Bi/Ge105: Evolution
Homework 6
Due Date: Wednesday, March 4, 2020

"We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes."

–Pierre Simon Laplace, A Philosophical Essay on Probabilities

1. The Tools of Population Genetics: A One-Locus, Two-Allele Model of Selection.

We took several lectures to introduce the rudiments of population genetics through the device of one-locus, two-allele models in which we imagine the evolutionary process as collapsing onto nothing more than the dynamics of only two alleles. In this problem, we are going to reconsider the key elements of how allele frequencies change with selection in the context of such models. We are going to focus on a diploid organism and we will use the notation f to characterize the frequency of the A_1 allele in the population. Within the confines of this model, the mean fitness is defined as

$$\bar{w} = f^2 w_{11} + 2f(1-f)w_{12} + (1-f)^2 w_{22}. \tag{1}$$

For the remainder of the problem, let's consider the case of heterozygous advantage and choose our fitnesses symmetrically as $w_{11} = w_{22} = 0.1$ and $w_{12} = 1.0$.

(A) In class we derived an expression for the change in allele frequency during one generation of the form

$$\Delta f = \frac{1}{\bar{w}}(f^2 w_{11} + f(1-f)w_{12}). \tag{2}$$

Provide a justification of this result by appealing to our simple picture of urns and mathematicizing that picture. With this result in hand, plot Δf vs

f given the fitness model we outlined above. Explain the trends in this phase portrait. What does this diagram tell us about the long time dynamics of this simple system?

(B) Plot the allele frequency as a function of generation number (i.e. $f_{n+1} = f_n + \Delta f$) given that the population starts out with an initial condition $f = 0.001$. Do the same for the initial condition $f = 0.99$.

(C) Plot the evolution of the mean fitness $\bar{w}$ over the generations. Give a qualitative explanation of the trends and what it implies about selection and the evolutionary process.

2. Timescale to evolve a promoter

In lab this week we began the process of thinking about how bacterial promoters evolve, with our inspiration coming from a recent paper from Jeff Gore's lab at MIT. As you are completing this problem set, your strains are actively evolving in the lab, and this problem serves as opportunity to think of how long this process might take. As a reminder of the experimental setup, Figure 1 shows the scheme by which we are evolving the strains. In the original experiments, they saw evolution of promoters within 7 to 90 days. In this problem will we think about this timescale, and whether it seems "fast" or "slow."

Our first task is to think about how many bacterial divisions (i.e. chances to accumulate a beneficial mutation in the promoter region) are occurring in our cultures each day.

(A) Given that we dilute by a factor of 100, how many divisions occur in our cultures each day? Other pieces of information you might want to know are that we are growing 3 mL cultures and Figure 2 shows the optical densities that these strains can grow to. For this, we like to use the rule of thumb that an OD_{600} of 1 corresponds to around a billion cells per mL.

With this number in hand, we can begin to think about how many mutations we expect to occur in our 103 bp region of interest.

(B) Recalling that the replication error rate is about 10^{-9} per bp in *E. coli*,

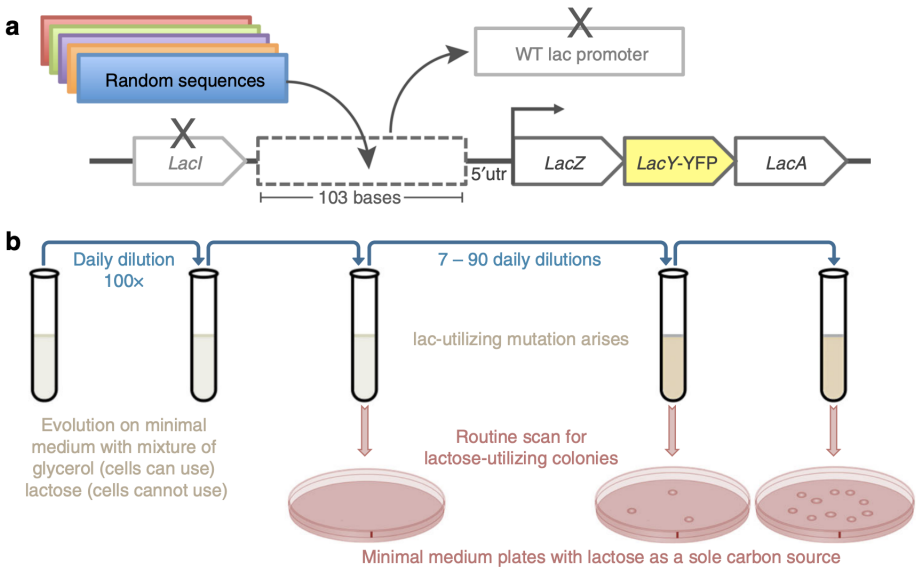

Figure 1: **Experimental set up from Yona et al. 2018.** The initial strains we are working with have one of four random sequences driving the expression of the lactose metabolizing genes. As such, these genes are unable to effectively express the *lacZYA* genes and can only metabolize the glycerol that is found in the media. Each day we will dilute the strains by a factor of 100 and periodically test if the strains have evolved the ability to metabolize lactose.

