MCB137L/237L: Physical Biology of the Cell
Spring 2022
Homework 11:
(Due 4/14/20 at 11:00pm)

Hernan G. Garcia

This homework draws from biological phenomena and physical models we explored throughout the semester. Note that this homework is for extra credit and that it is completely optional. As we are already going to drop your two lowest-scoring homework, if you complete this homework, we will use it to replace your third lowest-scoring homework. However, we will only do so if it actually *improves* your grade.

Note to MCB237L students: Do problem (1), and one out of the remaining three problems. You can choose to do more problems for extra credit (15% of total possible score per extra problem).

1. Counting Proteins with Partitioning Statistics

One of the great challenges in quantitative cell biology is to be able to turn the fluorescence values obtained from fusions to proteins to an actual absolute number of proteins. While there are many ways to "calibrate" such measurements using standards of a known concentration, in this problem, we will explore how we can use bacterial cell division, pure thought and the binomial distribution in order to calibrate a fluorescent protein. To make this possible, we will rely on the calculations on partitioning of carboxysomes that you did during discussion with Yovan and Yasemin.

(a) Begin by reading the paper by Rosenfeld *et al.* entitled "Gene Regulation at the Single-Cell Level" (posted on the website with the homework) and write a one paragraph commentary on the paper with special reference to how they used the binomial partitioning as a way to count repressor proteins. What is the experiment they did and what were they trying to learn?

In the rest of the problem we work out for ourselves the ideas about binomial partitioning introduced in the Rosenfeld *et al.* paper in order to consider the concentration of proteins as a function of time in dividing cells. In particular, the point of this problem is to work out the concentration of protein given that we start with a single parental cell that has N copies of this protein. In the Rosenfeld experiment, at some point while the culture is growing, the

production of the protein is stopped by providing a chemical in the medium and then the number of copies per cell is reduced as a result of dilution as the cells divide.

Interestingly, this problem opens the door to one of the most important themes in physics, namely, that of fluctuations. In particular, as the cells divide from one generation to the next, each daughter does not really get $N/2$ copies of the protein since the dilution effect is a stochastic process. Rather the partitioning of the N proteins into daughter cells during division follows the binomial distribution. Analyzing these fluctuations can actually lead to a quantification of the number of copies of a protein in a cell.

(b) We think of the N copies of the protein as being divided between the two daughters with N_1 going to daughter 1 and $N - N_1$ going to daughter 2. Explain how the probability of N_1 proteins going to daughter cell one is given by the binomial distribution

$$P(N_1, N) = \binom{N}{N_1} p^{N_1} q^{N-N_1}, \tag{1}$$

where the probability of a protein going to daughter cell 1 is p, and the probability of a one protein going to daughter 2 is $q = 1 - p$. For your explanation you can choose to show a formal mathematical derivation, or qualitatively walk us through the meaning of each term in the equation. Remember that, while for most of the course we could use the "stadium seating" approximation to think about how to place N_1 spectators in N seats, here N and N_1 are of comparable magnitude. This situation, which already encountered in the context of the DNA entropic spring, calls for the binomial coefficient $\binom{N}{N_1}$.

We can also calculate the mean of the probability distribution (also called the first moment of the distribution) by invoking a cool trick using the derivative with respect to p

$$\langle N_1 \rangle = \sum_{N_1=0}^{N} N_1 \binom{N}{N_1} p^{N_1} q^{N-N_1} = p \frac{\partial}{\partial p} \sum_{N_1=0}^{N} \binom{N}{N_1} p^{N_1} q^{N-N_1}. \tag{2}$$

This equation can be rewritten as

$$\langle N_1 \rangle = p \frac{\partial}{\partial p} \left((p+q)^N \right) = p N_{mother} (p+q)^{N-1}, \tag{3}$$

where we made use of the fact that

$$\sum_{N_1=0}^{N} P(N_1, N) = (p+q)^N. \tag{4}$$

Using $p + q = 1$, Equation 3 leads to

$$\langle N_1 \rangle = pN. \tag{5}$$

(c) Work out the expected averaged fluctuations squared in the partitioning process after each division by noting that the averaged fluctuations can be written as $\langle (N_1 - N_2)^2 \rangle$, where N_1 and N_2 are the number of proteins that end up in daughter cells 1 and 2, respectively. Show that, if $p = q = 0.5$, the partitioning error is given by $\langle (N_1 - N_2)^2 \rangle = N$. To make this possible, use the derivative trick twice such that

