APPHYS 237 / BIO 251, Problem Set 4

DUE: 6/04/24

Problem 1: Measuring the DFE for de novo beneficial mutations, Part II

This problem is a continuation of the barcoded lineage tracking problem from last week's homework, now with some applications to real data.

The file `levy_blundell_etal_2015_barcode_trajectories.txt` contains the raw read count trajectories obtained from one such experiment in yeast.[14] In this 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 $N_b \approx 7 \times 10^7$. 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

$$\hat{f}_{i,\tau} \equiv \frac{R_{i,\tau}}{D_\tau} \, . \tag{19}$$

(a) First, let's familiarize ourselves with the data. We have been thinking about theoretical frequency trajectories all quarter, so now is our change to see the real thing! Plot the frequency trajectories over time of 10 randomly selected barcodes (lineages) from the dataset. Then, choose the 10 highest-frequency barcodes at the last timepoint ($\tau = 112$) and plot their frequency trajectories over all timepoints. What do you notice about these two groups of barcodes? Do any of the randomly selected barcodes have trajectories similar to those of the high frequency set?

Since all of the barcodes start at low frequency, then the initial frequency dynamics of each barcoded lineage can be described by a variant of the branching process model we studied in class:

$$\frac{\partial f_i}{\partial t} = \underbrace{(X_i(t) - \overline{X}(t))f_i}_{\text{natural selection}} + \underbrace{\sqrt{\frac{f_i}{N_e}}\eta_i(t)}_{\text{genetic drift}} \, . \tag{20}$$

The main difference is that the effective fitness of a barcode, $X_i(t)$, is now a *time-dependent* quantity because each barcode could be composed of one or more sub-lineages: (i) the initial barcoded cells (with relative fitness 0), and (ii) additional beneficial mutations that arise in the experiment over time. The main goal of the original paper was to exploit this link to learn something about the adaptive landscape that the yeast experience *in vitro* – you'll repeat some of the key steps of their analysis in Problem 1 and Problem 2 of this week's homework.

(b) First, let's look at things by eye: if you examine the trajectories of the high frequency barcodes you plotted in part (a), can you guess where mutations started to accumulate in each of these lineages?

If we want to do this more rigorously, there are two main difficulties we have to deal with. The first is that the mean fitness of the population, $\overline{X}(t)$, will increase over time, as beneficial mutations

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

start to sweep through the population. The second is that we don't measure the frequency of lineage i directly, but only the noisy version in Eq. 19 that we observe from sequencing. Noise in these read count trajectories will reflect both the stochastic growth dynamics of the experiment (encapsulated by the N_e term in Eq. 20), as well as noise in the data generation process (DNA extraction, PCR amplification and sequencing).

Levy, Blundell, *et al* argued that this compound process is well approximated by a *second* 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 distribution of the frequency at the next timepoint, $p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau})$, can be approximated by a branching-process-like generating function:

$$H(z|\hat{f}_{i,\tau}) \equiv \int e^{-zf} p(f|\hat{f}_{i,\tau})\, df \approx \exp\left[-\frac{z\hat{f}_{i,\tau}[1+(X_{i,\tau}-\overline{X}_\tau)\Delta t_\tau]}{1+z\kappa_\tau/D_\tau}\right] , \tag{21}$$

where Δt_τ is the number of generations between the timepoints, $X_{i,\tau}$ is the instantaneous fitness of lineage i at timepoint τ, $\overline{X}_\tau$ is the instantaneous mean fitness of the population at that timepoint ($\overline{X}_\tau \approx \sum_i X_{i,\tau}\hat{f}_{i,\tau}$), and κ_τ is an effective noise parameter that captures the net effects of genetic drift *and* measurement noise. As we saw in class, this function is difficult to invert *exactly* to get the probability distribution $p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau})$. But for large $R_{i,\tau+1}$, it can be approximated by the asymptotic expansion,

$$p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau}) \sim \frac{\left[(1+(X_{i,\tau}-\overline{X}_\tau)\Delta t_\tau)\hat{f}_{i,\tau}\right]^{1/4}}{(4\pi\kappa_\tau/D_\tau)^{1/2} f_{i,\tau+1}^{3/4}} \exp\left[-\frac{\left(\sqrt{\hat{f}_{i,\tau+1}}-\sqrt{(1+(X_{i,\tau}-\overline{X}_\tau)\Delta t_\tau)\hat{f}_{i,\tau}}\right)^2}{\kappa_\tau/D_\tau}\right] \tag{22}$$

Both representations of this conditional probability distribution $p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau})$ will be useful at different stages of the problem below.

Our goal in the rest of Problem 1 is to infer the values of the mean fitness $\overline{X}_\tau$ and noise parameter κ_τ. There are two ways to approach this, and you only need to **choose one. Path 1** uses moment generating functions (parts (c) and (d)), or **Path 2**, which uses empirical read count distributions (parts (e) and (f)).

Pick **one** of the following paths:

Path 1: Moment generating functions (note, if you do parts (e) and (f), you may skip this part)

(c) We'll first use the measured data to verify that Eq. 21 is a good approximation. Consider the first timepoint ($\tau = 0$), where few of the lineages will have any beneficial mutations. This means that 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 Eqs. 21 and 22.

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. 21 as a linear function of $1/z$, so that you can use linear regression[15] to estimate the slope and intercept.

(d) If we continue to focus on rare lineages (e.g,, $20 \leq R_{i,\tau} \leq 60$), 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 (c). 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?

Path 2: Empirical read count distributions (note, if you already did parts (c) and (d), you may skip this part)

One method to infer mean fitness is to monitor the frequency decline of neutral lineages. Likewise, noise and fluctuations in these neutral frequency trajectories will capture the effects from growth-bottleneck cycles, amplification, and sequencing. We can use low abundance lineages as neutral markers, because the vast majority of lineages that are present at only 20-50 cells at the bottleneck will not accumulate a beneficial mutation before roughly 1000 generations.

