MCB137L/237L: Physical Biology of the Cell
Spring 2022
Homework 1: Biological Numeracy
(Due 1/27/22 at 11:00am)

Hernan G. Garcia

"The greats weren't great because at birth they could paint. The greats were great cause they paint a lot." - Macklemore

Homeworks in MCB137/237

Whether it is in the context of professional sports, art, or science, it pays to practice. The idea of the homeworks during our course is to give you a venue to get your 10,000 hours of calculations and estimates in as a means to become proficient in the mathematical and physical modeling of living systems. Sometimes, the problems will ask you to redo a derivation we did in class in a new way, and sometimes they will propose a whole new biological phenomenon to attack. Regardless, if you spend more than five hours on a homework set, it means that you should come to office hours. Make sure to start working on your problems early on!

The objective of this homework set is to get a feeling for the numbers in whatever problem you're considering in biology. Just like you always need to check the units in your calculations, a more subtle sanity check of your theoretical results stems from having some expectation about the order of magnitude you will obtain.

This first problem set involves a number of challenges in order-of-magnitude thinking. When doing street fighting estimates, the goal is to do simple arithmetic of the kind that all numbers are 1, few or 10. few $\times$ few $= 10$, etc. Please do not provide estimates with multiple "significant" digits that are meaningless. Be thoughtful about what you know and what you don't know. You may use the Bionumbers website http://bionumbers.hms.harvard.edu/ to find key numbers (examples are masses of amino acids (BNID 104877) and nucleotides (BNID 103828), the speed of the ribosome (BNID 100059), etc.), but please provide a citation to the Bionumber of interest as shown above. However, for many of these problems the essence of things is to do simple estimates, not to look quantities up.

Sometimes, the problems will be drawn directly from the 2nd edition of Physical Biology of the Cell (PBoC or PBoC2). In that case, I'll make the effort to scan the problems and include them as a figure. However, some of those problems might refer to information inside the book, which I will not scan. As a result, I highly recommend that you just get the book.

Homework submission: Gradescope will be used to submit and grade your homework. We will create two submissions for weekly homework (one for written pdf, the other one for submitting zip file for your code). When you have to write Python code in order to make plots, you don't need to include your actual code in the pdf document you submit. However, you need to submit all the codes you used to generate the plots in a separate zip file. All plots you generate need to have axes and lines that are clearly labeled. Please submit both pdf and original code before the deadline.

Finally, remember to write each problem on a different piece of paper so that you can upload them independently to Gradescope. This will make it easier for us to grade them.

How to join Gradescope:

1. Go to website: https://www.gradescope.com
2. Create an account
3. Add class with entry code: **4PE35G**
4. Please update your "Student ID" in the account settings

Extra Credit. Provide comments on the parts of chap. 2, "Setting the Scales of Living Things" of the upcoming third edition of *Physical Biology of the Cell* you will find with the online link to this homework. Note that this is an unfinished draft of the chapter. Figure placements are not necessarily correct and there are still a number of internal discussions amongst the author team about how to finish things off. We are especially interested in mistakes, flaws in logic, confusing figures, unclear discussions, etc., but are happy to entertain comments at all scales. This extra credit will constitute an additional 15% on your score on the homework. Please comment on the PDF directly, or use pen and paper and then scan the document. **Submit your comments directly via email to Hernan (hggarcia@berkeley.edu, not through GradeScope.**

1. Benjamin Franklin and Molecular Dimensions.

In his travels between America and Europe, Benjamin Franklin was subjected to the vicissitudes of the sea which led him to reflect on his reading of Pliny the Elder and claims of how oil was known to smooth the waves. Upon arriving in England, Franklin took the concept to the test. He tells us of his experience thus: "At length at Clapham where there is, on the common, a large pond, which I observed to be one day very rough with the wind, I fetched out a cruet of oil, and dropped a little of it on the water. I saw it spread itself with surprising swiftness upon the surface... the oil, though not more than a teaspoonful, produced an instant calm over a space several yards square, which spread amazingly and extended itself gradually until it reached the leeside, making all that quarter of the pond, perhaps half an acre, as smooth as a looking glass."

(a) Though Franklin himself never made the estimate (that was to await Lord Rayleigh), use Franklin's description of the experiment to work out the thickness of the oil film (the

height of a lipid!) that covered the surface of Clapham common pond.