$$\langle N_1^2 \rangle = \sum_{N_1=0}^{N} N_1^2 \binom{N}{N_1} p^{N_1} q^{N-N_1} = p \frac{\partial}{\partial p} \left[p \frac{\partial}{\partial p} \left(\sum_{N_1=0}^{N} \binom{N}{N_1} p^{N_1} q^{N-N_1} \right) \right] \tag{6}$$

as well as the result $\langle N_1 \rangle = pN$ described above. In addition, use the fact that $N = N_1 + N_2$, in order to calculate the average partioning error as

$$\langle (N_1 - N_2)^2 \rangle = \langle [N_1 - (N - N_1)]^2 \rangle = \langle (2N_1 - N)^2 \rangle. \tag{7}$$

Remember that $\langle N \rangle = N$, as N is a constant in our problem.

(d) Next, look at the Rosenfeld paper and explain how measuring fluorescence variations can be used to calibrate the exact number of copies of the fluorescent protein in a cell. Specifically, assume that the fluorescence intensity in each cell can be written as $I = \alpha N$, where α is an as-yet unknown calibration factor and N the number of proteins in the cell. Explain what this equation means and why you think it is justified. Derive an expression relating I_1, I_2 and I_{tot} using the result of part (c). Make a qualitative schematic showing a plot of $\langle (I_1 - I_2)^2 \rangle$ versus I_{tot} and explain how to get the calibration factor α from this plot. Note that we're asking to draw up an explanation, not to actually make a plot with Python..

(e) Now we are going to repeat the Rosenfeld experiment numerically in order to *fit* the calibration factor. Consider a fluorescent protein such that the calibration factor between the intensity and the number of fluorophores is 50, that is $I = 50N$. Generate intensity data by choosing $N_1 + N_2 = 10, 50, 100, 1000$ and 5000 and for each case, "partition" the proteins from the mother cell to the two daughters 100 times (i.e. as if you are looking at 100 mother cells divide for each choice of the protein copy number). To make this possible, flip a coin for each molecule in order to decide whether the molecule is going to daughter cell 1 or 2 (and remember how we've done something similar to this earlier in the course when modeling diffision as coin flips). Then, make a plot of the resulting $\langle (I_1 - I_2)^2 \rangle$ vs I_{tot} just as we did analytically in the previous problem. What I mean is that you need to make a plot of all of your simulation results. Then, do a fit to your "data" using a numpy function (see the note below) and see how well you recover the calibration factor that you actually put in by hand. Plot the fit on the same graph as all of the "data".

Note: You can use `numpy.polyfit` to perform a linear fit to your "data" using the syntax `numpy.polyfit(x, y, deg)` where `x` is the data x-coordinate, `y` is the data y-coordinate, and `deg` is the degree of the polynomial you'd like to fit to your data (for instance, you would use `deg = 1` for a linear fit). You can also use `numpy.linalg.lstsq` if you'd rather phrase the problem as a matrix equation (this is reasonably simple to do as well, and an example of a linear fit performed using this function is provided in the Numpy documentation linked

to above).

2. The Bleach-Chase Method

In class we talked about the molecular census. Synthesis of new proteins is not the only kinetic process that governs how many proteins will be found in a given cell. An additional kinetic process of great importance is the decay of the proteins. In a recent paper by Eden *et al.* (see the course website for this paper), the rate of protein decay was measured using a technique called "bleach-chase".

(a) Read the paper and write a one paragraph summary of what the paper is about, the essence of the method and the results. Note also that on all of the derivations that follow in this problem, you will be graded not only on having the right equations but also upon the clarity and logic of your presentation. Your job is to use the mathematics to explain how this method works and to explain precisely what is measured and how it is analyzed. This problem is an example of how we can take something right out of today's most very recent research results and turn it into a little story that you could explain to your scientifically-minded friends. Further, the question of how to actually go about measuring protein degradation rates is very hard and this paper shows us a very interesting answer to that question.