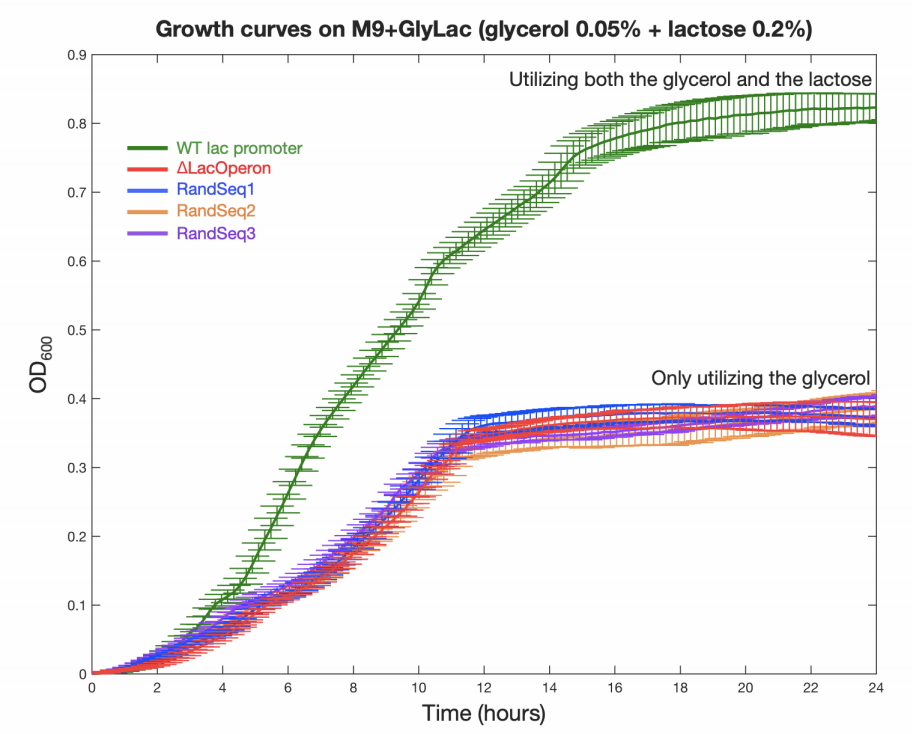

Figure 2: **Growth curves of the Gore strains as determined by optical density (OD) over time.** While a wildtype strain can metabolize both glycerol and lactose found in our media, the random sequences can only make use of the glycerol.

and using the number you found above, how many mutations do you expect to occur in our promoter region of interest?

The next surprising piece of this story was that it often only took a single point mutation to take a random sequence to a functional promoter. As a worst case scenario, let's consider that only **one** such mutation exists that can cause such an improvement.

(C) Given the number of mutations you expect to occur in our promoter region during a day of growth, how many of these mutation do you expect to be the **one** mutation we are looking for to make an active promoter region? Or rather, what is the chance of finding that specific mutation during a day of evolving?

At this point, you might be surprised by how readily this desired mutation can arise, but there is one last wrinkle to address. The dilution of 100 is both a blessing and a curse. It simultaneously serves as the mechanism by which our cells can accumulate mutations, but also as a way to remove genetic diversity from our population. While our desired mutation may very well appear somewhere in the culture, we also need to consider the chance that it gets carried over to the next day. To think about this, we will harken back to our discussions of the Luria-Delbrück experiment, where the timing of when a mutation occurred has important implications for how well-represented it is in the final population.

(D) Let's consider the fact that we are diluting by a factor of 100 each day. This means there are roughly 7 doublings (since $2^7 \approx 100$) each day. Fill out the table below as a way to guide your thinking about the relative importance of a mutation appearing earlier rather than later.

From filling out the table, we can begin to appreciate that while a mutation that occurs early on will be more well-represented in the final culture, it's also the case that the desired mutation is more likely to occur at a later doubling due to the sheer number of cells that are at play by this point.

(E) Building off of your table in part (D), compute the total number of cells you expect in the final culture to have your mutation of interest. This amounts to computing something like a weighted average, where you account

Doubling #	**# cells produced**	**Chance a cell has desired mutation**	**# doublings remaining**	**# ancestors cell will have in final culture**
1				
2				
3				
4				
5				
6				
7				

for the number of cells that are produced at each doubling as well as the number of progeny you expect such a cell to have. Lastly, what is the chance that one of these mutated cells gets carried over in the 1/100 of culture that starts the next day's experiment?

(F) Given the chance of your desired mutation appearing and surviving one day to the next, how many days do you expect this evolution experiment to take? How does your number jibe with the number of days seen in the original experiment?

(G) Discuss at least two assumptions that you made during this estimate and how they might have affected your solution.