(b) Using a typical molecular mass for a lipid (say, 1000 g/mol), work out the number of lipid molecules that covered that surface of the pond and use that number to compute the area per lipid. How do your results compare to the modern values for the size of lipids?

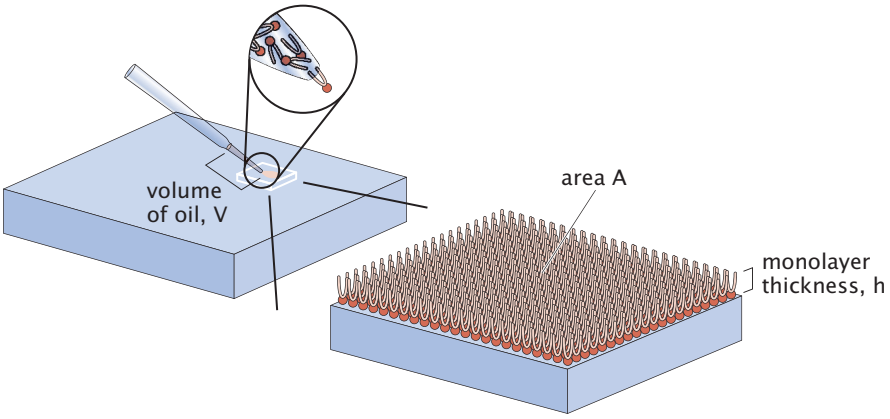

Figure 1: Putting oil on water to measure molecular dimensions. Here we see that the lipid molecules form a monolayer.

2. Street fighting the ribosome.

One of the most important molecular assemblies in the cell is the ribosome. The number of ribosomes per cell dictates how fast cells can grow. *E. coli* growing with a division time of 24 minutes have 72,000 ribosomes per cell, and slow growing *E. coli* with a division time of 100 minutes have a factor of ten fewer ribosomes with a count of $\approx$ 6800 ribosomes.

(a) In this part of the problem, we will use our street fighting skills to explore the ribosomal density in another organism as shown in Figure 2, and then see how well our results from the electron microscopy study square with the numbers quoted above. By examining the figure, make an estimate of the number of ribosomes per μm^3 and compare that result to the numbers quoted for *E. coli* above.

(b) In a beautiful turn of the millennium paper by Tania Baker and Stephen Bell whose abstract is shown in Figure 3, they imagined a world in which DNA polymerase was the size of a FedEx truck and explored what copying DNA would look like. Write a one-paragraph abstract of your own which carries out a similar analysis, but this time for the ribosome.

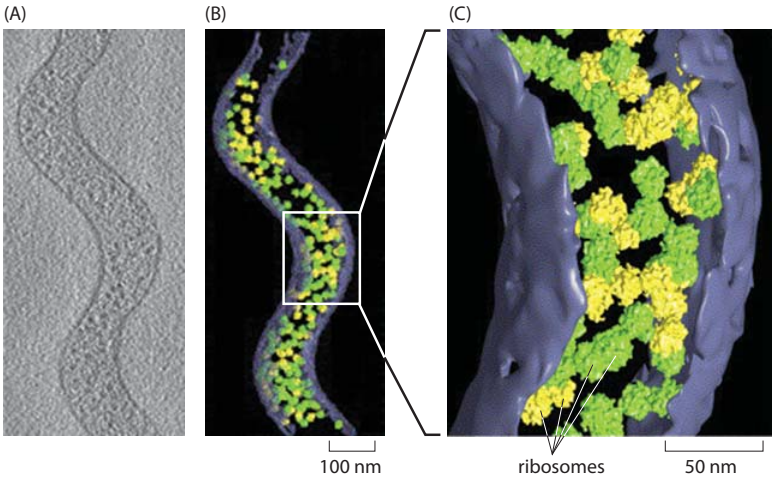

Figure 2: Cryo EM study of a bacterial cell. These images are of the tiny bacterium, *Spiroplasma melliferum*. Using algorithms for pattern recognition and classification, components of the cell such as ribosomes were localized and counted. (A) Single cryo-electron microscopy image. (B) 3D reconstruction showing the ribosomes that were identified. Ribosomes labeled in green were identified with high fidelity while those labeled in yellow were identified with intermediate fidelity. (C) Close up view that you should use to make your count. Adapted from JO Ortiz *et al.*, J. Struct. Biol. 156, 334-341 (2006).