(b) To make the degradation process accessible, these cells have a fluorescent protein fused to some protein of interest whose degradation rate is the subject of enquiry. (NOTE: one of the things that they worried about and we should worry about too as readers is that maybe the act of tethering a big fluorescent protein to the protein of interest would alter its degradation rate, the very quantity we are trying to measure.) In the absence of any photobleaching, the evolution of the total number of fluorescent proteins, N_f, follows the simple dynamical equation

$$\frac{dN_f}{dt} = \beta - \alpha N_f, \tag{8}$$

which acknowledges a rate of protein production β and a degradation rate α. Explain what this equation means and solve it as a function of time assuming that the initial number of proteins is zero. You can solve it analytically or numerically using Matlab. If you solve it numerically, plot it for a reasonable choice of parameters such as the Bicoid production and degradation rates we considered for our homework assignments earlier in the semester. What is the steady-state value of the number of fluorescent proteins per cell? What is the characteristic time scale to reach this steady state in terms of parameters α and β?

(c) The problem of trying to infer the degradation rate from just looking at the amount of fluorescent protein is, however, that its dynamics is dictated not only by the degradation rate, but also by the production rate. The question is: can we create an experimental condition where we measure a quantity that is solely determined by the degradation rate? The experimental idea is that we have two populations of identical cells. What this means is that these populations have the same genomes, have been subjected to the same growth conditions and environmental stimuli. In practice, what this really means is that the average

fluorescence intensity of our protein of interest in the two populations is the same. At a certain instant in time, we then photobleach one of the two populations so that their fluorescent intensity is now reduced relative to its initial value and relative to the value in the unphotobleached population. We now have two populations. First, we have the number of unbleached molecules, N_u. Second we have the number of bleached molecules, N_p with a conservation law which is that the total number of proteins of our species of interest is given by $N_u + N_p$.

A similar equation to Equation 8 describes the dynamics of the unphotobleached molecules in the cells that have been subjected to photobleaching, with the number that are unphotobleached given by N_u and described by the dynamical equation

$$\frac{dN_u}{dt} = \beta - \alpha N_u. \tag{9}$$

On the other hand, the number of photobleached proteins are subject to a different dynamical evolution described by the equation

$$\frac{dN_p}{dt} = -\alpha N_p, \tag{10}$$

since all that happens to them over time is that they degrade. Explain why there are two populations of proteins within the photobleached cells and why these are the right equations.

(d) Notice that, while the evolution of the unbleached population still depends on both the production and degradation rates, the time evolution of the bleached species is only dependent on the degradation rate. As a result, if we could track the amount of unbleached molecules as a function of time we would have direct access to the degradation rate α. The trick used by the authors in the paper is to evaluate the difference in the number of fluorescent proteins in the two populations. An absolutely critical assumption then is that

$$\frac{dN_f}{dt} = \frac{dN_u}{dt} + \frac{dN_p}{dt} \tag{11}$$

The point is that over time after photobleaching, the photobleached cells will become more fluorescent again as new fluorescent proteins are synthesized.

The key idea here is then to plot the difference between the intensity of the cells that were not disturbed by photobleaching and those that were. In particular, show the simple result that

$$\frac{d(N_f - N_u)}{dt} = -\alpha(N_f - N_u). \tag{12}$$

Note that this quantity is directly experimentally accessible since it calls on us to measure the level of fluorescence in the two populations and to examine the difference between them. Further, note that the dynamics depends only upon the one parameter that we are trying to measure. Integrate this equation by solving it analytically or numerically in Matlab, and show how the result can be used to determine the constant α that characterizes the dynamics of protein decay. Explain what you would actually plot if you were making the measurements and how that would yield the parameter α.

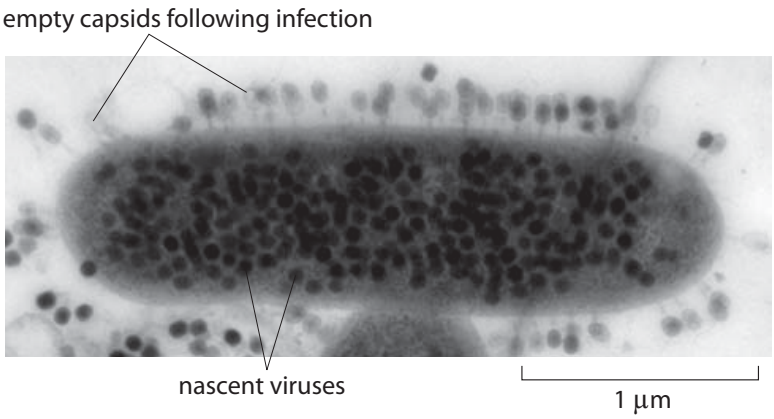


Figure 1: Synthesis of new viruses in an infected bacterium.

