \hat{f}^s_{i,\tau}\,, \tag{25}$$

where $\hat{f}^0_{i,\tau}$ and $\hat{f}^s_{i,\tau}$ are both described by Eq. 23 with $X_{i,\tau} = 0$ and $X_{i,\tau} = s$, respectively. Derive an exprssion for the generating function of $\hat{f}_{i,\tau+1}$, conditioned on the values of $\hat{f}_{i,\tau}$, $\hat{f}^0_{i,\tau}$, and $\hat{f}^s_{i,\tau}$. What is the effective lineage fitness $X_{i,\tau}$?

Unfortunately, we don't observe the sublineages $\hat{f}^0_{i,\tau}$ and $\hat{f}^s_{i,\tau}$ directly, so we'll have to estimate them from the observed values of $\hat{f}_{i,\tau}$. If the beneficial mutation establishes at time t_0, its frequency at later timepoints will be given by

$$f(t|s,t_0) \approx \frac{c}{N_b s} e^{\int_{t_0}^{t}(s-\overline{X}(t'))dt'}\,, \tag{26}$$

[17] E.g., using the `linregress` function in the SciPy `stats` package.

where c is an $\mathcal{O}(1)$ constant that depends on the variance in offspring number in the experiment ($c \approx 1.8$ here, see SI p. 11 in Levy, Blundell, *et al* 2012). We can therefore approximate

$$\hat{f}^s_{i,\tau} \approx \begin{cases} 0 & \text{if } t_\tau < t_0 \\ f(t_\tau|s,t_0)/\hat{f}_{i,\tau} & \text{if } f(t_\tau|s,t_0) < \hat{f}_{i,\tau}, \\ 1 & \text{else.} \end{cases} \tag{27}$$

This completely specifies the model. The probability of observing a given lineage trajectory, conditioned on s and τ, is given by

$$p(\{\hat{f}_{i,\tau}\}|s,t_0) \approx p(\hat{f}_{i,0}) \prod_\tau p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau},s,t_0)\,. \tag{28}$$

(e) Parameter estimation can be done with a standard Bayesian approach. Write a formal expression for the posterior probability, $p(s,\tau|\{f_{i,\tau+1}\})$, relative to the posterior probability without a beneficial mutation ($t_0 = \infty$). You may leave your answer as a function of $p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau},s,t_0)$ and the prior probabilities $p_0(s,\tau)$. This ratio is known as the *posterior odds ratio.*

Numerically calculate the posterior odds ratio for trajectory 14 in the data file. For simplicity, we'll discretize (s,t_0) values into a grid with spacing $\delta t_0 = 1$ and $\delta s = 0.005$, and we'll assume a flat prior

$$\frac{p_0(s,t_0)}{p_0(t_0=\infty)} \approx \begin{cases} cf_0 N_b U_b^0 s \cdot \delta s \cdot \delta t_0 & \text{for } 0 \leq s \leq 0.4 \text{ and } -250 \leq \tau < 100 \\ 0 & \text{else,} \end{cases} \tag{29}$$

where f_0 is the typical frequency of a lineage in the initial pool, and $U_b^0 \sim 10^{-5}$. For which values of s and τ is the posterior odds ratio the highest? Does this make sense given the shape of the trajectory?

(f) Now use your approach in (e) to estimate (s,t_0) values for all of the trajectories in the experiment. Set $t_0 = \infty$ if the posterior odds ratio is less than one; otherwise take the values of (s,τ) that maximize the posterior odds ratio. How many beneficial mutations do you detect?

(g) Finally, we can use your detected beneficial mutations to estimate the distribution of fitness effects, $U_b\rho(s)$. The number of beneficial mutations in an interval $s \pm \delta s$ that establish and rise to detectable frequencies is given by

$$n(s) \approx \left[N_b \int_0^{t^*(s)} e^{-\overline{X}(t)}\,dt \right] \cdot U_b\rho(s)\delta s \cdot \frac{s}{c} \tag{30}$$

where t^* is the latest the mutation could establish and still perturb the frequency of the lineage. Write an approximate expression for $t^*(s)$, and then rearrange Eq. 30 to write $U_b\rho(s)\delta s$ as a function of the observed values $n(s)$. Plot your estimated DFE using the beneficial mutations you detected in (f).

Problem 2: Genealogies from sequences of neutral mutations

In class, we saw how we can use coalescent theory to go from genealogies to sequences of neutral mutations. In this problem, we will consider how to go in the opposite direction. Suppose we draw a sample of $n = 6$ individuals from a population and observe mutations at one or more sites. We'll consider a few different imaginary scenarios with $S = 1$, 2, and 3 variable sites.

```
(a) A     (b) AG     (c) AG     (d) AG     (e) AGTG
    A         TC         AG         AG         AGCG
    A         AG         AC         AC         ACCG
    T         TC         TC         TC         TCCA
    T         AG         TC         TC         TCCG
    T         TC         TC         TG         TCCA
```

(a) Draw two genealogies that are consistent with the mutation pattern in (a), assuming that each mutation happens only once ($\mu T_c \ll 1$).

(b) Repeat for pattern (b) above.

(c) Repeat for pattern (c) above.

(d) Try to repeat for pattern (d). Is it possible to draw a consistent genealogy where each mutation happens only once? How is (d) different from (c) and (b), in terms of the number of distinct haplotypes that are observed? (A version of this idea, known as the ***four gamete test*** is frequently used to diagnose recombination or recurrent mutation events in DNA sequence data.)

(e) Draw a genealogy that is consistent with the mutation pattern in (e).