Cell, Vol. 92, 295–305, February 6, 1998, Copyright ©1998 by Cell Press

Polymerases and the Replisome: Machines within Machines

Review

Tania A. Baker* and Stephen P. Bell
Department of Biology
*Howard Hughes Medical Institute
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Synthesis of all genomic DNA involves the highly coordinated action of multiple polypeptides. These proteins assemble two new DNA chains at a remarkable pace, approaching 1000 nucleotides (nt) per second in *E. coli*. If the DNA duplex were 1 m in diameter, then the following statements would roughly describe *E. coli* replication. The fork would move at approximately 600 km/hr (375 mph), and the replication machinery would be about the size of a FedEx delivery truck. Replicating the *E. coli* genome would be a 40 min, 400 km (250 mile) trip for two such machines, which would, on average make an error only once every 170 km (106 miles). The mechanical prowess of this complex is even more impressive given that it synthesizes two chains simultaneously as it moves. Although one strand is synthesized in the same direction as the fork is moving, the other chain (the lagging strand) is synthesized in a piecemeal fashion (as Okazaki fragments) and in the opposite direction of overall fork movement. As a result, about once a second one delivery person (i.e., polymerase active site) associated with the truck must take a detour, coming off and then rejoining its template DNA strand, to synthesize the 0.2 km (0.13 mile) fragments.

In this review we describe our current understanding of the organization and function of the proteins of the replication fork and how these complexes are assembled at origins of replication. Understanding the architecture of DNA polymerases is relevant to RNA polymerases as well, as the core of the polynucleotide polymerization machine appears to be similar for all such enzymes. In the discussion of the replisome, we particularly focus on features shared by the machinery from different organisms.

Polymerases: Template-Directed Phosphoryl Transfer Machines

Synthesis of the new DNA strands occurs as a result of a collaboration between the synthetic capacities of multiple polymerases. Two types of polymerases are required: primases, which start chains, and replicative polymerases, which synthesize the majority of the DNA (Kornberg and Baker, 1992). The replication fork, however, contains at least three distinct polymerase activities: a primase and a replicative polymerase for each of the two template strands. In *E. coli*, primase is a single polypeptide, and the replicative polymerase is a dimer of DNA polymerase (pol) III core and several accessory proteins that together form the pol III holoenzyme (reviewed in Marians, 1992; Kelman and O'Donnell, 1995). Similarly, phage T4 has one primase and one replicative polymerase that appears to function as a dimer (Alberts, 1987; Munn and Alberts, 1991). The situation in eukaryotic cells is slightly different (Stillman, 1994). The primase is in a tight complex with a DNA polymerase (pol α) and eukaryotic cells have two distinct replicative polymerases: polymerase δ (pol δ) and polymerase ϵ (pol ϵ).

All the replicative polymerases have one large subunit that contains the polymerase active site and, with the exception of pol α–primase, the same subunit or an associated polypeptide carries a proofreading 3′→5′ exonuclease. The polymerase subunits also interact with proteins that dramatically influence their association with DNA. In *E. coli*, the replicative polymerase is found in a complex with proteins that control polymerase processivity; this holoenzyme, consists of 10 distinct polypeptides (Kelman and O'Donnell, 1995). In contrast, neither the T4 nor the eukaryotic polymerases copurify in a complex with the processivity factors (Alberts, 1987; Stillman, 1994). Therefore, these proteins are called accessory proteins rather than subunits (see Table 1).

Polymerase Architecture. The central feature of all the known polymerase structures is the existence of a large cleft comprised of three subdomains referred to as the fingers, palm, and thumb by virtue of the similarity of the structures to a half-opened right hand (Figure 1; polymerase structures are reviewed in Joyce and Steitz, 1994, 1995; Sousa, 1996). A diverse set of polymerases—

Figure 3: Abstract of a paper from Tania Baker where she maps the action of DNA polymerase onto human length scales to give a sense of its amazing properties. This parable is the basis of your own analysis of the ribosome. Adapted from Baker TA and Bell, SP Cell, Vol. 92, 295:305, February 6, (1998).

3. Composition of a cell.

Here we are going to do a rough atomic census of living material by thinking about the principal ingredients of a cell. To get a sense of the chemical makeup of the dry mass of a cell, we are going to focus only on proteins and nucleic acids.