(e) In the paper, they discuss both degradation and dilution. Explain the distinction between these two ideas. Comment on which you think is dominant in bacteria based on what you learned from this paper.

3. Physical Biology of Viruses

In this problem, we take a random walk through the physical biology of viruses, honoring them as one of the most sophisticated, interesting and scary parts of the biological world.

(a) One of the most important properties of a given infection is the so-called "burst size", the number of new viruses produced per infected cell. One of the original hypotheses (which you will refute here) for what controls the burst size is the available volume within the host cell. Given that for a typical bacteriophage infection the burst size is roughly 100 viruses, what fraction of the volume is taken up by the newly synthesized viruses? Figure 1 shows an electron microscopy image of an infected bacterium.

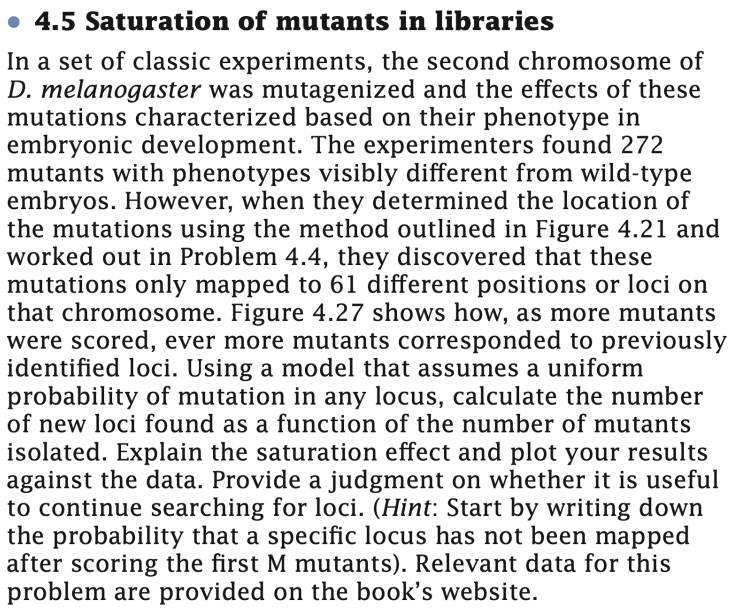
(b) How are viruses transmitted? Three key routes are through the respiratory tract, the digestive tract and the reproductive tract. In all three cases, our bodies are set up with a number of different tricks to resist infection including mucus and ciliary transport in our respiratory and digestive tracts and harsh conditions in our digestive tract such as low pH. The current coronavirus epidemic is apparently passed through the respiratory tract and in this part of the problem, we appeal to Figure 2 for a look at the distribution of droplet sizes. How many particles are contained in a typical cough or sneeze? How much volume is that? Given a viral concentration in sputum of 10^6 to 10^{11} RNAs/ml, estimate how many virions of SARS-CoV-2 will be carried in a typical droplet. A very interesting source of information on this is the work of Prof. Lydia Bourouiba from MIT who does visualization experiments on humans coughing. You can also see this excellent brief interview with Bourouiba on the physics of sneezing and coughing.

TABLE II. Numbers of Particles in Different Initial Diameter Ranges Emitted in One Cough and One Sneeze According to Duguid

Diameter Range (μm)	Number of Particles in a Cough	Number of Particles in a Sneeze
1–2	50	26,000
2–4	290	160,000
4–8	970	350,000
8–16	1600	280,000
16–24	870	97,000
24–32	420	37,000
32–40	240	17,000
40–50	110	9000
50–75	140	10,000
75–100	85	4500
100–125	48	2500
125–150	38	1800
150–200	35	2000
200–250	29	1400
250–500	34	2100
500–1000	12	1000
1000–2000	2	

Source: Data from Duguid, "The Size and Duration of Air-Carriage of Respiratory Droplets and Droplet-Nuclei." *Journal of Hygiene* 4:471–480, Table 3 (1946).

