APPHYS 237 / BIO 251, Problem Set 2

DUE: 4/28/24 at the beginning of class.

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

(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 expected 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 connect the phenotypic mutation rate $\hat{U}_{\Delta URA3}$ to a per-base-pair mutation rate, Lang and Murray sequenced 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 URA3, and presumably

[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.

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 804bp, so there are a total of 2412 possible single nucleotide variants that could be produced. Based on the wildtype sequence of URA3, Lang and Murray calculated that 123 of these potential variants are nonsense mutations (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: 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.[4] 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 in the flask. 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? Why or why not?

[4] 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 3: Neutral mutation accumulation in individuals vs populations

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

(a) Suppose that you know the population frequencies of the mutations at each site ℓ in the genome ($\ell = 1, \ldots, L$), which we'll denote by $f_\ell(t)$. Write a formula for the average number of mutations in a randomly sampled individual from the population as a function of $f_\ell(t)$. Call this number $M_1(t)$.

(b) Write an analogous formula for the average number of mutations that are *shared* by a randomly sampled pair of individuals in the population. Call this number $M_2(t)$. What about a random sample of n individuals?

(c) Write a stochastic differential equation for $f_\ell(t)$ for an arbitrary site ℓ, assuming that it evolves neutrally. For this problem, you may neglect any correlations between $f_\ell(t)$ at different sites. Use your stochastic differential equation to derive a *deterministic* differential equation for the average frequency, $\langle f_\ell(t) \rangle$. Solve this equation and show how $M_1(t)$ grows with time.

(d) 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 4: Measuring the fitness effects of all single gene knockouts

In Problem 6 of Problem Set 1, we worked out the mathematics of the *pooled fitness assay*. These experiments are often performed in the context of large *deletion screens*. Several gene-editing methods now exist for creating large pools of mutant strains (or *libraries*), in which each strain has a particular gene disrupted and replaced with a known sequence containing a random DNA barcode. In this case, the mutant strains are typically referred to as *gene deletions* or *knockouts*. By PCR amplifiying and sequencing the barcode region,[5] one can easily and cost-effectively track the frequencies of thousands of gene deletion mutants together in a single experiment – and therefore estimate their fitness effects. The goal of this problem is to give you a feel for what these numbers look like.

The text file `qian_etal_2012_deletion_fitnesses.txt` contains results from one such knockout experiment peformed in yeast.[6] In this experiment, a library of ∼4600 strains (each with a single gene deletion) was propagated 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) for each of the two replicates are listed in the text file. Each of these measurements will involve some amount of measurement error, so we can write the observed values as

$$\begin{aligned} \hat{s}_{i,1} &= s_i + \epsilon_{i,1} \,, \\ \hat{s}_{i,2} &= s_i + \epsilon_{i,2} \,, \end{aligned} \tag{12}$$

[5] We will officially introduce these techniques in more detail during the sequencing lectures. For now, you can just think of this as a way to count frequencies of many different types at relatively high resolution.

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

where s_i is the "true" fitness of gene deletion i and $\epsilon_{i,r}$ is a random error term with mean $\langle \epsilon \rangle = 0$ and distribution $p(\epsilon)$. Without loss of generality, we can rewrite this pair of numbers as an average

$$\overline{s}_i \equiv \frac{\hat{s}_{i,1} + \hat{s}_{i,2}}{2}\,, \tag{13}$$

and a difference

$$\Delta_i \equiv \hat{s}_{i,2} - \hat{s}_{i,1}\,. \tag{14}$$

which satisfy $\langle \overline{s}_i \rangle = s_i$ and $\langle \Delta_i \rangle = 0$. As above, the measurement errors in $\hat{s}_{i,1}$ and $\hat{s}_{i,2}$ will cause the sample average $\overline{s}_i$ to fluctuate around its true value s_i. Estimating the fitness effects of different gene deletions will therefore require us to distinguish these values from measurement noise. We will work through a crude approach for doing this below.

(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, defined by $\overline{\epsilon}_i \equiv \overline{s}_i - s_i$. Use this result to estimate the empirical distribution of $\overline{\epsilon}_i$ from the data, assuming that this distribution is the same across all genes.

(b) Using your result in (a), plot the number of genes you would expect to see with $|\overline{s}_i| \geq s$ if all the gene deletions were actually neutral ($s_i = 0$). Compare this prediction 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 (DFE) of spontaneous beneficial mutations for yeast grown in this environment:

$$U\rho(s)ds \equiv \text{per generation rate of producing a mutation with fitness effect } s \pm ds \tag{15}$$

We will consider a more direct way of measuring the DFE in Problem Set 3.

(e) The limited resolution in this experiment could come from one of two sources: (i) evolutionary noise due to genetic drift or (ii) frequency estimation noise (e.g. due to finite sequencing coverage). Assuming perfect frequency estimation, estimate the effective population size N_e required to produce a frequency change as large as the one produced by the minimum resolvable fitness effect, $s_{\rm err} \sim \langle |\overline{\epsilon}_i| \rangle$. You may assume that all deletion strains start at the same frequency. Similarly, in the absence of genetic drift ($N_e = \infty$) estimate the frequency resolution δf required to show that the fitness effect of a truly neutral deletion is $\leq s_{\rm err}$.