(a) Provide a simple and clean estimate for the volume and mass of a typical bacterium such as *E. coli*.

(b) One of the key rules of thumb we will invoke over and over again is a knowledge of the concentration of one molecule per *E. coli* cell. Using the volume from part (a), work out a simple estimate for the concentration of 1 molecule per *E. coli* cell. Remember that we are in street-fighting mode and thus your answer should be 1, few or 10 in nM, μM, mM or M.

(c) Assume that 1/3 of the mass of a bacterium is dry mass and for simplicity, we ascribe all of that dry mass either to proteins or nucleic acids. We will take our elemental composition of a "typical" amino acid to be $N_1C_5O_2H_8$ and a "typical" nucleotide to be $P_1N_5O_7C_{10}H_{14}$. Given that roughly half the dry mass of the cell is protein, work out the number of proteins and hence, the number of amino acids per cell.

(d) As an alternative approach to estimating the total number of proteins in *E. coli*, assume that the bacterium is tightly packed with proteins (think of golf balls in a bathtub). How does this compare to the estimate from part (c)?

(e) Work out the number of nucleotides in the genome of our bacterium of interest.

(f) Finally, figure out how many ribosomes are needed, translating at roughly 15 aa per second to translate all of those proteins. How many nucleotides are present in all of these ribosomes?

(b) Given all of these numbers from the rest of this problem, you are now able to work out the overall composition of a cell. Provide an approximate formula for the stoichiometry of a bacterium.

4. To build a cell.

Minimal growth medium for bacteria such as *E. coli* includes various salts with characteristic concentrations of mM and a carbon source. This carbon source is typically glucose and it is used at 0.2% (a concentration of 0.2 g/100 mL).

(a) Make an estimate of the number of carbon atoms it takes to make up the macromolecular contents of a bacterium such as *E. coli*.

(b) Make an estimate of the number of nitrogen atoms it takes to make up the macromolecular contents of a bacterium such as *E. coli*.

(c) How many cells can be grown in a 5 mL culture using minimal medium before the medium exhausts the carbon? Note that this estimate will be flawed because it neglects the *energy* cost of synthesizing the macromolecules of the cell. Similarly, given that the recipe for minimal media requires ammonium chloride NH_4Cl at a concentration of 100 mM, how many cells can be grown in a 5 mL culture using minimal medium before the medium exhausts the nitrogen?

(d) In rapidly dividing bacteria, the cell can divide in times as short as 1200 s. Make a careful estimate of the number of sugars (glucose) needed to provide the carbon for constructing the macromolecules of the cell during one cell cycle of a bacterium. Use this result to work out the number of carbon atoms that need to be taken into the cell each second to sustain this growth rate.

(e) These problems are intended to get you thinking about the wondrous process whereby cells convert a clear liquid with simple chemical ingredients into biomass as shown in Figure 4. Amazing! Now, work out an estimate related to the volume of the headspace you see in Figure 4 which has oxygen available for cell growth. Specifically, if 6 O_2 molecules are consumed for every sugar, make a simple estimate of the required volume of headspace needed to sustain cell growth. Note that our estimate about O_2 usage is crude and sloppy. To really do this carefully, we need to acknowledge the use of glucose both in providing building materials (i.e. carbon skeletons) as well as the energy needed to synthesize a cell. The estimate we do here is intended to give an impression of the magnitudes, and specifically to get a sense of the aeration requirements when we do a liquid culture growth procedure.

5. Sizing up the Central Valley.

California's Central Valley is one of the most potent agricultural regions in the world. In this problem, you are going to evaluate many of the key factors associated with its enormous productivity without any data aside from a single satellite image of the region as shown in Figure 5. Note that the key point here (and what you will be graded for if you care about such things) is the logical flow of your estimates, not the particular numerical values you found.

(a) Water usage. Using what you know about watering and the growth of plants, make an estimate of the amount of water used to irrigate the agriculture of the Central Valley.

(b) Nitrogen usage. Since the beginning of the twentieth century, we have doubled the number of occupants that can be fed on earth as a result of the Haber-Bosch process and the synthetic fixation of nitrogen. In this part of the problem, begin by estimating the number of kilograms of biomass per square meter that is produced per year. From that number, figure out how many kilograms of nitrogen are contained per square meter of biomass. Then, make an estimate of how much fertilizer is used for each square meter and hence for the entirety of the Central Valley.