Problem 3: 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.[18] 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.

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

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

$$\mathrm{MI}(a_i, a_j) = -\sum_{a,a'} \Pr(a_i = a,\, a_j = a') \log \left[\frac{\Pr(a_i = a,\, a_j = a')}{\Pr(a_i = a)\Pr(a_j = a')} \right] \tag{31}$$

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 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 ∼1000 gene families over the total number of generations that have elapsed since the origin of life (∼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?

Problem 4: The effective strength of genetic drift under background selection

In this problem, we will derive an effective model for the frequency trajectory of a neutral mutation in a genome with a large number of strongly deleterious sites. Suppose that we have an asexual population at mutation-selection balance ($Nse^{-U_d/s} \gg 1$), and a neutral mutation arises in one of the individuals in the population.

(a) What is the probability that the mutation arises in a genetic background with k deleterious mutations?

The neutral mutation will found a new lineage, which will initially be described by a coupled branching process,

$$\frac{\partial f_k}{\partial t} = [-sk - (-U_d)]f_k + U_d f_{k-1} - U_d f_k + \sqrt{\frac{f_k}{N}}\eta_k(t) \tag{32}$$

where f_k is the frequency of individuals that have the focal neutral mutation and k deleterious mutations, and the total frequency of the neutral mutation is $f(t) = \sum_k f_k(t)$. The stochastic dynamics of this process are rather complex[19], but significant insights can be gained by dropping the stochastic term and considering the deterministic limit.

[19] For more information, see Cvijovic, Good, and Desai (Genetics, 2019).

(b) Solve for the time-dependent values of $f_k(t)$, subject to the initial condition $f_k(0) = f_0\delta_{k,k_0}$. **Hint:** this is easiest to do by calculating the generating function $G(z,t) = \sum_k e^{-zk} f_k(t)$. Can you recognize the result as a known distribution, except with a shift and an overall prefactor?

(c) Use your result in part (b) to calculate the total mutation frequency, $f(t) = \sum_k f_k(t)$. [**Hint:** this is very easy to do with the generating function, since $G(0,t) = \sum_k f_k(t)$.] What do the trajectories look like for $k = 0$ and $k_0 > 1$, respectively? How quickly are the $k_0 > 1$ lineages purged from the population? What does the f_k distribution look like at long times when $k_0 = 0$?

(d) We can gain some insight into the effects of stochasticity with the following crude[20] argument. Suppose the neutral mutation arose in the $k = 0$ class, and is currently residing at a total frequency f in the population. From your answer in part (c), we expect the frequencies f_k to look mutation-selection balance with an overall prefactor f. Now suppose that we suddenly perturb each class by some small amount δf_k. Based on your answer to parts (b) and (c), how do the long-term values of $f_k(t)$ and $f(t)$ depend on the δf_k, and what is the characteristic timescale τ_{relax} over which these long-term values are attained?

(e) Your results in (d) suggest that, if we coarse-grain over timescales longer than τ_{relax}, then the population frequency $f(t)$ should look like a mirrored version of the k_0 class, which is smaller by a factor of $e^{-U_d/s}$. Use this result to write down a coarse-grained diffusion model for $f(t)$ that is valid when $f(t) \ll 1$. What is the effective population size? Is there a critical frequency below which we expect the coarse-grained model to break down?

Problem 5: Measuring recombination rates using the decay of LD

Problem 6: Sexual vs asexual selection on a highly polygenic trait

Suppose that we create a population by crossing two diverged strains of yeast, and we evolve the resulting hybrid offspring in an environment that selects for higher values of a particular trait. We'll assume that the fitness components of this phenotype are controlled by a large number L of mutational differences between the two strains, each contributing a small fitness effect $\pm s$. For simplicity, we'll assume that the positive and negative mutations are evenly distributed between the two parents, and that the recombination rate is sufficinetly high that the different mutations are assigned to offspring independently. Under these assumptions, the variance in fitness of the offspring are normally distributed with mean 0 and variance $V = Ls^2/4$. The goal of this problem is to consider what happens in the so-called ***infinitesimal limit***, where we let $L \to \infty$ and $s \to 0$ while keeping the variance $V = Ls^2/4$ constant. (Formally, we can achieve this by setting $s = \sqrt{V/L}$ and thinking about an asymptotic expansion for large L.)

(a) Let's first consider the case where we evolve the hybrid offspring asexually. For simplicity, we'll neglect the possibility of additional mutations in the offspring, so that we essentially have a pooled fitness assay similar to Problem 5 of Problem Set 1. What is the initial rate of fitness increase of the population ($\partial_t \overline{X}$)? What is the maximum fitness X^* that will typically exist in a population founded by N_0 hybrid offspring? (i.e., the maximum fitness that can be achieved before we have to wait for additional mutations?)

(b) Now let's imagine that the evolution step is performed with continual rounds of sexual reproduction, with a sufficiently high rate of recombination that the fitness-influencing sites

[20] See Cvijovic et al (2019) for a more rigorous derivation.

are effectively unlinked ($r_{ij} \gg \sigma$). How does the mean fitness of the population grow in this scenario? How long do we have to wait before the population reaches the maximum value X^* that existed in the initial pool? How much do the frequencies of mutations change over this timescale?

(c) Continuing with the scenario in (b), how long would we have to wait for the rate of adaptation to decrease significantly from its initial value? How much fitness has been gained by this time? How does this compare (at an order-of-magnitude level) to the maximum fitness that could be created by reshuffling mutations in a pool of N_0 hybrid offspring?