Figure 2: Distribution of droplet sizes after a sneeze.



4.5 Saturation of mutants in libraries

In a set of classic experiments, the second chromosome of *D. melanogaster* was mutagenized and the effects of these mutations characterized based on their phenotype in embryonic development. The experimenters found 272 mutants with phenotypes visibly different from wild-type embryos. However, when they determined the location of the mutations using the method outlined in Figure 4.21 and worked out in Problem 4.4, they discovered that these mutations only mapped to 61 different positions or loci on that chromosome. Figure 4.27 shows how, as more mutants were scored, ever more mutants corresponded to previously identified loci. Using a model that assumes a uniform probability of mutation in any locus, calculate the number of new loci found as a function of the number of mutants isolated. Explain the saturation effect and plot your results against the data. Provide a judgment on whether it is useful to continue searching for loci. (*Hint*: Start by writing down the probability that a specific locus has not been mapped after scoring the first M mutants). Relevant data for this problem are provided on the book's website.

Figure 3: Problem 4.5 from PBoC2.

4. Saturation of mutant libraries

One of the most important aspect of genetic screens (and life in general) is to recognize when you've reached the point of diminishing returns. To explroe this in the context of the genetic screen by Wieschaus and Nüsslein-Volhard, do problem problem 4.5 form PBoC shown in Figure 3. Note that Problem 4.4 mentioned to in the statement refers to Problem 3 of Homework 10. Figure 4.21 from PBoC is shown in Figure 4 while Figure 4.27 is shown in Figure 5.

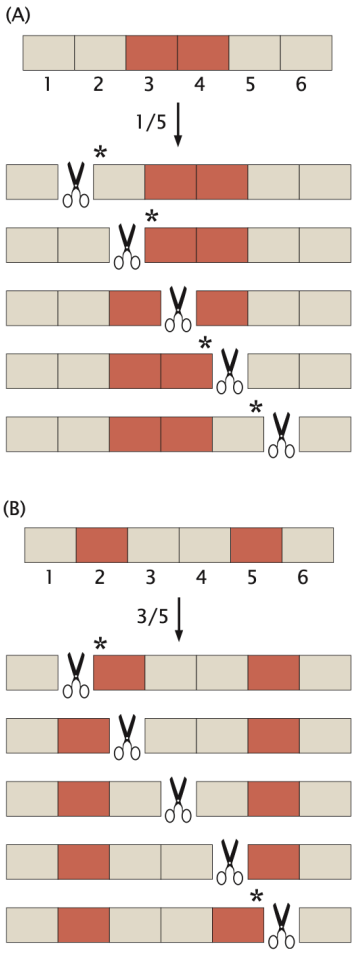


Figure 4.21: Concept of mutation correlation and physical proximity on the gene illustrated by labeling a string in two distinct points and then making random cuts of the string. The probability that the two labels will remain on the same part of the string *after* the cut depends upon their physical proximity *before* the cut. (A) Two mutations (red boxes) that are close to one another are likely to remain on the same part of the string (four times out of five). (B) Two mutations that are further apart are more likely to be separated (remaining together only two times out of five).

Figure 4: Figure 4.21 from PBoC2.

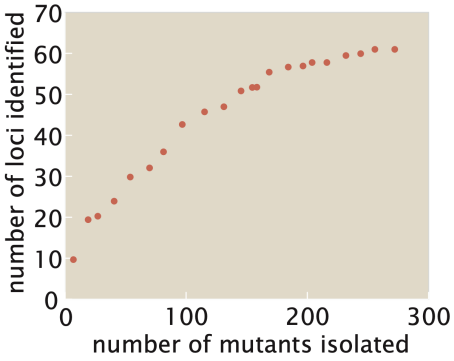


Figure 4.27: Saturation of a mutant library. Number of different identified loci as a function of the number of mutants isolated. (Adapted from C. Nusslein–Volhard et al., *Roux's Arch. Dev. Biol.* 193:267, 1984.)

Figure 5: Figure 4.27 from PBoC2.