Figure 4: Growth of *E. coli* in rich media. The tube on the left shows roughly 5 mL of growth media just after inoculation. The tube on the right shows such media after saturation due to exponential cell growth and division.

(c) Pesticide usage. Undertake an estimate similar to that in the first two parts of the problem to figure out how much pesticide is used on the Central Valley every year.

(d) Do NOT do this part until you have done parts (A) - (C). Look up some source of data on each of these three questions and compare your results to the data. Please do not redo your estimate.

6. The pandemic elephant in the room.

We are living through a global pandemic that has changed all of our lives in far reaching ways. As a result, each week, we will have at least one problem that reminds us of the pandemic, and asks us to think about it quantitatively. In this problem, we are going to explore the mass of an individual SARS-CoV-2 virion, the total mass of such viruses within a given individual at the peak of their infection and the total mass of all the SARS-CoV-2 viruses on the planet.

(a) Given the roughly ≈ 100 nm diameter of a single SARS-CoV-2 virion, work out a simple estimate for its mass. What fraction of that mass corresponds to the genome? To answer the latter question, use simple rules of thumb for the mass of a nucleotide and use the fact that this virus is a single-stranded RNA virus with a roughly ≈ 30 kb genome.

(b) There are a number of cell types in different tissues that are susceptible to infection by SARS-CoV-2. For our purposes, we are going to focus on the most massive such tissue,

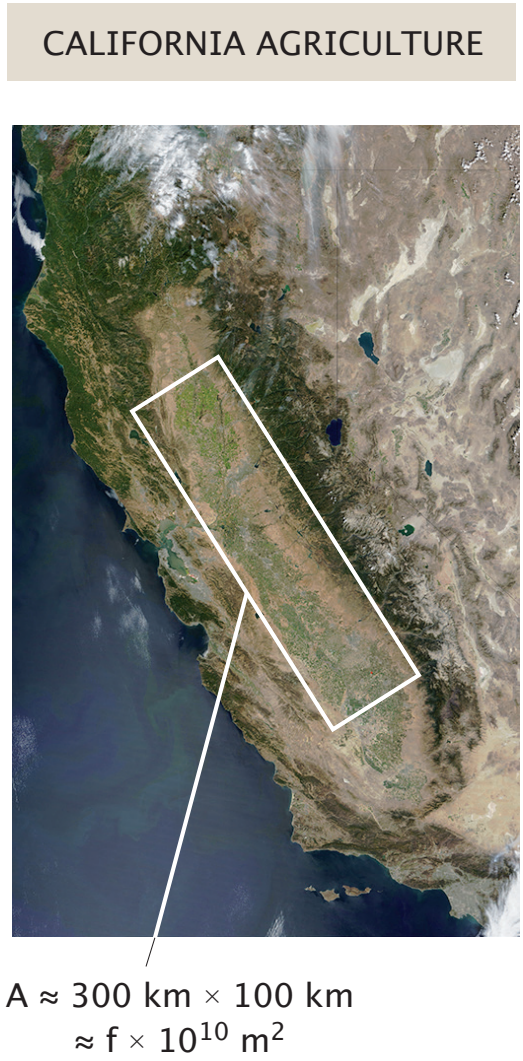

Figure 5: Satellite image of California's Central Valley.

namely, the lungs. There are several different assays for measuring the viral load within an infected individual. One method is to use RT-PCR to amplify their nucleic acid content with the result that there are between $10^6 - 10^8$ RNA copies per gram of lung tissue. Alternatively, infectious virions are measured by using cells in tissue culture and figuring out at what concentration of viruses half of the tissue culture cells will be infected, the so-called TCID50 (tissue-culture infectious dose). Samples from lung tissue yield the range of $10^2 - 10^4$ TCID50 per gram of lung material. Using these results, estimate the total number of virions in the lung and comment on the difference between the RNA-based assay and the infection assay. Given these numbers, what is the total mass of viruses within an infected individual at the peak of their infection? To the extent that our estimate is correct, what fraction of virions are actually infectious?

(c) Use the results of the previous two parts of the problem to estimate the total mass of all the SARS-CoV-2 viruses that have been present in the human population since the beginning of the pandemic.