(e) First, for each time point τ, form a list of all barcodes with exactly R_τ reads at time point τ, where $20 \leq R_\tau \leq 40$ reads. Calculate $p(R_{\tau+\Delta t}|R_\tau)$, the distribution of these reads at the subsequent time point $\tau+\Delta t$. We can compare this empirical distribution with our prediction for the number of reads at time point $\tau + \Delta t$, using the probability distribution in Eq. 22. We will infer the best fit pair $(\overline{X}_\tau, \kappa_\tau)$ by minimizing the distance between the empirical distribution and the predicted distribution of read counts at time $\tau + \Delta t$, which is defined to be the summed square of differences between the predicted distribution and the measured distribution:

$$\text{distance} = \sum_j^{2R_\tau} (\text{Measured number of barcodes at } R_j - \text{Predicted number of barcodes at } R_j)^2 \tag{23}$$

For each initial R_τ, each time point will yield a best-fit pair $(\overline{X}_\tau, \kappa_\tau)$. Estimate this best-fit pair for each $R_\tau \in [20, 40]$.

(f) For each time point, average your best fit $(\overline{X}_\tau, \kappa_\tau)$ over all initial read counts to obtain a single estimate of κ_τ and $\overline{X}_\tau$. 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? Do you obtain the same results as the inset in Figure 2a of the original paper (Levy, Blundell, *et al* 2012)?

Problem 2: Measuring the DFE for de novo beneficial mutations, Part III

We can now use the fitted values of κ_τ and $\overline{X}_\tau$ from Problem 1 to scan for the smaller set of outlier lineages that acquired a beneficial mutation. To do so, let's imagine that a beneficial mutation with

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

effect s arose and established in lineage i some time t_0. The frequency at later times is therefore given by

$$f(t|s,t_0) \approx \frac{c}{N_b s} e^{\int_{t_0}^{t}(s-\overline{X}(t'))dt'}\,, \tag{24}$$

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 the effective fitness of the entire lineage as

$$X_{i,\tau} \approx s \cdot \min\left\{ \frac{f(t_\tau|s,t_0)}{\hat{f}_{i,\tau}},\, 1 \right\} \tag{25}$$

which interpolates between $X_i(t) = 0$ at small times and $X_i(t) = s$ at long times when the beneficial mutation has fixed within the lineage.

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

$$p(\{\hat{f}_{i,\tau}\}|s,t_0) \approx p(\hat{f}_{i,0}) \prod_{\tau} p(\hat{f}_{i,\tau+1}|\hat{f}_{i,\tau},s,t_0)\,, \tag{26}$$

where the conditional probability is given by Eq. 22, with an effective fitness given by Eq. 25. We'll now use this model to infer the best fit values of s and t_0 for each lineage given the observed values of $\hat{f}_{i,\tau}$

(g) Parameter estimation can be done with a standard Bayesian approach. Write a formal expression for the posterior probability, $p(s,t_0|\{f_{i,\tau+1}\})$, 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 $p_0(s,t_0)$. 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

$$\frac{p_0(s,t_0)}{p_0(t_0=\infty)} \approx \begin{cases} c f_0 N_b U_b^0 s \cdot \delta s \cdot \delta t_0 & \text{for } 0 \leq s \leq 0.4 \text{ and } -250 \leq t_0 < 100 \\ 0 & \text{else,} \end{cases} \tag{27}$$

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 t_0 is the posterior odds ratio the highest? Does this make sense given the shape of the trajectory?

(h) Now use your approach in (d) to estimate (s,t_0) values for the first 1000 trajectories in the experiment. Set $t_0 = \infty$ if the posterior odds ratio is less than one; otherwise take the values of (s,t_0) that maximize the posterior odds ratio. How many beneficial mutations do you detect? Extrapolating the run time from this pilot data, estimate how long it would take your program to process all ∼500,000 trajectories in the experiment?

Bonus: estimate (s,t_0) values for all ∼500,000 trajectories in the experiment.

(i) 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

$$n(s) \approx \left[N_b \int_0^{t^*(s)} e^{-\overline{X}(t)}\, dt \right] \cdot U_b\rho(s)\delta s \cdot \frac{s}{c} \tag{28}$$

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. 28 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 3: 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)	(b)	(c)	(d)	(e)
A	AG	AG	AG	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 4: 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 $\pm s/2$. For simplicity, we'll assume that the positive and negative mutations are evenly distributed between the two parents, and that the recombination rate is sufficiently 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 6 of Problem Set 1. What is the initial rate of fitness increase of the population ($\partial_t \overline{X}$)?

(b) If the asexual population was founded by a large but finite number of hybrid offspring, N_0, there will be a maximum possible fitness $x_{\max}$ within the initial hybrid pool. Using extreme value theory, we can show that the expected value of $x_{\max}$ is given by

$$x_{\max} \approx \sqrt{2\sigma^2 \log N_0} \tag{29}$$

However, some of these individuals will drift to extinction while rare, while others will establish and start to grow to higher frequencies. One can correct for this in our extreme value theory calculation, leading to the maximum established fitness,

$$x_{\max}^{\text{est}} \approx \sqrt{2\sigma^2 \log N_0 \sigma} \tag{30}$$

How long does it take for the mean fitness of the population to reach fitness $x_{\max}^{\text{est}}$ if it grew at the initial rate of adaptation the whole time. Can you speculate what happens to the rate of adaptation after this point (in words)?

(c) 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 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 asexual fitness from Eq. 30? How much do the frequencies of mutations change over this timescale? Based on your answers, what would you conclude about the efficiency of adaptation in sexual vs asexual populations?